文章编号:2095-2163(2019)05-0176-04 中图分类号:TP181 文献标志码:A

集成学习在糖尿病预测中的应用

张玉玺,贺 松,尤思梦
(贵州大学 医学院,贵阳 550025)

摘 要: 糖尿病、高血压和心脑血管病并称为影响人类健康的三大杀手,不仅对患者的生命健康造成严重的威胁,也给患者的家庭造成严重的经济负担。对糖尿病做出准确的预测,意义深远。本文采用了KNN、支持向量机、逻辑回归、随机森林、集成学习五种方法对糖尿病数据进行预测,分别取得了71.86%,72.29%,74.46%,71.87%,76.62%的准确率。结果表明,集成学习预测效果最佳,验证了其优异性。

关键词: 集成学习;糖尿病;预测;机器学习

Application of integrated learning in diabetes prediction

ZHANG Yuxi, HE Song, YOU Simeng
(Medical College, Guizhou University, Guiyang 550025, China)

【Abstract】 Diabetes, hypertension and cardiovascular and cerebrovascular diseases are called three killers of human health, which not only posed a serious threat to the patient's life and health , but also caused a serious economic burden to the patient's family. Accurate prediction of diabetes has profound implications. In this paper, five methods including KNN, Support Vector Machine, Logistic Regression, Random Forest and Integrated Learning are used to predict diabetes data, and the accuracy rates of 71.86%, 72.29%, 74.46%, 71.87% and 76.62% are achieved respectively. The results show that the integrated learning has the best prediction effect and its excellent performance is verified.

【Key words】 integrated learning; diabetes; prediction; machine learning

0 引 言

目前,随着科学技术的发展,大数据信息时代已悄然来临,人工智能技术的研究也取得了长足进步,越来越多的学者将研究的关注点转到医疗智能诊断上来。作为人工智能技术的重要分支,机器学习也已广泛地被应用于医学模型的构建中,并发挥着不可替代的作用。机器学习[1-2](Machine Learning, ML)是一门交叉学科,涉及统计学、概率论等多个领域,该算法是从已有数据中挖掘分析获得规律,并利用这些规律对未知数据做出预测。

糖尿病是一种以高血糖为主要特点的代谢性疾病,典型特征为多尿、多饮、多食、体重减轻。国际糖尿病联盟(International Diabetes Federation, IDF)于2017发布的全球糖尿病地图数据表明,目前全球共有4.25亿成人(20~79岁)糖尿病患者,估计患病率为8.8%;中国成人糖尿病患者数量高达1.14亿,占全球成人糖尿病患者总数的1/4以上,这一数据仍在继续增长,预计到2045年将增至1.2亿[3]。而中国大多数的糖尿病患者在患病之前,自身既没有察觉、也没有明确意识,因此,对糖尿病进行早期的诊断则显得尤为重要。

本文选用了机器学习算法中的KNN、支持向量机、逻辑回归、随机森林四种分类算法构建糖尿病单一分类器,同时通过投票法作为结合策略结合上述四种分类算法构成分类投票聚合模型Voting。基于此,将运用前述五种分类器对糖尿病数据进行分析、预测,并运用10折交叉验证方法对各个模型进行评估比较,选出最好的糖尿病预测模型,以期为糖尿病的早期筛查与诊断提供辅助决策。本文拟展开研究论述如下。

1 机器学习算法

1.1 KNN算法

KNN(k-NearestNeighbor)算法,又叫K近邻算法,或者说K最近邻分类算法,是著名的模式识别

基金项目: 贵州省数字健康管理工程技术研究中心项目(黔科合G字[2014]4002号)。

作者简介: 张玉玺(1995-),男,硕士研究生,主要研究方向:机器学习、医学智能信息处理;贺 松(1974-),男,副教授,硕士生导师,主要研究方向:医疗大数据、数字图像处理;尤思梦(1999-),女,本科生,主要研究方向:医学信息处理。

通讯作者: 贺 松 Email:yuxi995@foxmail.com

收稿日期: 2019-07-12

统计学方法。KNN 算法在理论上比较成熟，是最简单的机器学习算法之一，在机器学习分类算法中占据着重要位置。K 最近邻指的是 K 个最近的邻居，也就是可以用最接近的 K 个邻居来表示每个样本。

KNN 中 K 个最近的邻居的选取是基于所选用的距离函数，计算时默认的是欧氏距离[4]。二维空间内点 $A(x_1,y_1)$ 与 $B(x_2,y_2)$ 之间的欧氏距离公式为：

$$\rho_{AB}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}, \tag{1}$$

三维空间内点 $A(x_1,y_1,z_1)$ 与点 $B(x_2,y_2,z_2)$ 之间的欧氏距离公式为：

$$\rho_{AB}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}, \tag{2}$$

n 维空间的欧氏距离公式为：

$$\rho=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}. \tag{3}$$

K 近邻算法中，K 值的选取对于整个算法起着决定性作用[5]。当 K 的取值过小时，一旦数据中有噪声存在，将会对预测结果产生比较大的影响。当 K 的取值过大时，容易受到样本均衡问题的影响，训练的模型会用较大邻域中的训练数据进行预测，模型的近似误差将会增大。

K 要尽量选择奇数。选偶数，很有可能会发生分类结果相等的情况，不利于模型的预测，而选择奇数则可以保证在预测结果的最后产生一个较多的类别。因此，研究必须要选择合适的 K 值来构建文中的 KNN 模型，本文通过 10 折交叉验证确定 KNN 模型的最优 K 值为 7。

1.2　支持向量机算法

支持向量机（Support Vector Machine，SVM）算法 1964 年由 Cortes 和 Vapnik[6] 提出，此后历经一系列改进和扩展，目前已经发展成较为成熟的机器学习模型。SVM 不仅能够实现分类、回归任务，而且能够进行异常值的检测，是机器学习领域中广为流行的模型。

支持向量机尝试找到一个最优超平面来对样本进行分割，分割的原则是间隔最大化，该超平面能够将正类和负类正确分隔开。虽然 SVM 分类器在许多数据上的表现都很好，可是仍需指出，现实中的大部分数据并不是线性可分的，这个时候满足这样条件的超平面就根本不存在，即特征空间存在超曲面将正类和负类分开。对于这种情况，可以将训练样本从原始空间映射到一个更高维的希尔伯特空间（Hilbert space）中去，将其转化为线性问题，使得样本在这个空间中线性可分。

SVM 将非线性问题转化为线性问题的方法关键就是选择一个核函数，常用的核函数有线性核（linear）、多项式核（poly－nomial）、高斯 RBF 核和 Sigmoid 核函数。在本文 SVM 模型的构建中，研究选择的核函数是多项式核。

1.3　逻辑回归算法

逻辑回归（Logistic Regression，LR）算法，又称对数几率回归，虽然名字中带“回归”字样，但其实际上却是一种分类学习方法，主要应用于两分类问题。逻辑回归由于具有计算速度快、解释性好以及容易扩展和实现等优点，常会应用于疾病诊断，经济预测等方面。逻辑回归算法使用 Sigmoid 函数作为研究中的预测函数，对于输入变量 x，通过线性函数 $y=wx+b$ 的运算，输出变量 y，y 则通过 Sigmoid 函数转换成标签化的结果。模型函数的阈值可以进行设置，当 Sigmoid 函数的输出值大于研究设定的阈值时，模型会将其判定为“1”这一个类别；否则判定为“0”这一类别，函数阈值是一个可调节的参数。其对应数学公式如下：

$$Sigmoid(x)=\frac{1}{1+e^{-x}}. \tag{4}$$

逻辑回归也会面临过拟合的问题，所以就要采取一定的措施来缓解模型过拟合。通用的方法是在逻辑回归的代价函数上，加入正则化项，从而能得到一个拟合较为适度的机器学习模型，常见的正则化手段有 L1 正则化和 L2 正则化[7]。

1.4　随机森林算法

随机森林（Ramdom Forest，RF）算法是由 Breiman 提出的一种基于 CART 决策树的组合分类器算法，可构造出多个树形分类模型。随机森林是一种集成学习算法，是由多个决策树合并在一起形成的组合识别模型。“随机”一词在这里有两层含义。第一层可以理解为在总训练样本中随机有放回地为森林中的每个决策树选取训练集；第二层是对森林中的每个决策树从所有样本属性中随机不放回地选择部分样本属性。

随机森林的每棵树都可以得出分类决策结果，通过采用森林内决策树投票，根据少数服从多数的原则，来判定待测样本的类别，而所有树中票数较高的类别即为最终结果。随机森林分类精度相对较高，具有不易过拟合、抗噪声能力强且易实现等特点[8]，但运算量也相对较大。

1.5 集成学习

1.5.1 集成学习原理

集成学习通过构建并结合多个学习器完成学习任务。与一般的学习方法不同,一般的学习方法是用训练数据构造一个学习器,而集成学习方法是构造多个学习器并通过一定的策略将其结合起来,上文中提到的随机森林算法就是最常见的集成学习算法。但在实际模型构建中,由于每个模型都有其各自的优势及局限性,研究只能得到多个在某些方面有偏好的学习器(弱学习器)。而集成学习则能将多个弱学习器相结合,以期得到一个稳定且在各个方面性能表现都比较出色的模型。在此情况下,集成学习能够综合各个学习器的预测结果,即使某一学习器因为自身不足导致分类错误,可是只要大部分的学习器预测正确,最终仍能得到正确的预测结果[9]。

1.5.2 集成学习结合策略

对于机器学习中的分类任务,最常用的结合策略是投票法,每个弱分类器给出自己的分类预测,再通过投票法结合后得出最终的结果。机器学习中的投票法也有不同的方式,最常见的是简单投票法,包括相对多数投票法和绝对多数投票法。对此可做阐释分述如下。

相对多数投票法中,每个分类器向其中一个类别投票,再将得票数最多的类别作为最终类别,不会出现无分类结果的情况,如果存在多个类别最终获得的票数相等且最高,就随机选择一个作为最终类别。设 $H(x)$ 为集成输出的类别,则相对多数投票法的计算公式为:

$$H(x) = C_{\arg} \max_j \sum_{i=1}^{T} h_i^j(x) , \tag{5}$$

绝对多数投票法是将获得一半以上票数的类别作为最终的类别,如果没有哪一个类别获得一半以上的票数,则最终无分类结果。绝对多数投票法的计算公式可表示为[10]:

$$H(X) = \begin{cases} C_j, & 若 \sum_{i=1}^{T} h_i^j(x) > \frac{1}{2} \sum_{q=1}^{l} \sum_{i=1}^{T} h_i^q(x); \\ 无类别, & 其它. \end{cases} \tag{6}$$

此外,还有一种相对复杂的投票机制是加权投票法,即将每个分类器的投票结果乘以一个权重 w_i,再将所有乘以权重后的结果求和,最终以最大的票数类别作为最终的类别。研究推得其计算公式可表示为:

$$H(x) = C_{\arg} \max_j \sum_{i=1}^{T} w_i h_i^j(x) . \tag{7}$$

在本文中,使用了 4 个单一分类器,即:KNN 分类器、SVM 分类器、逻辑回归分类器和随机森林分类器,通过把 4 个分类器的预测结果采用简单投票法中的相对多数投票法作为结合策略结合起来,得票数最多的类别作为集成模型最终的预测类别。

2 实验结果与分析

2.1 数据来源

本研究采用的数据来源于开放的皮马印第安人糖尿病数据集,该数据集由 768 个皮马印第安人糖尿病信息样本组成(样本均为女性)。其中,每个样本均包含 Pregnancies(是否怀孕)、Glucose(葡萄糖含量)、Blood Pressure(血压指数)、Skin Thickness(皮肤厚度指数)、Insulin(胰岛素含量)、BMI(体重指数)、Diabetes Pedigree Function(糖尿病谱系功能)、Age(年龄)共 8 个输入变量,8 个输入变量全部为连续型变量,无需设置哑变量,同时包含 Out come(结果)一个输出变量,当 Out come 的值为 1 时代表患糖尿病,当 Out come 的值为 0 时表示未患糖尿病。

2.2 模型建立

机器学习算法模型的预测能力与训练样本的数量关系密切,根据以往的经验和相关文献研究,将数据集按照 7:3 的比例进行划分,其中 70%(包含 538 条样本)作为训练集的数据资料,用来建造预测模型;另外 30%(包含 230 条样本)作为测试集数据资料,用来检测和评价模型的性能效果。

研究中,采用 Python 语言开发的 sklearn 机器学习库中的 KNN 算法、支持向量机算法、逻辑回归算法、随机森林算法,以是否怀孕、葡萄糖含量、年龄等 8 个特征作为自变量,患者是否患糖尿病作为因变量,分别构建 4 个单一分类器和以相对多数投票法作为结合策略的集成分类器。实验中,使用 10 折交叉验证对模型参数进行调优,以使模型具有最优的参数组合。

2.3 模型评估

本文主要通过准确率、灵敏度、ROC 曲线下面积等指标对构建的分类器模型进行性能评价,具体结果见表 1 和图 1。

由表 1 可知,在对糖尿病数据的预判上,集成模型 Voting 的效果是这 5 个模型中最好的,其准确率达到了 76.62%,比最高的单一分类器提升了

2.16%，其次是逻辑回归 74.46%，再次是支持向量机 72.29% 和随机森林 71.87%，KNN 的效果最差，为 71.86%。

AUC（ROC 曲线下面积）能够体现模型性能的优劣，图 1 显示的是各个分类器的 ROC 曲线。曲线越是靠近左上方，曲线下的面积就越大，表明该算法的预测效果越好。本实验采用 10 折交叉验证预测得到了各模型的 *AUC* 值。由表 1 和图 1 可知，5 种机器学习方法 *AUC* 值的排名依次是：集成模型 Voting 为 0.802，逻辑回归为 0.791，随机森林为 0.782，支持向量机为 0.718，KNN 为 0.717。

综上，研究将选择准确率最高、*AUC* 值最大的集成模型 Voting 作为最终的糖尿病数据预测模型。

表 1　各分类器性能比较

Tab. 1　The performance comparison of different classifiers

模型类型	准确率/%	*Precision*	*Recall*	*F* 值	*AUC*
KNN	71.86	0.66	0.46	0.54	0.717
支持向量机	72.29	0.66	0.47	0.55	0.718
逻辑回归	74.46	0.71	0.48	0.58	0.791
随机森林	71.87	0.70	0.37	0.49	0.782
Voting	76.62	0.79	0.47	0.59	0.802

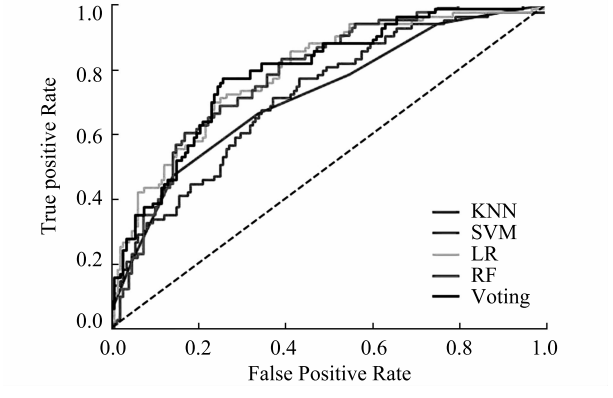

图 1　各分类器 ROC 曲线

Fig. 1　The ROC curve of different classifiers

3　结束语

本文阐述了机器学习中的 KNN、支持向量机、逻辑回归、随机森林四种算法以及集成学习的基本原理与特点，并基于糖尿病数据分别建立相应的模型，利用交叉验证对模型的参数进行了优化，通过准确率、*AUC* 值等模型评价指标对模型进行选择。结果表明以相对多数投票法作为结合策略的集成模型 Voting 具有更好的预测效果。由于数据集样本量有限，导致整体预测准确率偏低。但有理由相信，在有足够数据的情况下，将会构建出更加准确的预测模型。希望本次研究能够为糖尿病的预测提供一定的帮助，并能够为国内的医疗事业做出应有的贡献。

参考文献

[1] MORPURGO R, MUSSI S. An intelligent diagnostic support system[J]. Expert Systems, 2001, 18(1): 43-58.

[2] SELA R J, SIMONOFF J S. RE-EM trees: A data mining approach for longitudinal and clustered data [J]. Machine Learning, 2012, 86(2): 169-207.

[3] International Diabetes Federation. IDF Diabetes Atlas[M]. 8th ed. Brussels: International Diabetes Federation, 2017.

[4] 吴兴惠，周玉萍，邢海花，等. 机器学习分类算法在糖尿病诊断中的应用研究[J]. 电脑知识与技术，2018，14(35)：177-178，195.

[5] 杨帆，林琛，周绮凤，等. 基于随机森林的潜在 k 近邻算法及其在基因表达数据分类中的应用[J]. 系统工程理论与实践，2012，32(4)：815-825.

[6] CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[7] 史双睿. 异质集成学习器在鸢尾花卉分类中的应用[J]. 电子制作，2019(2)：45-47，79.

[8] 阚红星，张璐瑶，董昌武. 一种 2 型糖尿病中医证型的舌图像识别方法[J]. 中国生物医学工程学报，2016，35(6)：658-664.

[9] 曾一平. 基于集成学习的小麦识别研究[J]. 现代商贸工业，2019，40(17)：207-209.

[10] ZHOU Zhihua. Ensemble methods: foundations and algorithms [M]. Boca Raton: CRC Press, 2012.