Advances in Applied Mathematics 应用数学进展, 2020, 9(11), 1961-1969
Published Online November 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.911226

Hans汉斯

改进支持向量机在睡眠分期预测模型中的应用

许淑婷*，韩成志，郑斌斌，孙莹莹

杭州师范大学理学院，浙江 杭州
Email: 1484796314@qq.com, 1224764141@qq.com, 1250933904@qq.com, 2296933124@qq.com

收稿日期：2020年10月25日；录用日期：2020年11月13日；发布日期：2020年11月20日

摘　要

睡眠分期是评估人类睡眠质量和诊断相关疾病的关键，对于睡眠分期预测的研究已有诸多成果。文中以脑电信号作为睡眠分期的工具，在支持向量机(Support Vector Machines, SVM)分类应用于睡眠分期预测模型的研究基础上，为了减少睡眠分期预测模型的建立的时间，采用K边界近邻法(K Nearest Bound Neighbor, KNBN)支持向量预选取的方法构造支持向量候选集，建立基于KNBN-SVM的睡眠分期预测模型。实验结果表明，该睡眠分期预测模型的预测准确度理想，并且耗时大幅度缩短。KNBN-SVM方法有效地改进了基于标准SVM睡眠分期预测模型，具有实用价值。

关键词

支持向量机，K边界近邻法，预选取，睡眠分期预测，脑电信号

Application of Improved Support Vector Machine in Predicting Sleep Stages Classification

Shuting Xu*, Chenzhi Han, Binbin Zheng, Yingying Sun

College of Science, Hangzhou Normal University, Hangzhou Zhejiang
Email: 1484796314@qq.com, 1224764141@qq.com, 1250933904@qq.com, 2296933124@qq.com

Received: Oct. 25th, 2020; accepted: Nov. 13th, 2020; published: Nov. 20th, 2020

Abstract

Sleep staging is the key to assess human sleep quality and diagnose related diseases. There have

*第一作者。

文章引用: 许淑婷, 韩成志, 郑斌斌, 孙莹莹. 改进支持向量机在睡眠分期预测模型中的应用[J]. 应用数学进展, 2020, 9(11): 1961-1969. DOI: 10.12677/aam.2020.911226

been many results in the study of sleep state recognition. In this paper, EEG signals are used as a tool for sleep staging. Based on the research of support vector machine classification applied to sleep stages classification models, the K nearest boundary neighbor support vector pre-selection method is used to construct support vector candidate set in order to reduce the time to establish a model, build a sleep stage prediction model based on KNBN-SVM. Experimental results show that the prediction accuracy of the sleep staging prediction is ideal, and the time-consuming is greatly shortened. The KNBN-SVM effectively improves the sleep staging prediction based on the standard SVM, and has practical value.

Keywords

Support Vector Machine, K Nearest Bound Neighbor, Pre-Selection, Predicting Sleep Stages Classification, Electroence Phalogram Signal

1. 引言

睡眠是一种自发的、主动的调节过程，睡眠质量对人的身心状态有着重大影响。然而，随着生活节奏的加快，生存压力不断增大，人类的睡眠质量开始出现问题，这使人们在身体上、心理上饱受折磨，也严重影响人类的生活质量。因此，如何提高睡眠质量，减少睡眠相关疾病对健康的影响，受到了广泛的关注。

临床医学上不同时间的睡眠深度是睡眠质量判定的一个主要方面[1]，对人体睡眠的分期研究能够帮助对不同时间的睡眠深度进行准确的评估，是监测和研究人类睡眠质量和相关疾病的关键。人体在进入睡眠时大脑将以一种电信号(Electroence phalogram, EEG)形式在作用，脑电信号在不同睡眠分期所呈现的特点有所不同[2]。基于脑电信号进行自动分期，是评估睡眠质量、诊断和治疗睡眠相关疾病的重要辅助工具，相应的研究对于预防和治疗相关的疾病有着十分重要的临床意义与应用价值。

对于利用脑电信号对睡眠状态进行分期的研究，已有诸多研究成果。2007 年李谷等[3]在基于脑电信号 Hilbert-Huang 变换的睡眠分期研究中对 560 个睡眠脑电信号样本进行分期。2010 年刘慧等[4]在基于模糊熵的脑电睡眠分期特征提取与分类的研究中，对数据进行模糊熵的分析，并且利用支持向量机对睡眠状态进行分类。2012 年 Liang 等[5]使用多尺度熵和线性判别分析法来分析睡眠 EEG 信号并对其分期。2013 年周鹏等[6]使用主成分分析法处理脑电信号，并利用 SVM 实现睡眠分期。2015 年，Diykh M 等[7]对睡眠分期研究中，利用结构图相似和 K-means 的方法对提取的 EEG 信号的时频统计特征量进行分析，但是这个方法实时性不好并且实用价值不大。2016 年，张晓宇等[8]利用快速傅里叶变换处理脑电信号，并对频谱结果做符号化，将符号化之后的结果作为 BP 神经网络的输入参数实现睡眠自动分期。以上研究在睡眠状态分期预测中均取得较理想的效果，但在预测模型构建当中需要耗费大量时间，依然有很大的改进空间。

为了解决耗时过长的问题，本文在支持向量机分类应用于睡眠分期预测模型构建的研究基础上，采用 K 边界近邻法支持向量预选取算法[9]来构建支持向量候选集，将支持向量候选集输入到 SVM 中训练，减少训练样本，在保证良好分类预测准确度的前提下加快训练速度，大量减少睡眠分期预测模型构建的时间。

本文第 2 节介绍支持向量机和 K 边界近邻法支持向量预选取算法，第 3 节构建睡眠分期预测模型，第 4 节是睡眠分期预测模型数值实验。

2. 支持向量机和 K 边界近邻法支持向量预选取算法

2.1. 支持向量机简介

支持向量机是在统计学习的背景下产生的一种机器学习方法，由 Vapnik [10]和 Cortest 提出，是一种结构风险最小化的理论，主要用来解决非线性回归和分类问题。

假设训练样本 $\{x_i, y_i\}, i=1,2,\cdots,N, N\in Z^+, x\in R^n, y_i\in\{-1,1\}$，$R^n$ 表示输入模式的特征空间。SVM 中训练的目的是在泛化误差达到最小（或有上界）的前提下，找到最优判决函数 $f(x)$，将两类数据分开。假定具有最大分类间隔的超平面为 $\omega\cdot x+b=0$，该超平面可以满足上述条件，它是由法向量 ω 和截距 b 决定。为了找到该分离超平面，我们需要求解以下凸二次规划问题[11]：

$$\begin{cases}\min & \omega(\alpha)=-\sum_{i=1}^{N}\alpha_i+\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j),\\ \text{s.t.} & 0\le\alpha_i\le C,\ i=1,2,\cdots,N,\\ & \sum_{i=1}^{N}y_i\alpha_i=0,\end{cases}\tag{1}$$

其中 $C\ge 0$ 是惩罚因子，α_i 是拉格朗日乘子，$K(x_i,x_j)=\phi(x_i)^{\mathrm{T}}\phi(x_j)$ 是核函数。

通过对式(1)求解可得 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$。法向量 $\omega^*=\sum_{i=1}^{N}\alpha_i^* y_i\phi(x_i)$，选择 α^* 的一个正分量 $0<\alpha_j^*<C$，可求得截距 $b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i K(x_i,x_j)$。引入核函数 $K(x_i,x_j)$，由凸二次规划问题得到最优判决函数为：

$$f(x)=sign\left(\sum_{i=1}^{N}y_i\alpha_i^* K(x,x_i)+b^*\right)。\tag{2}$$

其中 $sign(\cdot)$ 是一个符号函数，$\alpha_i^*\ne 0$ 对应的样本被称为支持向量。

SVM 是二分类模型分类器，但也可以做多分类。目前建立 SVM 多分类器的方法有很多，常见的是利用多个二分类器组合构造一个多分类器，可以用于多分类数据的处理[12]。已有诸多学者将 SVM 分类器应用于睡眠状态的识别，为了保证识别的准确度，睡眠分期预测模型的建立需要大量睡眠脑电信号样本参与 SVM 训练，这将导致训练时间过长。

2.2. K 边界近邻法支持向量预选取算法

K 边界近邻法是一种支持向量机预选取方法，SVM 的最优分离超平面由支持向量来确定，而支持向量只占全部训练集的一小部分[9]。因此，通过选取距离每个样本最近的 k 个异类样本作为支持向量机候选集，并将其作为训练样本进行 SVM 训练，能够在不影响支持向量机分类性能的前提下，大大减少训练样本，提高训练速度[13]。

已知训练样本分为两类，正类样本 T_1 和负类样本 T_2，

$$T_1=\{x_1^+,x_2^+,\cdots,x_{N_1}^+\},\quad T_2=\{x_1^-,x_2^-,\cdots,x_{N_2}^-\},$$

其中 $x_i^+,x_j^-\in R^M$，$i=1,2,\cdots,N_1$，$j=1,2,\cdots,N_2$，M 是样本维度。

已知有两类样本 $x_i^+, x_j^- \in R^M$，当它们线性可分时，它们之间的样本距离可以用欧氏距离来表示：

$$d\left(x_i^+, x_j^-\right) = \left\| x_i^+ - x_j^- \right\|_2 = \sqrt{\sum_{m=1}^{M}\left(x_{im}^+ - x_{jm}^-\right)^2}, \tag{3}$$

当两类样本非线性可分时，通过映射函数 $\varphi(\cdot)$ 将原输入空间映射到高维的特征空间中，样本在高维的特征空间中变得线性可分，这时的样本距离被称为非线性距离：

$$d\left(x_i^+, x_j^-\right) = \sqrt{K\left(x_i^+, x_i^+\right) - 2K\left(x_i^+, x_j^-\right) + K\left(x_j^-, x_j^-\right)}, \tag{4}$$

其中 $K(\cdot,\cdot)$ 是核函数。常用的核函数有线性核，多项式核、高斯核以及 S 型核。其中，高斯核函数因对于高维样本数据的分类问题也能有很好的识别效果而被频繁使用。高斯核函数如下：

$$K\left(x_i^+, x_j^-\right) = \exp\left(-\frac{\left\| x_i^+ - x_j^- \right\|^2}{2r^2}\right), \tag{5}$$

其中 r 是一个常数。

在某一类中，位于该类边界上的向量称为边界向量。已知有正类样本集 T_1 和负类样本集 T_2，$d\left(x_i^+, x_j^-\right)$ 为 T_1 中的第 i 个样本与 T_2 中的第 j 个样本之间的距离。对于 T_1 中的每一个 x_i^+，将 T_2 与 x_i^+ 距离最小的 k 个样本定义为 T_2 的边界向量；同理，可以定义 T_1 的边界向量。当 i 和 j 遍历所有值后就可以得到边界向量集。

KNBN 方法的具体步骤：

(1) 从正类样本集 T_1 中选择一个样本 x_i^+，求 x_i^+ 与所有负类样本之间的距离，保留最近的 k 个负类样本，将它们放入边界向量集当中。

(2) 返回步骤(1)，直至遍历所有正类样本为止。

(3) 将所有负类样本按照步骤(1)和步骤(2)操作，保留离每个负类样本最近的 k 个正类样本，将它们放入边界向量集当中。

(4) 把上面得到的边界向量集当中相同的支持向量删去，进行唯一化处理，最终得到支持向量候选集。

如图 1 所示，选取的样本都是离选定样本最近的另类样本，所以 KNBN 选出的向量一定为与两类样本的分布无关。因此，选取适当的 k 值，边界向量集一定能包含所有的支持向量，构造出一个支持向量候选集。

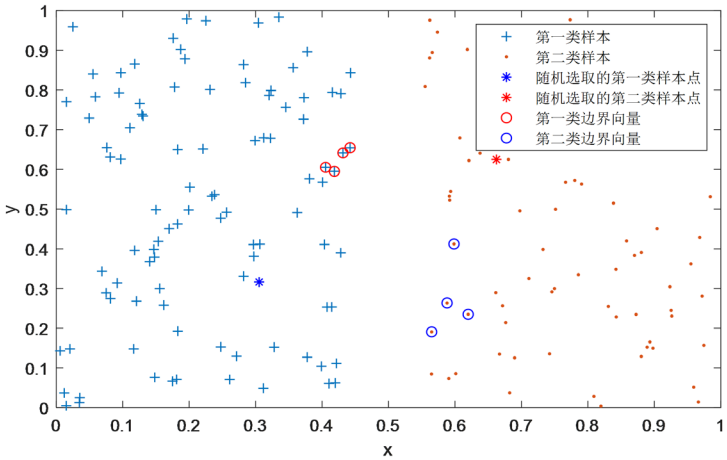


Figure 1. Diagram of KNBN support vector preselection
图 1. KNBN 支持向量预选取示意图

3. 睡眠分期预测模型

3.1. 睡眠分期基础理论

睡眠过程是动态并且复杂的。在国际睡眠分期的判读标准 R & K [14]中，将睡眠状态分为：清醒期，快速眼动期(REM)和非快速眼动期(NREM)；在非快速眼动期中，根据睡眠状态由浅入深的逐步变化，又进一步分为睡眠 I 期，睡眠 II 期，睡眠 III 期和睡眠 IV 期；其中睡眠 III 期和睡眠 IV 期又可合并为深睡眠期。睡眠分期如图 2 所示。

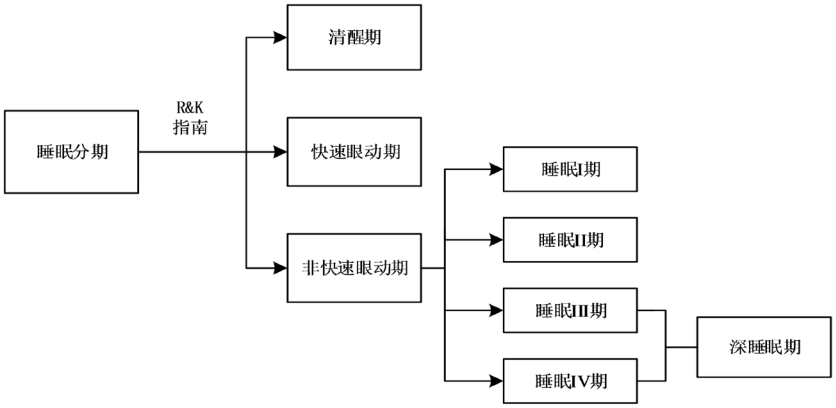
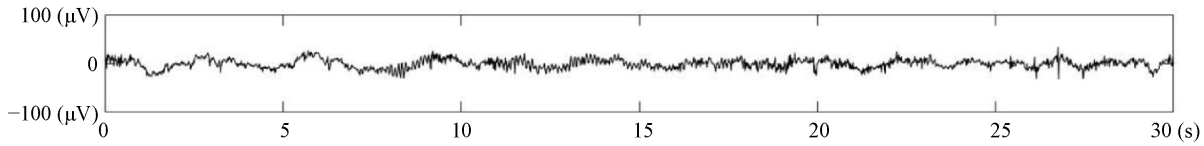
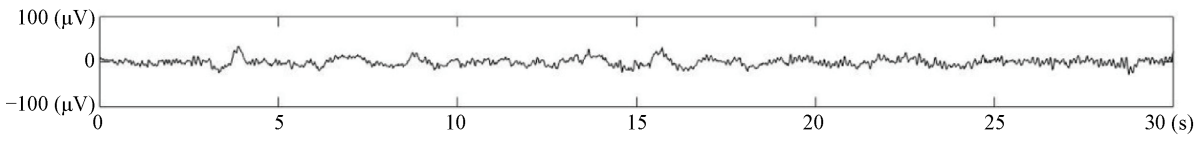


Figure 2. Diagram of sleep stages classification
图 2. 睡眠分期示意图

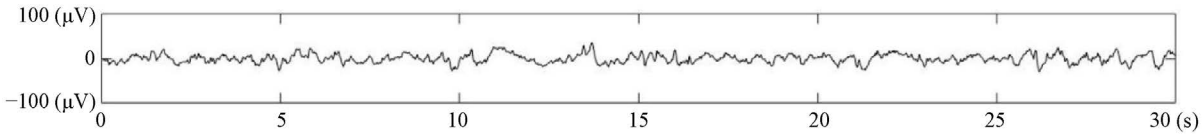
大脑的活动状态在各个睡眠时期具有显著差异，睡眠脑电信号是一种自发型的脑电信号，其中包含了睡眠时期大量的大脑信息。引言部分已经提到基于脑电信号对睡眠分期进行分析研究，是诊断和治疗睡眠疾病的重要辅助工具。而脑电信号作为一种生理信号，其具有随机性强，信号微弱，一般波幅范围在 0~200 μv，并且该信号极易受到噪声干扰。同时频率较低，频域特征比较突出，其范围一般为 0.5~100 Hz。此外，EEG 信号是一种典型的非线性信号，表现出较强的非平稳性[15]。



(a) 清醒期



(b) 睡眠 I 期



(c) 睡眠 II 期

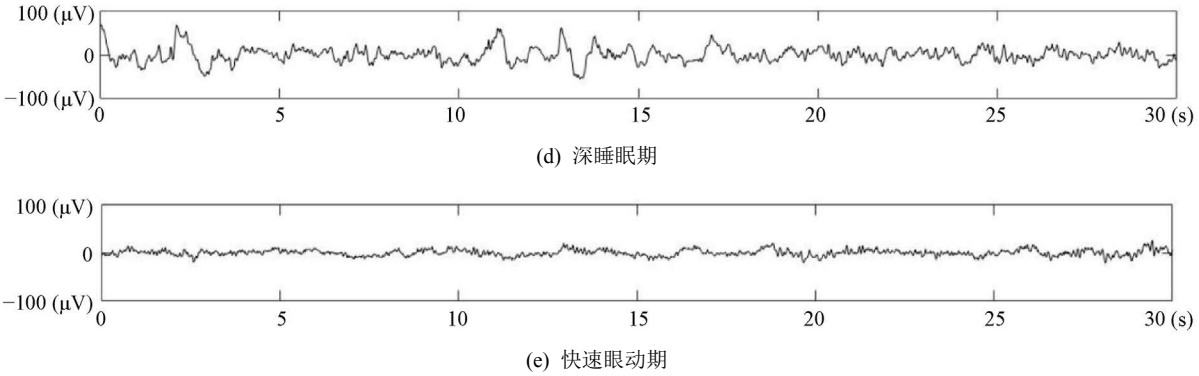


Figure 3. Time sequence of sleep EEG signals for each sleep stage
图 3. 各睡眠分期的睡眠脑电信号时序列

在睡眠过程中，主要经历了快速眼动期和非快速眼动期两个大时期，EEG 信号在不同睡眠分期所呈现的特点有所不同。如图 3 所示，自上而下依次为清醒期、睡眠 I 期、睡眠 II 期、深睡眠和快速眼动期的脑电信号时序列。

由于在不同睡眠期 EEG 信号有不同特征，因此我们可以利用其各个睡眠期 EEG 信号的特点，进行 SVM 分类。基于脑电信号的 SVM 睡眠分期模型，能够帮助更好的识别睡眠状态，有利于对相关睡眠疾病的防预和治疗。

3.2. 基于标准 SVM 的睡眠分期预测模型

机器学习中分类方法众多，SVM 是近年来使用较多的一种方法，尤其是 SVM 具有判别能力优良、泛化能力强、参数设置简单等优点。本文利用 SVM 的诸多优点与脑电信号时序列的特点，将 SVM 分类方法用于脑电信号数据的分类处理，进而以较高准确率识别睡眠状态。

本文基于标准 SVM 算法构建睡眠分期预测模型，下面是分期预测模型的步骤：

(1) 对 EEG 信号样本数据进行归一化处理；

(2) 归一化后的数据分成训练数据集和测试数据集；

(3) 将 EEG 信号样本中的训练数据集输入至 SVM 分类器当中，构建基于标准 SVM 的睡眠分期预测模型；

(4) 测试数据集代入步骤(3)中的睡眠分期预测模型进行预测，得到预测结果。

3.3. 基于 KNBN-SVM 的睡眠分期预测模型

基于标准 SVM 的睡眠分期预测模型能达到较为理想的预测效果，但在实验测试所采集到的 EEG 信号样本中，包含了一些无用的或是冗余样本。换言之，训练样本中包含了支持向量和非支持向量，因此 SVM 的训练中还包括了对非支持向量的优化，需要花费大量的训练时间。

本文使用 K 边界近邻法支持向量预选取方法(KNBN)对 EEG 信号训练样本进行支持向量预选取。通过 KNBN 方法可以减少 EEG 信号训练样本中的冗杂样本和对分类无用的多余样本，再用 SVM 来训练和预测，从而减少预测模型建立的训练时间并在一定程度上提高分类预测准确度。

下面为基于 KNBN-SVM 的睡眠分期预测算法步骤：

(1) 将 EEG 信号样本数据进行归一化处理，并分成训练样本和测试样本；

(2) 将训练数据集按睡眠分期标签分为五类，随机选取其中的某一类作为正类，其余四类作为负类，按照 2.2 中 KNBN 方法在两类数据集中预选取出边界向量；

(3) 当五类样本中的每一类样本都经历过作为正样本之后，得到一个原始训练数据集的边界向量集，将其中相同的样本唯一化处理，得到支持向量候选集；

(4) 将步骤(3)中得到的支持向量候选集作为训练数据集输入到 SVM 模型中，得到一个五分类的睡眠分期预测模型；

(5) 将测试数据集输入到步骤(4)中得到的睡眠分期预测模型中得到分类准确度及训练时长。

图 4 是基于 KNBN-SVM 的睡眠分期预测模型流程图：

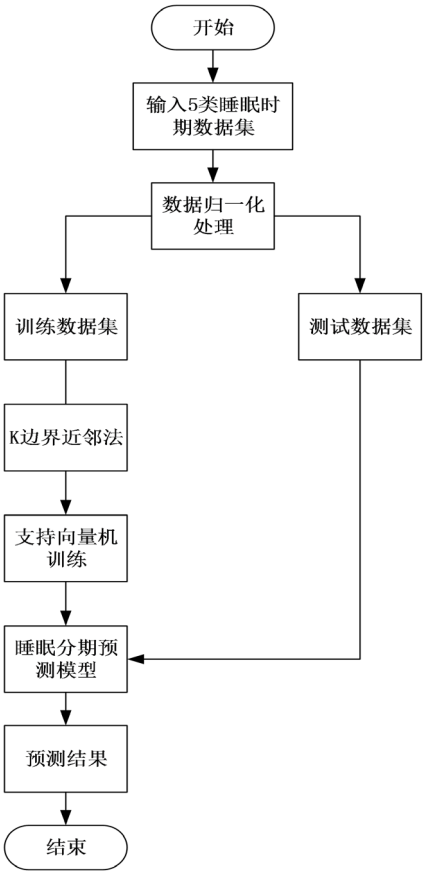


Figure 4. Flow chart of sleep staging prediction model based on KNBN-SVM
图 4. 基于 KNBN-SVM 的睡眠分期预测模型流程图

4. 睡眠分期预测模型数值实验

下面为了讨论 KNBN-SVM 算法对于睡眠分期预测的有效性和可行性，将 KNBN-SVM 算法和标准的 SVM 算法下构建的睡眠分期预测模型的实验结果进行分析和比较。本文是在 Matlab R2018a 环境中，2.3 GHz，Pentium，DualCPU，4 GB 内存的硬件平台上进行。

本文睡眠分期预测实验数据采用“‘华为杯’第十七届中国研究生数学建模竞赛”赛题 C 题的睡眠脑电信号数据。该数据中共包含五类睡眠状态数据集，分别为：清醒期、快速眼动期、睡眠 I 期、睡眠 II 期和深睡眠期，共有 3633 个样本，随机选取 3600 个 EEG 信号样本参与实验，其中 2000 个样本为训练样本，1600 个样本为测试样本。按照 3.2 中的流程构建基于标准 SVM 的睡眠分期预测模型，并代入

EEG 信号测试数据集得到基于标准 SVM 的睡眠分期预测模型的预测准确度以及训练耗时。对于KNBN-SVM 方法，将实验数据集按照 3.3 中的流程，先使用 KNBN 方法在原始的 EEG 训练样本集中预选取支持向量，再将支持向量候选集作为新的训练样本代入 SVM 中训练，构建基于 KNBN-SVM 的睡眠分期预测模型，最后代入 EEG 信号测试样本得到 KNBN-SVM 训练算法的预测准确度和训练总耗时，其中 KNBN-SVM 算法的训练时长为预选取边界向量集的时间和 SVM 训练时间的总和，KNBN 的参数值[9] $k=4$。SVM 和 KNBN-SVM 中数据训练采用的核函数均选为高斯核函数，其中 $r=1.3$。实验结果如表 1 所示，实验结果为 100 次独立实验结果的平均值。

从表 1 中可以看出，直接使用标准的 SVM 训练 2000 个训练样本，得到的分类预测准确率为 73.75%，训练耗时 864 ms；KNBN-SVM 方法中对训练数据进行 KNBN 支持向量预选取选取出 1205 个边界向量作为支持向量候选集。将支持向量候选集作为训练样本集输入到标准 SVM 当中训练，得到的分类预测准确率为 75.31%，训练总耗时 380 ms。实验结果表明，KNBN-SVM 训练算法的训练样本比标准的 SVM 算法的训练样本减少了 39.75%，训练总耗时减少了 484 ms，预测准确度提升了 1.56%。

Table 1. Comparison of sleep stages classification of SVM and KNBN-SVM
表 1. 标准 SVM 和 KNBN-SVM 的睡眠分期预测模型比较

训练算法	训练样本/个	边界向量数/个	测试样本量/个	训练时间/ms	预测准确度/%
标准 SVM	2000	无	1600	864	73.75%
KNBN-SVM	2000	1205	1600	380	75.31%

5. 结语

本文将支持向量机分类应用于睡眠分期预测模型的研究，使用脑电信号时序列作为睡眠状态识别的工具，构建基于标准 SVM 的睡眠分期预测模型和基于 KNBN-SVM 的睡眠分期预测模型，并进行比较。实验表明，基于标准的 SVM 的睡眠分期预测模型预测准确度较理想；基于 KNBN-SVM 的睡眠分期预测模型不仅相对于标准的 SVM 准确度有所提高，而且还大幅度的缩短了训练时间。因此，KNBN-SVM 方法适用于睡眠分期的研究，对于预防和治疗相关的疾病有着一定的应用价值。

参考文献

[1] 张泾周, 周钊, 滕炯华, 苗治平. 基于神经网络的睡眠分期处理算法研究[J]. 计算机仿真, 2010, 27(8): 141-144.

[2] 李晓博. 脑电信号的睡眠分期方法研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2019.

[3] 李谷, 范影乐, 李轶, 庞全. 基于脑电信号 Hilbert-Huang 变换的睡眠分期研究[J]. 航天医学与医学工程, 2007, 20(6): 458-463.

[4] 刘慧, 谢洪波, 和卫星, 王志中. 基于模糊熵的脑电睡眠分期特征提取与分类[J]. 数据采集与处理, 2010, 25(4): 484-489.

[5] Liang, S.F., Kuo, C.E., Hu, Y.H., *et al*. (2012) Automatic Stage Scoring of Single-Charnel Sleep EEG by Using Multiscale Entropy and Autoregressive Models. *IEEE Transactions on Instrumentation and Measurement*, **61**, 1649-1657. https://doi.org/10.1109/TIM.2012.2187242

[6] 周鹏, 李向新, 张翼, 明东, 董新明, 薛然婷, 王学民. 基于主成分分析和支持向量机的睡眠分期研究[J]. 生物医学工程学杂志, 2013, 30(6): 1176-1179.

[7] Mohammed, D., Li, Y. and Wen, P. (2016) EEG Sleep Stages Classification Based on Time Domain Features and Structural Graph Similarity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **24**, 1159-1168. https://doi.org/10.1109/TNSRE.2016.2552539

[8] 张晓宇. 基于脑电图的睡眠自动分期和特征分析[D]: [硕士学位论文]. 南京: 南京大学, 2016.

[9] 李庆, 胡捍英. 支持向量预选取的 K 边界近邻法[J]. 电路与系统学报, 2013, 18(2): 91-96.

[10] Cortes, C. and Vapinik, V.N. (1995) Support-Vector Network. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/BF00994018

[11] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 第七章, 95-135.

[12] 沈洋. 支持向量机多分类器的研究与应用[D]: [硕士学位论文]. 无锡: 江南大学, 2019.

[13] 韩成志, 郑恩涛, 马国春. 基于距离配对排序的支持向量预选取算法[J]. 应用数学进展, 2020, 9(2): 195-203.

[14] Vernon, M.K., Dugar, A., Revicki, D., *et al*. (2009) Measurement of Non-Restorative Sleep in Insomnia: A Review of the Literature. *Sleep Medicine Reviews*, **14**, 205-212. https://doi.org/10.1016/j.smrv.2009.10.002

[15] 李俊雨. 基于 EEG 信号时－频分析的深度睡眠过程控制机理研究[D]: [硕士学位论文]. 西安: 陕西科技大学, 2018.