第47卷第1期 中国科学技术大学学报 Vol. 47, No. 1
2017年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2017

文章编号:0253-2778(2017)1-0001-09

支持向量机在高考成绩预测分析中的应用

张 莉,卢星凝,陆从林,王邦军,李凡长

(苏州大学计算机科学与技术学院,江苏苏州 215006)

摘要:支持向量机作为一种机器学习算法因其良好的推广性和强大的非线性处理能力而令人瞩目.为此将支持向量机与国家高考的实际数据相结合,以具体高校的高考模拟考试成绩为主要训练数据,进行学生的高考成绩预测.实验考虑了三种情形.一是通过六次模拟考试的特征分来预测高考的特征分;二是通过六次模拟考试和高考的特征分来预测高考的录取批次;三是通过六次模拟考试的特征分和高考的预测特征分来预测高考的录取批次.通过与神经网络算法的比较,实验结果均表明了支持向量机方法的稳定性和良好的预测性.

关键词:支持向量机;高考;预测;神经网络;机器学习

中图分类号:TP 391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.01.001

引用格式:张莉,卢星凝,陆从林,等. 支持向量机在高考成绩预测分析中的应用[J]. 中国科学技术大学学报,2017,47(1):1-9.

ZHANG Li, LU Xingning, LU Conglin, et al. National matriculation test prediction based on support vector machines[J]. Journal of University of Science and Technology of China, 2017,47(1):1-9.

National matriculation test prediction based on support vector machines

ZHANG Li, LU Xingning, LU Conglin, WANG bangjun, LI Fanzhang

(*School of Computer Science and Technology, Soochow University, Suzhou* 215006, *China*)

Abstract: Support vector machine(SVM), one of machine learning methods, is very impressive for its good generalization and powerful nonlinearly processing ability. SVM was combined with national matriculation, where scores of six mock exams are taken as training data to predict the final admission scores. Three situations were considered. First, the scores of NMT were predicted using scores in six simulation tests. Second, the admission batch was predicted by using scores in six simulation tests and NMT. Third, the admission batch was predicted by using scores in six simulation tests and the estimated scores in NMT. In all experiments, SVMs were compared with neural networks (NNs). Experimental results show that SVMs are much more stable and have better prediction ability.

Key words: support vector machine; national matriculation test; prediction; neural network; machine learning

收稿日期:2016-03-01;**修回日期**:2017-09-17

基金项目:国家自然科学基金(61373093,61672364),江苏省自然科学基金(BK20140008),江苏省高校自然科学研究项目(13KJA520001),江苏省青蓝工程资助.

作者简介:张莉(通讯作者),女,1975年生,博士/教授.研究方向:机器学习、数据挖掘. E-mail: zhangliml@suda.edu.cn

0 引言

尽管应试教育下的高考存在诸多弊端,但不可否认的是高考提供了一个较为公平的竞争平台,为社会基层的百姓提供了一条向高层流动的路径.高考是由国家教育部统一组织的,国家或者部分省份统一命题的,社会关注度极高的一次考试,所以每个城市、每个学校都会在高考之前进行多次模拟考试,以希望达到一些目的.比如,可以帮助学生发现自己在知识点、心理上、答题策略上存在的不足,及时调整学习的重点方向;也可以帮助教师及学校及时调整教学计划、合理安排教学进度、关注重点学生,以让尽可能多的学生充分发挥出自己的潜能.

我们注意到每年高考结束以后,有或多或少这样的报道:就读于某中学的某某学生,在得知自己高考成绩后,从教学楼跳下去世……这些新闻报道,除了让我们深感惋惜之外,也让我们开始关注高考成绩的预测问题.如果让平时的模拟考试不仅仅在内容上给学生、学校一些及时的反馈与帮助,也能在高考成绩的心理预期上给考生及其家长带来更准确的预估,那么也可以在考生对自己的人生规划上提供更早、更多、更科学的帮助,从而减少悲剧的发生.

无论是在国内还是在国外,都有很多关于高考的研究,国外的相关研究主要是关于学术能力测试[1-2]以及托福考试[3-5]等方面,已取得的研究成果证实预测是可能的,并且具有广阔的应用前景.本文主要关注国内的状况,不对国外的研究展开讨论.下面就中国大陆在高考预测方面的工作进行讨论.

利用计算机预测高考成绩比较早的是1985年的李以渝[6],在他之前的研究仅局限在对高考成绩的统计工作[7-8].文献[6]的不同之处是开始借助机器学习的方法来进行高考成绩的分布预测,虽然高考成绩的分布预测与高考成绩预测是有所区别的,但不可否认的是机器学习方法已经被应用在了高考预测中.

卫子光认为,人才的预测与高考也有直接的关系[9].他们从化学高考的层面,进行了关于化学高考成绩预测效度的初步分析.对于效度的理解,文献[10]给出的定义是:衡量考试的正确性、有效性的指标,考试的正确性与有效性是衡量考试科学性的主要指标之一.文献[10]还表明,一般关于高考效度的研究分为两类:一类是衡量高考选才是否是有效;另一类研究学生在高考中所取得的高考成绩与大学期间的学业成绩之间的关联程度有多大.在高考效度研究中,必不可少的组成包括:教育的公平性、教育的成本以及基础教育等方面.另外,在英语[11-12]和数学[13]学科上,关于高考预测效度的分析也有所体现.

本文主要讨论模拟考试与高考成绩之间是否有强关联.2004年,亓鲁霞[14]指出,高考与模拟考试存在一定的区别.区别在于模拟考试主要是为了模拟高考,帮助学生和教师练习与平时的反馈;而高考的主要功能却是为了帮助高等教育学校选拔合格的、优秀的高中毕业生.研究表明,虽然模拟考试与高考有区别,但还是可以依据考生模拟考试的成绩来比较准确地预测出高考的成绩,模拟考试确实具备预测功能.在英语教育方面,Heaton指出,模拟考试可以帮助教师和考生准确找到班级及考生个人存在的薄弱知识点,进而有针对性地去解决这些薄弱环节[15].他还指出模拟考试可以让教师知道考生们认为课程的哪些部分是难点,教师也能够通过模拟考试思考教学大纲和正在运用的教学材料与教学方法是否有效,让考生和教师产生理性的成绩预测;因此,虽然模拟考试与高考存在着一定的差异,但模拟考试的成绩与高考的成绩高度相关,可以借助于模拟考试的成绩来预测高考成绩.

文献[16]对高考预测系统的设计是基于Web挖掘的方法.该方法先挖掘出Web页面上高考招生的数据,然后利用回归分析来预测高考分数线.该系统的分析对象是山东省的考生.吴金财曾针对2010年和2011年深圳市考生的模拟考试的成绩,提出了“均分相对值”和“有效平均分”的概念,探讨对考生评价和分析时如何使用平均分[17].李敬文等将灰色系统理论的知识和模糊数学理论的知识相结合,借助基于模糊灰色理论的知识预测大学各专业的上线分数线[18],成功建立了一种关于高考数据预测的模型.

在预测过程中,大部分方法均采用了数理统计中的线性回归分析,这自然也是机器学习处理预测问题的一种手段.机器学习方法除了线性回归分析之外,还有其他很多方法,如神经网络(neural network)、决策树(decision tree, DT)、支持向量机(support vector machine)、贝叶斯学习(Bayesian learning)等[19-23].文献[24]借助“学生成绩管理系统”这一研究背景,对决策树算法进行了改进,建立了一种根据学生成绩数据库信息的成绩预测模型,

综合分析预测学生的高考成绩. 预测的目标是上线情况,也就是录取批次. 通过实验分析,得到预测上线率的准确率在 80%以上,能基本实现预测的目的. 张琼借助 Clementine 软件,利用贝叶斯网络分类器,对高考的录取批次[25]进行了预测. 实验数据是来源于某高中的学生三年的成绩及基本信息,包括他们的高考成绩.

在统计学习理论(statistical leaning theory)的基础上发展出的支持向量机(support vector machine)理论[26-27],被广泛地认为是一种普适且非常有效的学习方法. 在结构风险方面,支持向量机可以很好地实现最小化原则. 结构风险最小化这一原则的实现,使得支持向量机在处理小样本方面具备更大的推广优势. 因支持向量机良好的推广性和实用性,研究人员将注意力放在了其算法的进一步改进提升上. Zhang 等采用 Mercer 正定核提出隐空间支持向量机(hidden space support vector machine),能够将任意非线性函数实现非线性映射[28]. 有学者构建了比支持向量机更加稀疏的学习机,1-范数支持向量机和稀疏支持向量机[29-30]. 稀疏支持向量机已经从理论上被证明比支持向量机具有更好的稀疏性[31].

在有关分类问题[32],回归估计问题[33],新颖检测问题[34],聚类问题[35]以及半监督学习问题[36]等方面,支持向量机的理论有针对性地得到发展,能很好地解决问题. 有关支持向量机及其变形方法的应用非常多,不一一列举.

鉴于支持向量机在分类和回归估计等问题中的不错表现,本文选择将支持向量机方法作为对高考成绩预测的工具. 实验中对将收集到的普通高校学生的六次模拟考试的特征分作为主要训练数据,对高考的特征分以及录取批次进行预测. 通过与神经网络算法的比较,来验证支持向量机在国家高考成绩预测的可行性.

1　支持向量机

在处理最基本的回归估计问题、密度估计问题、分类问题等方面,SVM 都有很好的应用. 分类问题和回归估计问题和本文的研究相关,下面我们分别介绍 SVM 处理分类和回归估计的算法.

1.1　支持向量机分类

1995 年,Cortes 等首先提出了支持向量机[27]. 在统计学的 VC 维理论以及结构风险最小概念基础上,SVM 将向量映射到高维空间,并在该空间建立分隔超平面将数据分开,并使得这个超平面具有最大的边缘(margin).

在分类问题中,假设有训练数据集 $\{(x_i,y_i)\}_{i=1}^{n}$,其中 $x_i \in \mathbb{R}^m$ 表示 m 维数据特征,$y_i \in \{-1,+1\}$ 代表数据类别,n 是数据个数. SVM 通过最小化下面的优化问题来解决两类分类问题.

$$\begin{aligned}&\min_{w,b,\xi}\ \frac{1}{2}\ \|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i\\ &\text{s.t.}\quad y_i(w\cdot x_i+b)\geqslant 1-\xi_i,\\ &\qquad \xi_i\geqslant 0,\ i=1,\cdots,n\end{aligned}\tag{1}$$

式中,$\boldsymbol{w}\in\mathbb{R}^m$ 是超平面的法向量,b 是阈值,ξ_i 是引入的松弛变量,$C>0$ 是惩罚因子. 采用拉格朗日乘子方法和核技巧,求得优化问题(1)的对偶规划:

$$\begin{aligned}&\max\ \sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\\ &\text{s.t.}\ \ \sum_{i=1}^{n}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C,i=1,\cdots,n\end{aligned}\tag{2}$$

式中,α_i 是拉格朗日乘子,$K(x_i,x_j)$ 是满足 Mercer 条件的核函数. SVM 的决策函数可以表示为

$$f(x)=\operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b\Big)\tag{3}$$

式中,sgn 表示二值函数,其值为 $\{-1,+1\}$.

1.2　支持向量机回归

假定,我们已知一组独立同分布的训练样本集 $X=\{(x_i,y_i)\mid x_i\in\mathbb{R}^d,y_i\in\mathbb{R},i=1,\cdots,l\}$,$x_i$ 表示训练样本,其对应的响应或者输出为 y_i,l 表示总的样本个数,d 表示样本的维数(特征)个数. 在分类问题中,y_i 取的是离散的-1 或者$+1$;在回归估计问题中,y_i 可以在整个实数集上取值.

相对于支持向量机分类算法,支持向量机回归估计算法也需要控制函数的复杂度以及最小化经验风险,因而对任意的损失函数,支持向量机回归估计可被表示为如下的优化问题[33]:

$$\begin{aligned}&\min\ \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}[c(\xi_i)+c(\xi_i^*)]\\ &\text{subject to}\ \boldsymbol{w}^{\mathrm{T}}x_i+b-y_i\leqslant\varepsilon+\xi_i\\ &\qquad y_i-\boldsymbol{w}^{\mathrm{T}}x_i-b\leqslant\varepsilon+\xi_i^*\\ &\qquad \xi_i\geqslant 0,\xi_i^*\geqslant 0,\ i=1,\cdots,l\end{aligned}\tag{4}$$

式中,$\varepsilon\geqslant 0$ 是预定义的可允许误差,即函数输出与期望输出 y 之差可以小于等于 ε;$C>0$ 是正则因

子，用来平衡函数复杂度和经验风险之间的关系；$c(\cdot)$ 表示损失函数，估算经验风险或者损失的多少.

类似于支持向量机分类算法，这里也采用拉格朗日乘子技术得到原规划(4)式的 Wolfe 对偶规划，我们再引入核技巧，求解得到最终的回归估计函数：

$$f(x)=\sum_{x_i\in SV}(\alpha_i^*-\alpha_i)K(x_i,x)+b \qquad (5)$$

阈值 b 的计算同样需利用 KKT 条件和边界支持向量. 对所有的边界支持向量 x_i，我们进行如下计算：

$$b=y_i-\sum_{x_j\in SV}(\alpha_j^*-\alpha_j)K(x_j,x_i)+\varepsilon \qquad (6)$$

最后对阈值取平均得到最终结果.

2 基于 SVM 的高考成绩预测

本文实验全部在 Window 7 环境中实现，以 Matlab 为主要编程工具. SVM 的训练和测试利用 Chang 等公布的 LibSVM [38] 实现.

2.1 数据预处理

江苏省高校招生录取依据的是考生语数外三门总分的成绩(特征分)和两门选修的等级. 其中特征分总分是 480 分，根据分数的高低一般分三个等级：本一、本二和专科. 两门选修的等级一般划分为 A^+、A、B、C、D. 关于特征分，本文主要关注特征分所处的录取批次，两门选修的等级我们只关注是否达到 B 等级，对个别高校要求更高等级和稍低等级的我们不作考虑，只关注一般性的要求.

实验数据收集于江苏省海门市四甲中学参加完整考试的 538 名考生. 作为地级市三星级重点中学，该校的生源属于中等水平，且其多年的高考录取成绩一直相对比较稳定，在高考录取中各批次的分布是比较均匀，各批次的录取比例略高于江苏省的全省录取比例，具有较强的代表性. 本文的实验框架如图 1 所示.

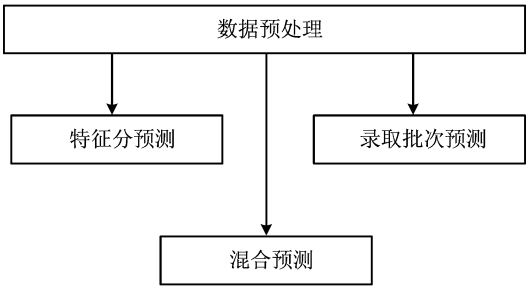


图 1 实验框架图

Fig. 1 Experiment framework

考生成绩数据的表现形式有两种：数值型和枚举型. 特征分是数值型变量，其取值是实数；等第是枚举型变量，其取值范围为{本一，本二，专科}；选一等第和选二等第也是枚举型变量，取值范围为{A^+，A，B，C，D}. 由于这里涉及的枚举型变量均是字母或者文字，不便于用计算机操作，而且支持向量机所处里的数据必须是数值型的，因而我们简单地把枚举型变量数值化，如表 1 所示. 等第和选一/二等第是属于不同的特征维，因而数值的重复并不会产生影响. 此外由于 A^+ 和 A 具有相似的结果，因而把它们都数值化为 1. 表 2 是数值化的 9 位考生的信息，表中列出了所有的 28 个特征维数及其编号，由于我们只预测特征分和录取批次(等第)，因而最后 2 维特征即 27 和 28，在实验中没有用到. 由于江苏省高考文科和理科总分一致，语文数学的基础分和附加分比例一致，等级考试按照学生成绩比例划分，所以将本文的高考数据文理科按照相同情况处理.

2.2 高考特征分预测

对高考进行特征分预测时，首先需要组成训练数据集 $X=\{(x_i,y_i)\mid x_i\in\mathbb{R}^d,y_i\in\mathbb{R},i=1,\cdots,l\}$. 取第 1 维到第 24 维特征作为样本的特征，即 $x_i\in\mathbb{R}^{24}$，y_i 对应于第 25 维特征.

本文对于所有的 538 个数据进行处理. 为了保证能较好地训练，取前 500 个作为训练样本，即 $l=500$；剩下的 38 个作为测试数据，用来测试所得预测模型. 此外，我们对训练样本进行了归一化处理，即把输入值的范围控制在 0 到 1 之间. 需要注意的是，测试样本的归一化是在训练样本基础上进行的，而测试样本的输入却有可能在[0,1]区间之外.

实验采用平均相对方差(average relative variance，ARV)作为指标来衡量模型的性能，定义如下：

$$\mathrm{ARV}=\frac{\sum_{t=1}^{T}(y_t-f(x_t))^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2} \qquad (7)$$

式中，x_t 是用于测试的样本，y_t 是对应的真实输出，$f(x_t)$ 是 y_t 的预测值，$\bar{y}=\frac{1}{T}\sum_{t=1}^{T}y_t$，$T$ 是测试样本的个数.

表 1　枚举型变量特征处理

Tab. 1　Feature values of enumerated variables

原始	本一	本一	专科	A^+	A	B	C	D
数值	1	2	3	1	1	2	3	4

表 2　数据预处理后 9 位考生的成绩特征

Tab. 2　Nine samples after data preprocessing

特征	学生								
	1	2	3	4	5	6	7	8	9
1	301	250	262	207	245	268	221	271	331
2	2	3	3	3	3	3	3	3	2
3	1	2	1	4	2	2	2	2	2
4	1	3	1	4	3	2	2	2	2
5	336	305	288	249	289	302	270	318	328
6	2	3	3	3	3	2	3	2	2
7	2	2	1	4	4	1	3	2	2
8	1	3	1	4	2	2	2	2	2
9	272	253	269	240	151	282	247	296	312
10	3	3	3	3	3	2	3	2	2
11	2	2	1	4	4	1	3	2	2
12	2	4	1	3	4	2	4	2	2
13	294	291	307	241	245	271	206	301	334
14	3	3	2	3	3	3	3	3	2
15	1	3	2	4	4	2	3	2	2
16	2	3	1	4	3	1	3	2	2
17	293.5	244	286	246.5	246.5	273.5	260	294	290
18	2	3	2	3	3	2	3	2	2
19	1	2	1	3	3	1	3	2	1
20	1	2	1	3	3	1	3	3	2
21	323	291	304	240	266	300	238	294	318
22	2	3	2	3	3	2	3	3	2
23	1	2	1	4	4	1	3	4	2
24	1	2	1	4	4	1	3	2	2
25	316	306	335	267	276	311	274	314	324
26	2	3	1	3	3	2	3	2	2
27	1	2	1	3	3	1	3	2	2
28	1	2	1	3	3	1	3	2	2

支持向量机涉及两个参数:正则因子 C 以及核参数. 这里采用拟合能力较好的高斯核,因而我们要确定 γ 的取值. 一般来说,首先给定参数的一定取值范围,然后用交叉验证的方法来从中确定较好的参数. 本实验采用 10 倍交叉验证,把已知训练集再进行 10 等分,取其中的 9 份来训练模型,1 份来验证. 用 10 次实验的平均结果来选取最优参数,其中 C 的取值范围为 $\{0.1,1,10,100\}$, γ 的取值范围为 $\{2^{-8},2^{-7},\cdots,2^{8}\}$. 支持向量机回归估计得到的 ARV 曲线图见如图 2 所示. 对不同的 C,每条曲线均会有一个最小值,如 $C=10$ 时,ARV 曲线的最小值在 $\gamma=2^{-1}$ 处,最小值为 0.373 8. 在 $C=100$ 时,最小的 ARV 值为 0.349 2,此值也是在给定参数范围内的最好性能. 由于 $C=100$ 时的曲线不够平滑,可能存在过拟合,因此关于支持向量机回归估计的参数,我们选择能使得 ARV 曲线较为平滑,且值较低的,即 $C=10$. 此外,核参数确定为 $\gamma=2^{-1}$.

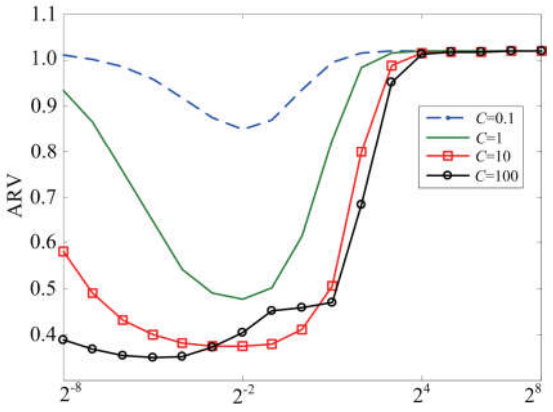


图 2 支持向量机特征分预测的 10 倍交叉验证结果

Fig. 2 Ten-fold cross validation prediction results of NMT score obtained by SVM

在神经网络中的可调参数比较多,但是最影响性能的应该是隐藏层节点的个数 n_{hidden}. 同样,采用 10 倍交叉验证,其中 n_{hidden} 取值于给定的范围集合:$\{5,10,15,20,25,30,35,40\}$. 此外,由于神经网络对权值的初始值敏感,所以对每个节点数,我们都运行 10 次,以 10 次平均来代表该节点的性能. 神经网络得到的 ARV 曲线图见如图 3 所示. 由图 3 可以看到,在 $n_{\text{hidden}}=10$ 时,ARV 值是最小的,因而隐藏层节点数确定为 10. 其他的网络参数,采用默认值.

在确定好方法的参数后,把选好的参数用来训练各自对应的方法,然后用测试样本集来测试. 最后的结果见表 3 所示,其中神经网络的结果是 10 次平均的结果,所以会有标准差. 从表中的数据可以看出,支持向量机比神经网络具有更好的平均相对方差,即 ARV. 理论上 ARV 越接近 0,表明预测的准确性越高;反之接近 1,则说明预测的效果不好. 当然如果 ARV 大于 1,则说明预测没有任何的意义.

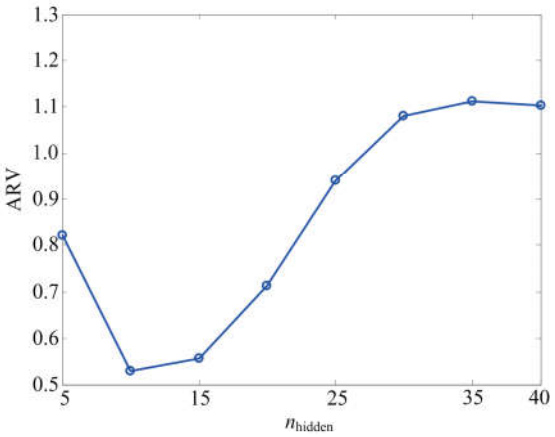
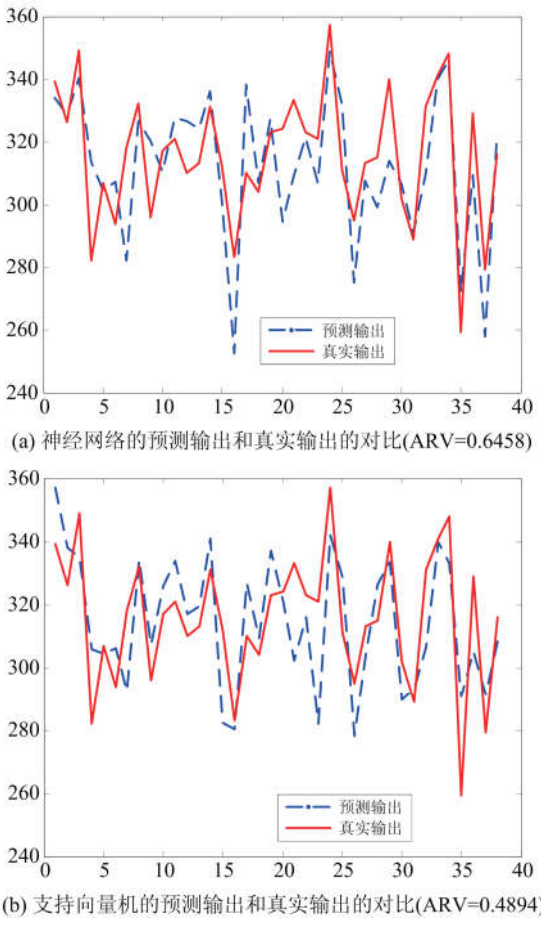


图 3 神经网络特征分预测的 10 倍交叉验证结果

Fig. 3 Ten-fold cross validation prediction results of NMT score obtained by NN



(a) 神经网络的预测输出和真实输出的对比(ARV=0.6458)

(b) 支持向量机的预测输出和真实输出的对比(ARV=0.4894)

图 4 两种方法在测试集上的预测效果图

Fig. 4 Prediction results obtained by two methods on the test set

这两种方法各自的预测效果如图 4 所示. 图中的实线表示真实的输出,而虚线则表示预测的输出. 图 4(a)是神经网络的结果,图 4(b)是支持向量机的结果. 由于神经网络的随机性,图 4(a)展示的是该方法获得较好性能时的预测结果. 由图 4 可以看出,神经网络的拟合幅度较大,在曲线的 V 形处,常常超出真实值. 相对而言,支持向量机的预测相对保守一些.

表 3　两种方法在特征分预测上的性能对比

Tab. 3　Performance comparison of two methods on NMT score prediction

	支持向量机	神经网络
平均相对方差	0.4894	0.7026 ±0.0846

2.3　高考录取批次预测

在上一节中,我们实现了支持向量机对特征分的预测,并和神经网络对比,得到了不错的结果. 下面对高考的录取批次进行分类预测,即按照他们考取的学校来进行分类,分为一本、二本和专科三类.

实验中,假定录取批次的预测是在得知了特征分之后进行的. 因此,我们取了第 1 维到第 25 维特征作为样本的特征,即 $x_i \in \mathbb{R}^{25}$, y_i 对应于第 26 维特征. 和 2.2 节的实验一样,我们也取前 500 个作为训练样本,剩下的 38 个作为测试数据. 也对数据进行了归一化处理. 此时的训练集合表示为

$$X = (x_i, y_i) \mid x_i \in \mathbb{R}^{25}, y_i \in \{1,2,3\},$$
$$i = 1, \cdots, 500.$$

集合 {1,2,3} 代表了集合{本一, 本二, 专科}. 这是一个多分类问题,我们采用 one-against-one 方法来处理支持向量机多分类问题. 分类问题的性能衡量指标为分类的精度(accuracy, AC),即

$$\mathrm{AC} = \frac{\text{正确识别的样本个数}}{T} \tag{8}$$

式中, T 是测试样本的总个数.

我们同样采用 10 倍交叉验证的方法设置算法参数,结果如图 5 所示. $C = 1$, $C = 10$ 和 $C = 100$ 的曲线比较相似,也比较平滑. 在这三者中取最好的,即 $C = 100$ 时的曲线. 在此曲线上,当 $\gamma = 2^{-5}$ 时支持向量机的 10 倍交叉验证精度为87.22%,因此我们确定好支持向量机多分类算法的参数为 $C = 100$,且 $\gamma = 2^{-5}$.

神经网络的隐藏层节点 n_{hidden} 同样在给定范围集合中取值,对每个节点数,我们都运行 10 次,以 10 次平均来代表该节点的性能. 神经网络得到的 10

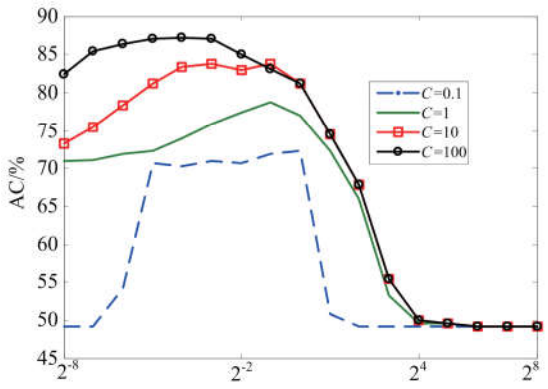


图 5　支持向量机录取批次预测的 10 倍交叉验证结果

Fig. 5　Ten-fold cross validation prediction results of admission bath obtained by SVM

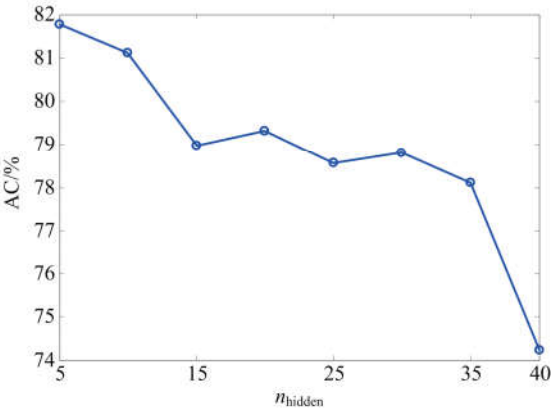


图 6　神经网络录取批次预测的 10 倍交叉验证结果

Fig. 6　Ten-fold cross validation prediction results of admission bath obtained by NN

倍交叉验证曲线图如 6 所示. 当 $n_{\text{hidden}} = 5$ 时,精度为 81.78%是最高的. 因而隐藏层节点数确定为 5,其他的网络参数采用默认值.

在确定好方法的参数后,先用选好的参数来训练各自对应的方法;然后用测试样本集来测试. 最后的结果如表 4 所示,其中神经网络的结果是 10 次平均的结果,带有标准差. 从表中的数据可以看出,神经网络不稳定,其标准差为 8.12%,这意味着神经网络也有与支持向量机差不多性能的情况,但是大部分情况下,性能较弱.

表 4　两种方法在录取批次预测上的性能对比

Tab. 4　Performance comparison of two methods on admission batch prediction

参数	支持向量机	神经网络
精度/%	86.84	78.42 ±8.12

支持向量机的分类结果为 86.84%,也就是说在测试的 38 人中,只有 5 人的录取批次弄错了. 表

5列出了分类错误的5个数据的部分特征信息，包括6次模拟考试的录取批次(第2、6、10、14、18以及22维)和高考的录取批次(第26维).可以看出这几个考生不太稳定，比如529号.平时的模拟考试录取批次为二本或专科，预测出来的结果是二本，但出人意料的是，最后他/她上的是一本.其他的错分样本，也是属于较难预测的.

表5 错分数据的部分特征信息

Tab.5 Feature information of some misclassified samples

特征	503	517	518	529	534
2	1	2	2	2	1
6	1	1	2	3	2
10	2	2	2	2	1
14	2	3	3	2	1
18	2	1	2	2	2
22	1	2	2	2	2
26	1	3	3	1	1

2.4 混合预测

混合预测是先进行特征分的预测，然后再预测录取批次.这样做是非常有意义的，即根据考生以往模拟考试的成绩来预测他/她的高考成绩，然后再由模拟考试成绩和预测的高考成绩一起来预测他/她的录取批次.

特征分的预测和录取批次的预测在前面都已经完成，不需要再训练模型，直接采用2.3节中的分类模型.也就是说，训练样本集不变，仍然是前500个数据，数据的特征为前25个特征.对支持向量机算法，参数设置为$C=100$，且$\gamma=2^{-5}$.对神经网络方法，令隐节点个数为5.与2.3节不同的是测试数据的设计.原始测试数据的第25维特征被2.2节的预测值所代替，也就是说采用预测的特征分来进行分类，测试结果如表6所示.从结果看，虽然预测的情形不太乐观，但是支持向量机仍然比神经网络要好.预测的效果不很理想，主要的原因可能还是因为预测特征分的不精确性导致的.

表6 两种方法在混合预测上的性能对比

Tab.6 Performance comparison of two methods on mixture prediction

	支持向量机	神经网络
精度(%)	68.42	58.16±6.38

3 结论

本文将支持向量机应用到高考成绩预测中，包括对特征分的预测、录取批次的预测以及兼顾两者的混合预测.用支持向量机分类来预测高考的录取批次，以回归分析来预测高考特征分，并成功进行混合预测，并和神经网络的方法进行了比对.实验结果表明，支持向量机比神经网络具有更好的预测性能，且预测具有一定的可行性.然而研究过程中，我们也发现工作还存在很多不足，在数据的收集、对混合预测的效果提升等方面需要进一步的研究.考虑到不同学校不同年份数据的差异，我们希望可以在下一步的工作中引入更多的数据来检测预测性能.

参考文献(References)

[1] STUDLEY R., GREISER S. UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California [J]. Educational Assessment, 2002, 8(1):1-26.

[2] BRIDGEMAN B, BURTON N, CLINE F. Substituting SAT II: Subject tests for SAT I: Reasoning tests: Impact on admitted class composition and quality [J]. Research in Higher Education, 2003, 44(1):83-98.

[3] AL-MUSAWI N M, AL-ANSARI S H. Test of English as a Foreign Language and First Certificate of English Tests as predictors of academic success for undergraduate students at the University of Bahrain [J]. System, 1999, 27(3): 389-399.

[4] STANSFIELD C W, HEWITT W E. Examining the predictive validity of a screening test for court interpreters [J]. Language Testing, 2005, 22(4): 438-462.

[5] FEELEY T H, WILLIAMS V M, WISE T J. Testing the predictive validity of the GRE exam on communication graduate student success: A case study at University at Buffalo [J]. Communication Quarterly, 2005, 53(2):229-245.

[6] 李以渝. 用计算机预测高考成绩分布[J]. 预测, 1985, (6):30-32.

[7] 邓少敏, 张震强, 杨小春. 微型计算机高考分数统计系统 [J]. 江西师范大学学报: 自然科学版, 1985, (4):18-22.

[8] 符华儿. 用微型计算机统计高考分数 [J]. 广西科学院学报, 1982, (1):49-57.

[9] 卫子光, 韩家勋, 苑庆兰, 等. 化学高考预测效度研究初探[J]. 化学教育, 1995, 16(11): 4-11.

[10] 吴根洲. 高考效度研究文献述评 [J]. 教育测量与评价：理论版，2009，(2)：49-51.
[11] 许之所，张丽芳. 高考英语试卷预测效度实证研究[J]. 武汉理工大学学报，2004，17(2)：247-249.
[12] 闪豆豆. 高考英语模拟考试的预测效度分析 [D]. 南京：南京师范大学，2011.
[13] 彭立. 2007 年湖南省高考数学效度研究[D]. 长沙：湖南师范大学，2009.
[14] 亓鲁霞. 意愿与现实：中国高等院校统一招生英语考试的反拨作用研究[M]. 北京：外语教学与研究出版社，2004.
[15] HEATON J B. Writing English Language Tests [M]. London and New York：Longman Group UK Limited，1975.
[16] 韩向峰，刘希玉. 基于 Web 挖掘的高考预测系统的设计与实现[J]. 计算机应用研究，2004，21(8)：160-162.
[17] 吴金财. 考试成绩分析与评价：如何使用平均分——以深圳市高考模拟考部分成绩数据为例[J]. 教育测量与评价：理论版，2013，(8)：57-60.
[18] 李敬文，陈志鹏，李宜义，等. 组合预测模型在高考数据预测中的应用研究 [J]. 计算机工程与应用，2014，50(7)：259-262.
[19] ALPAYDM E. Introduction to Machine Learning [M]. Massachusetts：The MIT Press，2004.
[20] MITCHELL T M. Machine Learning [M]. New York：The McGraw-Hill Companies，Inc.，1997.
[21] DUDA R O，HART P E，STORK D G. Pattern Classification [M]. 2nd ed. New York：John Wiley & Sons，Inc.，2001.
[22] 边肇棋，张学工. 模式识别[M]. 第二版. 北京：清华大学出版社，2000.
[23] 陈志泊. 数据仓库与数据挖掘 [M]. 北京：清华大学出版社，2009.
[24] 周琦. 改进的 C4.5 决策树算法研究及在高考成绩预测分析中的应用[D]. 硕士学位论文，广西大学，2012.
[25] 张琼. 基于贝叶斯方法的高考成绩类别预测 [J]. 太原师范学院学报，2009，8(2)：41-43.
[26] BOSER B E，GUYON I M，VAPNIK V N. A training algorithm for optimal margin classifiers [C]// Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. Pittsburgh：ACM Press，1992：144-152.
[27] CORTES C，VAPNIK V N. Support-vector networks [J]. Machine Learning，1995，20(3)：273-297.
[28] ZHANG L，ZHOU W D，JIAO L C. Hidden space support vector machines [J]. IEEE Transactions on Neural Networks，2004，15(6)：1424-1434.
[29] ZHOU W D，ZHANG L，JIAO L C. Linear programming support vector machines [J]. Pattern Recognition，2002，35(12)：2927-2936.
[30] MANGASARIA O L. Generalized support vector machines [J]. Advances in Large Margin Classifiers，1999，26(5)：135-146.
[31] ZHANG L，ZHOU W D. On the sparseness of 1-norm support vector machines [J]. Neural Networks，2010，23：373-385.
[32] BURGES C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery，1998，2(2)：121-167.
[33] DRUCKER H，BURGES C J C，KAUFMAN L，et al. Support vector regression machines [J]. Advances in Neural Information Processing Systems，2008，28(7)：779-784.
[34] TAX D，DUIN R. Data domain description by support vectors [C]// Proceedings of ESANN99. Brussels：D. Facto Press，1999：251-256.
[35] BEN-HUR A，HORN D，SIEGELMANN H T，et al. Support vector clustering [J]. Journal of Machine Learning Research，2002，2(2)：125-137.
[36] BENNETT K P，DEMIRIZ A. Semi-supervised support vector machines [J]. Advances in Neural Information Processing Systems，1999，9(2)：368-374.
[37] NOBLE W S. What is a support vector machine [J]. Nature Biotechnology，2006，24(12)：1565-1567.
[38] CHANG C C，LIN C J. A library for support vector machines [J/OL]. (2010. 2. 22) [2014. 3. 10]http://www. csie. ntu. edu. tw/～cjlin/libsvmtools/.