文章编号：2095-2163(2021)01-0064-05 中图分类号：TP391 文献标志码：A

早期糖尿病风险预测模型的比较研究

王成武，晏峻峰
（湖南中医药大学 信息科学与工程学院，长沙 410208）

摘 要： 糖尿病是一种比较常见的慢性疾病，并且存在较长的无症状阶段。本文主要介绍了机器学习中的 5 种分类算法，分别是朴素贝叶斯、支持向量机、逻辑回归、决策树和集成分类器 Random Forest，并在 Weka 数据挖掘平台上，对糖尿病数据进行挖掘分析，根据混淆矩阵、Kappa 系数、ROC 曲线、均方根误差以及相对绝对误差这几个性能指标对分类器效果进行分析，找到最适合糖尿病疾病预测的算法，为当今医疗行业其他疾病数据的挖掘分析提供思路。

关键词： 糖尿病；机器学习；集成分类器；数据挖掘；Weka

Comparison of early diabetes risk prediction models

WANG Chengwu, YAN Junfeng
(School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China)

【Abstract】 Diabetes is a relatively common chronic disease, and there is a long asymptomatic stage. This article mainly introduces five classification algorithms in machine learning, which are Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest, an integrated classifier. On the Weka data mining platform, the diabetes data is mined and analyzed. The effect of the classifier is analyzed according to the confusion matrix, Kappa coefficient, ROC curve, root mean square error and relative absolute error, and the most suitable algorithm for diabetic disease prediction is achieved, which could provide ideas for the current medical industry data mining.

【Key words】 diabetes; machine learning; integrated classifier; data mining; Weka

0 引 言

糖尿病是一种终身疾病，可引发心脏病、血管疾病等并发症[1]，不仅影响了患者的生活质量，也会带来相应的经济负担，所以进行早期糖尿病风险预测具有十分重要的意义。

作为重要的数据挖掘技术，机器学习等人工智能技术，在糖尿病预测与治疗上应用得很多。例如，Purushottam 等人[2]分别用 C4.5 算法和 Partial Tree 算法自动提取糖尿病预测规则来预测患者的糖尿病风险。Santhanam 等人[3]用遗传算法对糖尿病数据集进行维数约简并利用支持向量机进行了糖尿病的预测。胡玮[4]基于改进邻域粗糙集和随机森林算法进行了糖尿病的预测研究。黄艳群等人[5]利用患者相似性建立了个性化糖尿病预测模型。

本文将机器学习技术应用在早期糖尿病风险预测数据集上，构建多种分类模型，通过各种性能评价指标对模型进行分析，选择最优分类模型，该模型可通过评估症状来检查用户患糖尿病的风险。

1 基本原理及方法

1.1 实验数据

本文选取的是 UCI 机器学习库中的早期糖尿病风险预测数据集，共包含 520 个样本，其中阳性样本为 320 个，阴性样本为 200 个，每条样本数据包含 16 个特征属性和一个类属性，分别是：Age（年龄）、Gender（性别）、Polyuria（多尿症）、Polydipsia（烦渴）、sudden weight loss（体重减轻）、weakness（虚弱）、Polyphagia（多食症）、Genital thrush（生殖器鹅口疮）、visual blurring（视觉模糊）、Itching（瘙痒）、Irritability（烦躁）、delayed healing（延迟康复）、partial paresis（部分偏瘫）、muscle stiffness（肌肉紧张）、Alopecia（脱发）、Obesity（肥胖）、class（类别）。

1.2 算法原理

1.2.1 朴素贝叶斯

朴素贝叶斯算法（Naive Bayes，NB 算法）是常用的概率分类算法[6]，朴素贝叶斯具有一些明显的特征，例如计算的速度非常快、准确率高、方法简单

基金项目：湖南省教育厅科研重点项目（18A219）。

作者简介：王成武（1997-），男，硕士研究生，主要研究方向：机器学习、数据挖掘；晏峻峰（1965-），女，博士，教授，博士生导师，主要研究方向：人工智能及其应用。
通讯作者：晏峻峰 Email：junfengyan@hnucm.edu.cn

收稿日期：2020-10-28

等特点，在一般贝叶斯理论的基础上，朴素贝叶斯中的‘朴素’一词就是假定样本中的属性彼此独立地对其产生影响，并不考虑属性之间的依赖关系，在实际应用中对于大部分比较复杂的问题都是很有成效的。

基于属性条件独立性假设，在样本分类任务中，计算样本 w 所属类别的概率 $P(c|w)$，计算方式为：

$$P(c|w)=\frac{P(c)P(w|c)}{P(w)}=\frac{P(c)}{P(w)}\prod_{i=1}^{n}P(w_i|c), \tag{1}$$

其中，n 表示属性个数，W_i 表示样本 w 在第 i 个属性上的取值。$P(w)$ 在所有类别中都是相同的，因此在公式(1)的基础上知朴素贝叶斯分类器的基本表达式：

$$R(w)=\max\ P(c)\prod_{i}^{n}P(w_i|c). \tag{2}$$

1.2.2　支持向量机

支持向量机(Support Vector Machine，SVM)是一种常用的二分类模型[7]，通过使用给定的样本数据集进行建模，在样本空间中找到一个最优的划分超平面，该平面产生的分类结果是最具有鲁棒性的，并且对未见示例有最好的泛化能力。SVM 是针对线性可分情况进行分析的，对于非线性分类问题，可以通过核函数将低维样本空间映射到高维特征空间，这样高维特征空间即可采用线性算法对样本的非线性特征进行线性分析。常用的核函数有以下几种：

(1)线性核函数，对应公式为：

$$ker(x,y)=xy, \tag{3}$$

(2)多项式核函数，对应公式为：

$$ker(x,y)=(xy+1)^d, \tag{4}$$

(3)径向基核函数，对应公式为：

$$ker(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right), \tag{5}$$

(4)拉普拉斯核函数，对应公式为：

$$ker(x,y)=\exp\left(-\frac{\|x-y\|}{\delta}\right), \tag{6}$$

(5)Sigmoid 核函数，对应公式为：

$$ker(x,y)=\tanh[\beta(xy-\theta)]. \tag{7}$$

1.2.3　逻辑回归

逻辑回归(logistics regression)属于监督学习方法，是一种广义的线性回归分析模型，主要用于概率预测或分类。逻辑回归最基本的学习算法是极大似然，即假设数据是伯努利分布，通过极大似然函数来推导损失函数，使用梯度下降来求解参数，以此来对数据进行二分类。逻辑回归中常用建模函数的数学表达式如下：

$$f(x)=\frac{1}{1+e^{-x}}, \tag{8}$$

$$h(x)=f(\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_{16}x_{16}). \tag{9}$$

其中，$f(x)$ 指观测个体患上糖尿病的概率，是一个 Sigmoid 函数；$x_1,x_2,\cdots,x_{16}$ 是糖尿病数据集的 16 个特征属性；θ 是权重参数。将 Sigmoid 函数与线性回归两者结合，使最终预测概率的值处于 0～1 之间：若大于 0.5，将其归为 Positive 类；若小于 0.5，则归为 Negative 类。

1.2.4　决策树 J48

ID3 算法中根据信息增益评估和选择特征，每次选择信息增益最大的特征作为判断模块建立子结点，使用信息增益的缺点是偏向于具有大量值的属性，而且该算法不能够处理连续分布的数据特征，于是就有了 C4.5 算法，在 WEKA 中称为 J48 算法，该算法是在 ID3 算法的基础上进行改进而产生的[8]，算法中包含 ID3 算法的所有功能，除此之外，还可以利用信息增益率来选择属性，合并具有连续属性值、处理含有未知属性值的训练样本等。

1.2.5　随机森林

随机森林(Random Forest)是由 Breiman 提出的[9]，是一种组合分类器，其基本单元就是决策树。将决策树作为个体学习器，加入了随机样本选择和随机特征选择策略。对于本文而言，即随机地从 16 个属性特征中选择 m 个属性($m<16$)，并且使用有放回的抽样策略从数据集中选取样本。在新数据集上训练决策树，通过每棵决策树的预测结果来决定测试样本最终的预测结果。算法的整体流程如图 1 所示。

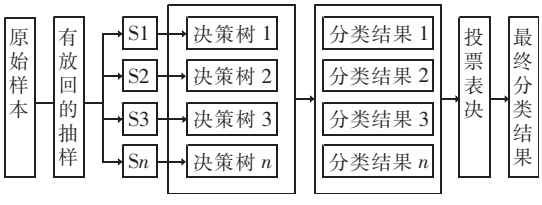

图 1　随机森林算法流程图

Fig. 1　Random Forest algorithm flow chart

1.3　性能指标

1.3.1　混淆矩阵

混淆矩阵可用来判断分类器的优劣，详见表 1。所有评价指标具体定义如下。

表 1 混淆矩阵

Tab. 1 Confusion matrix

	预测为正例	预测为假例
真实为正例	真正例 *TP*	伪反例 *FN*
真实为假例	伪正例 *FP*	真反例 *TN*

(1)精确率(*Precision*):预测结果为正例样本中真实为正例的比例,公式如下:

$$precision=\frac{TP}{TP+FP}, \tag{10}$$

(2)召回率(*Recall*):真实为正例的样本中预测结果为正例的比例,公式如下:

$$recall=\frac{TP}{TP+FN}, \tag{11}$$

(3) *F*:为精确率(*Precision*)和召回率(*Recall*)两者的调和平均值,公式如下:

$$F=\frac{2*precision*recall}{precision+recall}. \tag{12}$$

1.3.2 *Kappa* 系数

Kappa 系数是一种计算分类精度的方法,用于衡量模型预测结果和实际分类结果是否一致,其计算公式为:

$$Kappa=\frac{P_a-P_e}{1-P_e}. \tag{13}$$

其中, P_a 为实际一致率, P_e 为理论一致率。*Kappa* 系数的取值在 0 ~ 1 之间,若 $Kappa \geqslant 0.75$,则表明分类器的一致性很好。

1.3.3 ROC 曲线

受试者工作特征曲线(receiver operating characteristic curve,ROC 曲线),用来比较 2 个分类模型有效性的可视化工具,*AUC* 表示 ROC 曲线下的面积,取值在 0.5 ~ 1 之间。*AUC* 可以直观地评价分类器的好坏,值越大越好。

1.3.4 均方根误差

均方根误差(*RMSE*)是对样本数据集抽样误差的度量,其数值越小表示模型越稳定。

1.3.5 相对绝对误差

相对绝对误差(*RAE*)是预测数值与实际数值两者差的绝对值,数值越小则表明模型越优。

2 实验结果与分析

使用 WEKA 数据挖掘平台对 6 个分类器进行分析。在 Test options 栏目下选择十折交叉验证法,依此选择分类算法进行实验,其中 SVM 算法的核函数选用径向基核函数。实验产生的各性能指标的结果见表 2 和表 3。

表 2 精度指标

Tab. 2 Accuracy index

算法	*TP* Rate	*FP* Rate	*Precision*	*Recall*	*F*-Measure
Naive Bayes	0.871	0.120	0.878	0.871	0.872
SVM	0.942	0.055	0.944	0.942	0.943
Logistics	0.923	0.084	0.923	0.923	0.923
J48	0.960	0.035	0.961	0.960	0.960
Random Forest	0.962	0.037	0.962	0.962	0.962

由表 2 分析可知,精确率(*Precision*)为预测出的真阳性病例在预测为阳性病例中的比例,召回率(*Recall*)为预测出的真阳性病例在实际真阳性病例中的比例,精确率和召回率是相互影响的,理想情况下是两者都高,但一般情况下是精确率高,召回率就低;反之,召回率高,精确率就低。在各种疾病的监测研究中,一般采用的方法是在保证精确率的条件下,提升召回率。精度指标中的 *F* 值综合了精确率和召回率,可以用来综合评价实验结果的质量。可以看出 Random Forest 的精确率、召回率和 *F* 值远远高于 Naive Bayes、Logistics 等分类器。

表 3 分类结果

Tab. 3 Classification result

算法	*Accuracy*	*Kappa*	*RMSE*	*RAE*	*ROC Area*
Naive Bayes	0.871	0.734	0.318	0.315	0.946
SVM	0.942	0.879	0.240	0.122	0.944
Logistics	0.923	0.838	0.252	0.235	0.969
J48	0.960	0.916	0.198	0.116	0.966
Random Forest	0.975	0.919	0.196	0.081	0.998

在此基础上,对表 3 所得实验结果进行分析,可得各项指标的阐释分述如下。

(1)分类器准确率(*Accuracy*):由表 3 中数据分析可知,在这 5 个分类器中,Random Forest 算法对样本分类的准确率最高,其次是 J48 算法,Naive Bayes 算法的准确率较差,于是通过属性约简的方式来优化 Naive Bayes 算法的预测结果,即找出预测效果最好的属性集,使用 CfsSubsetEval 属性评估器和 GreedyStepwise 搜索方法进行属性选择,根据最终的属性集进行实验,得出 Naive Bayes 算法的准确率为 0.88,相比之前的准确率略有提升,但还是远不及其它分类器的预测结果。总地来说,集成分类器 Random Forest 的识别准确率要高于一般的单一分类器。

(2) *Kappa* 系数比较(*Kappa*):若分类器与随机一个分类器的分类结果全一致,则 *Kappa* 系数为 1,

反之为 0。所以 *Kappa* 系数越大，表明分类的效果越好。由表 3 中数据可以得知 Random Forest 算法的 *Kappa* 系数值最大，故该算法相比其它算法在此数据集上建立的模型更好。

（3）均方根误差比较（*RMSE*）：由表 3 中数据可知 Random Forest 算法和 J48 算法所建立模型运行产生的 *RMSE* 值是最低的两个，其次是 Logistics 和 SVM，两者的结果较为相近，5 种分类器中的 Naive Bayes 的 *RMSE* 值最大，效果最差。

（4）相对绝对误差比较（*RAE*）：由表 3 中数据可知 SVM 和 J48 的 *RAE* 值较为相近，预测结果相差不大，5 种分类算法中 Random Forest 的 *RAE* 值最小，表明该模型最优，所预测的数据值最为贴近实际值。

（5）ROC 曲线面积比较（*ROC Area*）：曲线图的横纵坐标分别表示模型预测数据的假阳性率和真阳性率，ROC 曲线越靠近纵轴，表示模型越好。5 类预测模型 ROC 曲线图如图 2 ~ 图 6 所示。

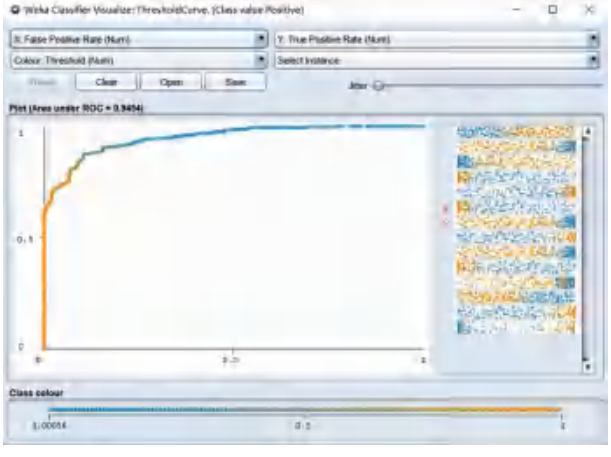

图 2　Naive Bayes 预测模型 ROC 曲线

Fig. 2　ROC curve of Naive Bayes prediction model

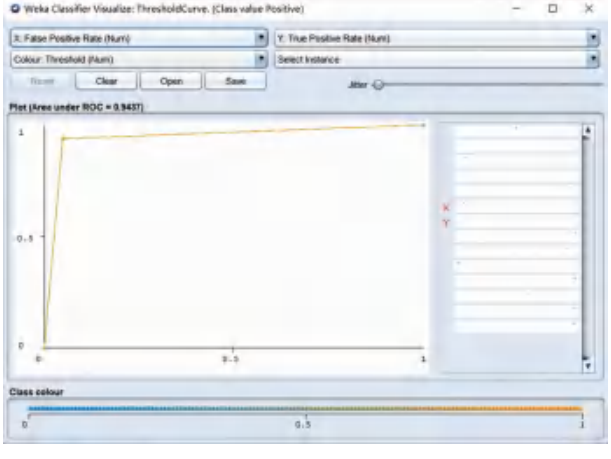

图 3　SVM 预测模型 ROC 曲线

Fig. 3　ROC curve of SVM prediction model

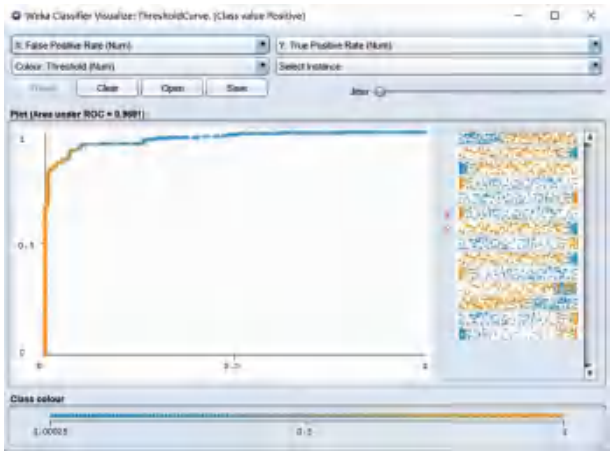

图 4　Logistics 预测模型 ROC 曲线

Fig. 4　ROC curve of Logistics forecast model

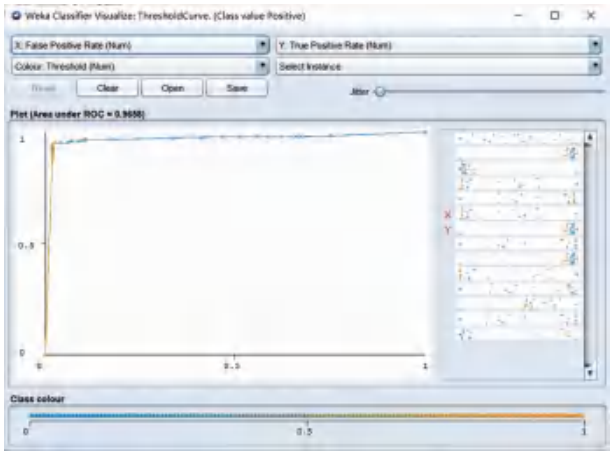

图 5　J48 预测模型 ROC 曲线

Fig. 5　ROC curve of J48 prediction model

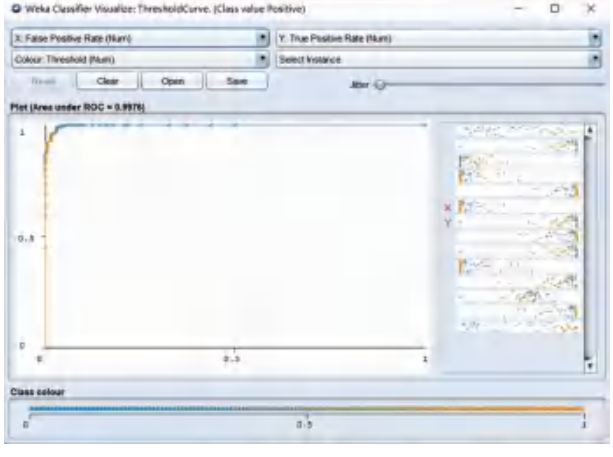

图 6　Random Forest 预测模型 ROC 曲线

Fig. 6　ROC curve of Random Forest prediction model

由图 2 ~ 图 6 可看出，Random Forest 预测模型的 ROC 曲线最为靠近纵轴，所以该算法的建模效果最优，其次是 Logistics 预测模型。同样地，该结果表明集成分类器的建模效果要高于一般的单一分类器。

为了验证这几种模型在不同数据量的数据集上的表现是否具有一致性，分别随机抽取 320，420 个样本，重复以上实验，选取 *F-Measure* 作为此次实验中模型的性能评价指标，最终的训练结果如图 7 所示。

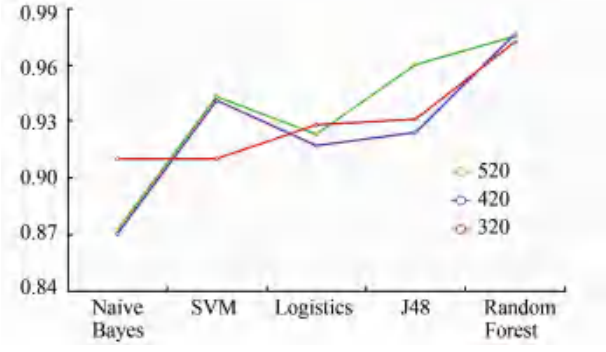

图 7 *F* 值变化曲线

Fig. 7 *F* value change curve

从图 7 中可以看出，随着数据集的减少，各模型的分类效果是有所变化的，在样本数据集数量为 320 的时候，朴素贝叶斯算法的分类效果相比之前有较大的上升幅度，而支持向量机算法在数据集减少到 320 的时候分类效果相比之前有较大的下降幅度。数据集大小为 520 和 420 时，各模型分类效果的变化趋势基本一致，而数量为 320 的时候 Logistics 的分类效果是胜于 SVM 算法的，但总地来说，不论数据集是多少，分类效果最优的还是集成分类器 Random Forest。

3 结束语

本文基于 WEKA 数据挖掘平台，使用 5 种分类算法对早期糖尿病风险预测数据集进行分析，并利用多种评价指标来确定分类效果。从实验结果可以看出，集成分类器 Random Forest 在该糖尿病数据集上的分类效果最佳。故今后医疗行业其它疾病的预测，可根据实际情况，通过结合策略将多个单一分类器整合起来形成集成分类器，以此来提升模型的分类精度。

参考文献

[1] 刘月. 基于数据挖掘技术的 2 型糖尿病的预测与健康管理研究[D]. 秦皇岛：燕山大学，2018.

[2] PURUSHOTTAM, SAXENA K, SHARMA R. Diabetes mellitus prediction system evaluation using C4. 5 rules and partial tree [C]// 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions). Noida, India：IEEE, 2015：1-6.

[3] SANTHANAM T, PADMAVATHI M S. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis[J]. Procedia Computer ence, 2015, 47：76-83.

[4] 胡玮. 基于改进邻域粗糙集和随机森林算法的糖尿病预测研究[D]. 北京：首都经济贸易大学，2018.

[5] 黄艳群，王妮，张慧，等. 利用患者相似性建立个性化糖尿病预测模型[J]. 医学信息学杂志，2019，40(1)：54-58.

[6] KONONENKO I. Seminaive bayesian classifier[C]// Proc. of the 6th European Working Session on Learning. Berlin, Heidelberg：Springer, 1991：206-219.

[7] 兰欣，卫荣，蔡宏伟，等. 机器学习算法在医疗领域中的应用[J]. 医疗卫生装备，2019，40(3)：93-97.

[8] 高海宾. 基于 Weka 平台的决策树 J48 算法实验研究[J]. 湖南理工学院学报(自然科学版)，2017，30(1)：21-25.

[9] 刘文博，梁盛楠，秦喜文，等. 基于迭代随机森林算法的糖尿病预测[J]. 长春工业大学学报，2019，40(6)：604-611.

(上接第 63 页)

[5] FOUEDJIO F. Clustering of multivariate geostatistical data[J]. WIREs Comput. Stat., 2020, 12：e1510.

[6] MAHMOOD S, MUELLER K. Taxonomizer：Interactive construction of fully labeled hierarchical groupings from attributes of multivariate data[J]. IEEE Transactions on Visualization and Computer Graphics, 2020：2875-2890.

[7] 吴笑丰. 基于微信公众号的中学校园失物招领系统设计[J]. 科技传播，2020，12(6)：153-154.

[8] 王瑞东. Java web 软件框架技术探讨[J]. 中国新通信，2019，21(9)：46.

[9] 梁弼，张紫桂，熊伦. 一种轻量级的多层 Web 应用架构研究及使用[J]. 陕西科技大学学报，2020，38(4)：166-171.

[10] 韩振才. 大数据时代下计算机软件技术的应用[J]. 电子技术与软件工程，2020(15)：52-53.