

## 基于集成学习的 2 型糖尿病患者降糖药用药方案智能分类探讨

宋亚男<sup>1</sup>, 金昕晔<sup>2</sup>, 张颖<sup>3</sup>, 陈康<sup>2</sup>, 应俊<sup>1</sup>, 薛万国<sup>1</sup>, 母义明<sup>2</sup>

<sup>1</sup>解放军总医院 医疗大数据中心, 北京 100853; <sup>2</sup>解放军总医院第一医学中心 内分泌科, 北京 100853;

<sup>3</sup>解放军总医院第一医学中心 眼科, 北京 100853

**摘要:** **目的** 探讨集成学习中的 Adaboost 算法在 2 型糖尿病患者降糖药用药模式分析中的应用。**方法** 收集解放军总医院第一医学中心 2013 - 2017 年的 2 型糖尿病患者住院患者病例资料 3 005 例, 随机选择 1 697 例为训练集, 1 308 例为测试集, 根据医嘱用药、生化检验、基本体征、人口统计学等资料, 应用 Adaboost 算法建立学习模型, 对患者用药模式进行分类, 并计算模型的准确性和 Kappa 系数。**结果** Adaboost 模型预测的用药分类准确率为 64.2%, Kappa 系数为 0.36。通过 Adaboost 模型分析, 发现与降糖药用药相关的重要变量有尿肌酐、糖化血红蛋白、肌酸激酶同工酶、空腹血糖等。**结论** Adaboost 算法在降糖药用药方案的预测方面具有较好的效果, 集成学习方法在患者用药决策方面具有一定可行性。

**关键词:** 2 型糖尿病; 降糖药; Adaboost 算法; 多分类学习

**中图分类号:** R 587.1 **文献标志码:** A **文章编号:** 2095-5227(2019)08-0719-06 **DOI:** 10.3969/j.issn.2095-5227.2019.08.004

**网络出版时间:** 2019-08-22 17:13 **网络出版地址:** <http://kns.cnki.net/kcms/detail/10.1117.R.20190822.1712.004.html>

**引用本文:** 宋亚男, 金昕晔, 张颖, 等. 基于集成学习的 2 型糖尿病患者降糖药用药方案智能分类探讨 [J]. 解放军医学院学报, 2019, 40 (8): 719-724.

## Intelligent classification of hypoglycemia treatment plan for patients with type 2 diabetes based on ensemble learning

SONG Ya'nan<sup>1</sup>, JIN Xinye<sup>2</sup>, ZHANG Ying<sup>3</sup>, CHEN Kang<sup>2</sup>, YING Jun<sup>1</sup>, XUE Wanguo<sup>1</sup>, MU Yiming<sup>2</sup>

<sup>1</sup>Medical Big Data Center, Chinese PLA General Hospital, Beijing 100853, China; <sup>2</sup>Department of Endocrinology, the First Medical Center, Chinese PLA General Hospital, Beijing 100853, China; <sup>3</sup>Department of Ophthalmology, the First Medical Center, Chinese PLA General Hospital, Beijing 100853, China

Corresponding author: MU Yiming, Email: muyiming@301hospital.com.cn

**Abstract: Objective** To apply Adaboost in the determination of hypoglycemia treatment plan in patients with type 2 diabetes. **Methods** Clinical data about 3 005 patients with type 2 diabetes hospitalized in the first medical center of Chinese PLA General Hospital from 2013 to 2017 were collected, including medical prescriptions, biochemical testing results, clinical manifestations, demographic characteristics, etc. Adaboost algorithm was used to establish the machine learning model and classify the treatment plan of the patients, with 1 697 cases as training set and 1 308 cases as testing set randomly, and then accuracy and Kappa coefficient of the model were computed. **Results** The prediction accuracy of the model by Adaboost was 64.2% and the Kappa coefficient was 0.36. After analyzing the model established by Adaboost, we found that UCr, HbA1c, CK-MB, FBG, etc. were significantly related to the treatment plan selecting. **Conclusion** To some extent, Adaboost algorithm is feasible and accurate in predicting hypoglycemia treatment plan.

**Keywords:** type 2 diabetes; hypoglycemic drugs; Adaboost algorithm; mult-class classification learning

**Cited as:** Song YN, Jin XY, Zhang Y, et al. Research of intelligent classification of hypoglycemic drugs program for patients with type 2 diabetes based on ensemble learning [J]. Acad J Chin PLA Med Sch, 2019, 40 (8): 719-724.

糖尿病是严重威胁全球卫生健康的慢性非传染性疾病之一。近 30 年, 我国糖尿病患病率显著增加, 成人患病率高达 9.7%, 糖耐量异常人群达到 15.5%。糖尿病中又以 2 型糖尿病发病人数最

多, 占糖尿病患者的 90% 以上<sup>[1]</sup>。糖尿病目前尚不能彻底根治, 应用降糖药物是主要的治疗手段, 合理的用药方案可以有效控制患者血糖水平, 降低并发症的发病率。与 1 型糖尿病相比, 2 型糖尿病的降糖药用药方案更加复杂<sup>[2]</sup>。随着糖尿病发病机制研究的深入, 越来越多的新药以及组合用药方式不断涌现, 增加了医生决策用药的难度。因此, 降糖药物的个体化选择、精准用药都是目前糖尿病患者和医生所面临的重要问题。由于 2 型糖尿病患者的临床数据结构复杂、样本量大, 且包含大量缺失或冗余信息, 传统的医学统计方法

收稿日期: 2019-04-08

基金项目: 北京市科学技术委员会重大项目 (D141107005314004); 解放军总医院医疗大数据研发项目 (2017MBD-020)

Supported by Beijing Municipal Science and Technology Commission (D141107005314004)

作者简介: 宋亚男, 女, 硕士, 助理工程师。研究方向: 机器学习、医疗大数据分析。Email: 15204163159@163.com

通信作者: 母义明, 男, 主任医师, 教授, 博士生导师。Email: muyiming@301hospital.com.cn

容易陷入各种各样的困境<sup>[3-7]</sup>。因而基于大样本数据挖掘技术对糖尿病用药模式进行分析具有较大的现实意义。Liu 等<sup>[8]</sup>在 2012 年研究了 2 型糖尿病的用药推荐, 首先用 K-近邻算法进行特征选择, 当推荐用药时, 先将病例数据用特征选择结果确定的 38 个特征向量表示, 然后计算欧氏距离来衡量病例与训练数据之间的相似性, 并在训练数据中找到 K 个最相似病例, 最后在此 K 个处方中选择最常见的用药进行推荐。结果显示, 在 484 个测试病例中, 有 89.4% 的用药推荐得到了临床医生认可, 证明了方法的有效性。但是, 基于欧氏距离的相似性度量对于不平衡样本分类敏感, 即当一类样本容量很大, 而其他类样本容量很小时, 会导致小类样本被误判为大类样本。Adaboost 算法作为集成学习的经典算法, 是一种增强型的机器学习方法, 其主要优势在于打破了样本原有的分布, 重新采样使分类器更多地关注难学习的小类样本<sup>[9]</sup>。因其在不平衡样本的分类问题中表现出优秀的预测性能, 使得 AdaBoost 被广泛应用于解决各类医疗问题<sup>[10-12]</sup>。因此, 本研究将利用 Adaboost 算法对大样本量的糖尿病临床数据进行分析, 研究临床用药与复杂的生理指标的相关性, 实现降糖药用药方案的分类预测。

## 资料和方法

**1 资料** 选取解放军总医院第一医学中心 2013 - 2017 年 2 型糖尿病住院患者的病例资料, 纳入标准: 1) 出院诊断为 2 型糖尿病或有 2 型糖尿病并发症; 2) 住院信息及实验室检查结果相对完整, 数据缺失小于总量的 30%; 3) 治疗符合表 1 中②、③、⑤、⑧四种用药方案。

**2 用药方案分类** 依据《中国 2 型糖尿病防治指南 (2017 年版)》的分类并结合医生的临床用药经验<sup>[13]</sup>, 将降糖药用药方案分为 8 种, 分别为: ① GLP-1 受体激动剂, ② 基础胰岛素 + 餐时胰岛

素, ③ 预混胰岛素, ④ 餐时胰岛素, ⑤ 基础胰岛素 + 口服药, ⑥ 单药口服, ⑦ 双药口服, ⑧ 多药口服 (表 1)。其中, 餐时胰岛素包括短效和速效胰岛素, 基础胰岛素包括长效和中效胰岛素, 口服药包括格列奈类、磺脲类、噻唑烷二酮类和双胍类。其中②、③、⑤、⑧类别, 在治疗过程中容易混淆, 存在决策难度。且这 4 种用药方案的例数占总数的 68%, 意味着其在临床上的使用频次较高, 对这 4 种用药方案进行正确鉴别, 将对临床用药决策提供较大帮助。因此, 本文仅采用②、③、⑤、⑧用药方案的病例作为分类建模的数据集。

**3 数据预处理** 1) 数据集进行缺失值处理。由于临床数据的不完整性, 存在大量特征指标缺失的情况。有些指标的数据缺失量虽然较大, 但却是风险预测模型的重要因素, 不能轻易删除, 需要进行缺失值插补处理。通过整理数据发现, 数据集中的缺失部分主要分布在生化检验变量中, 而生化检验指标都是连续变量, 因此按照 4 组类别的不同分布情况, 使用分组均值填补缺失值。2) 对数据做规范化处理。将性别、并发症等字符型变量转化为数值变量, 并对整体数据集进行标准化处理, 以消除指标之间的量纲和取值范围差异的影响。3) 由于样本数量的不均衡性, 本研究采用分层抽样的方法, 从 4 类样本各随机抽取 40%、60%、80% 和 90% 作为训练数据集, 其余的作为测试数据集。

**4 建模方法** 本文采用 Adaboost 算法建立机器学习模型, 对 2 型糖尿病降糖药用药方案进行分类决策。Adaboost 算法可应用多种基分类器构建模型, 本文选择 CART 分类器作为基分类器进行分析。具体的建模过程如下: 1) 初始时令训练样本集中所有样本的权重相等, 表示每个样本被基分类器选入作为训练子集的概率相同; 2) 在第 k 次迭代时, 根据样本权重选取样本点组成训练子集, 训练基分类器, 并用对所有样本进行分类, 提高被错误

表 1 降糖药用药方案分类  
Tab. 1 Classification of medication schemes for hypoglycemic drugs

Index	Name of classification	Description
①	GLP-1 receptor agonists	All prescriptions for GLP-1 receptor agonists
②	Basal insulin + dietary insulin	Basal insulin + short-acting or fast-acting insulin
③	Premixed insulin	Premixed insulin
④	Dietary insulin	Short-acting or fast-acting insulin
⑤	Basal insulin + oral anti-diabetic drug	(Short-acting insulin + isophane insulin) + oral anti-diabetic drug
⑥	One oral anti-diabetic drug	Take one oral anti-diabetic drug alone
⑦	Two oral anti-diabetic drugs	Combinations of two oral anti-diabetic drugs
⑧	Multiple oral anti-diabetic drugs	More than two oral anti-diabetic drugs

分类的样本权重，降低被正确分类的样本权重；3)用更新过权重的样本集生成下一个训练子集，训练下一个基分类器。最后，用迭代产生的多个基分类器对样本进行分类，并对分类结果进行带权重的多数投票，得到最终结果。根据 Adaboost 算法生成的强决策树分类模型，可以得到参与训练的数据集中各变量的重要性评分，评分原理是综合考虑最终的决策分类器的权重和变量的 Gini 指数，计算每个变量对分类结果的影响。评分越高，代表变量的重要性越大。

5 性能评价 本文通过计算准确率 Accuracy 和 Kappa 系数来评价模型的性能。准确率的计算方法为被正确预测的样本数除以所有的样本数，即模型的预测准确率。Kappa 系数是评价模型一致性的指标，Kappa 系数越高，代表一致性越好，意味着模型的性能更佳。其计算方法为：

$$Kappa=(p_0-p_c)/(1-p_c) \quad ①$$

其中， $p_0$  为每一类正确分类的样本数量之和除以总样本数。假设每一类的真实样本个数分别为  $a_1, a_2, \dots, a_c$ ，而预测出来的每一类的样本个数分别为  $b_1, b_2, \dots, b_c$ ， $n$  为测试样本总数， $c$  为分类的类别总数，则

$$p_c=(\sum_{i=1}^c a_i \times b_i)/(n \times n) \quad ②$$

为更好地说明 Adaboost 算法对四类用药方案分类预测的有效性，将采用单一的 CART 决策树模型和 Adaboost 算法生成的强分类器模型，分别对测试样本进行分类，对比两种模型的预测结果及模型性能。

### 结 果

1 纳入病例一般情况 本研究共纳入 3 005 例 2 型糖尿病患者，平均年龄 ( $53.2 \pm 12.0$ ) 岁，其中男性 1 922 例，女性 1 083 例。四种用药方案共包含的样本例数如表 2 所示，采用多药口服的病例数最少，仅有 245 例。纳入研究的特征变量如表 3 所示，主要包括诊断记录、用药记录、生化检验数据、体征数据、人口统计学数据等共计 51 项。

表 2 四种用药方案的样本数

Tab. 2 Quantity of samples for four medication schemes (n=3 005)

Index	Name of classification	Quantity of samples (n)
Class 1	Basal insulin + Dietary insulin	1 439
Class 2	Premixed insulin	769
Class 3	Basal insulin + Oral medicine	552
Class 4	multiple oral anti-diabetic drugs	245

2 建模数据构成 采用分层抽样方法抽取建模数据，训练数据集共 1 697 例，测试数据集共 1 308 例，见表 4。

3 模型预测准确率及一致性 利用训练数据集建立 Adaboost 分类模型，然后将测试数据集输入训练好的模型中得到分类预测结果，根据预测结果的混淆矩阵计算分类准确率和 Kappa 系数。两个模型输出的混淆矩阵如表 5、表 6 所示，在四类用药方案的测试数据集中，单一 CART 决策树模型正确识别的样本数分别为 620、136、59、4；而 Adaboost 模型正确识别的样本数为 625、138、72、5。由此可知，Adaboost 算法生成的强分类器模型对四类用药方案的识别准确率均有所提高。根据以上混淆矩阵，分别计算得两个模型的准确率和 Kappa 系数，即 Adaboost 模型准确率为 64.2%，Kappa 系数为 0.36；单一 CART 决策树模型准确率为 62.6%，Kappa 系数为 0.32。可见，Adaboost 模型的总体准确率和 Kappa 系数均优于 CART 模型，进一步证明了 Adaboost 在 2 型糖尿病降糖用药方案决策支持方面的有效性。

4 变量重要性排序 前 20 个变量的重要性排序如图 1 所示，尿肌酐 (urinary creatinine, UCr)、糖化血红蛋白 (glycosylated hemoglobin, HbA1c)、肌酸激酶同工酶 (creatine kinase isoenzyme, CK-MB) 的重要性评分排在前三位，是影响降糖用药的关键性指标。此外，游离 T4、游离 T3、肌酸激酶 (creatine kinase, CK)、血清白蛋白 (albumin, ALB)、空腹血糖水平 (fasting blood glucose, FBG) 等也是重要影响因素。

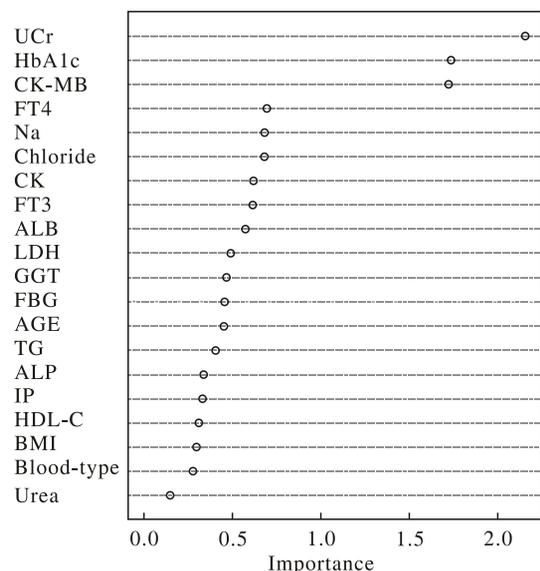


图 1 Adaboost 模型的前 20 个重要变量 Fig.1 The top 20 important variables in Adaboost model

表3 人口统计学与临床数据分析  
**Tab. 3 Demographic and clinical characteristics of samples using different medication schemes**

Variable	All	Class 1	Class 2	Class 3	Class 4	Missing rate (%)
Demographic variable						
Age (yrs)	54.07 ± 13.74	52.7 ± 14.89	55.65 ± 12.21	54.31 ± 12.57	56.62 ± 12.91	0.00
Sex (n, %)						
male	1 922(64.0)	903(62.8)	480(62.4)	384(69.6)	155(63.3)	0.00
female	1 083(36.0)	536(37.2)	289(37.6)	168(30.4)	90(36.7)	0.00
Vital signs						
BMI	26.23 ± 3.41	25.95 ± 3.75	26.51 ± 2.93	26.46 ± 3.15	26.49 ± 3.17	0.57
Biochemical indicator						
HbA1c (%)	8.76 ± 2.03	9.48 ± 2.20	7.97 ± 1.61	8.5 ± 1.59	7.75 ± 1.49	5.99
Tbil (μmol/L)	10.9 ± 4.60	10.74 ± 4.65	10.89 ± 4.38	11.2 ± 4.77	11.15 ± 4.57	1.03
Dbil (μmol/L)	3.17 ± 1.47	3.09 ± 1.49	3.15 ± 1.43	3.39 ± 1.49	3.25 ± 1.42	1.10
ALP (U/L)	70.25 ± 20.76	73 ± 22.18	67.61 ± 19.06	67.04 ± 18.13	69.89 ± 21.02	1.60
Urea (mmol/L)	5.69 ± 2.02	5.8 ± 2.22	5.88 ± 2.06	5.37 ± 1.54	5.20 ± 1.41	1.43
GGT (U/L)	30.46 ± 21.67	30.5 ± 21.93	28.71 ± 20.94	31.7 ± 20.97	32.92 ± 23.57	3.19
TG (mmol/L)	1.91 ± 1.24	1.94 ± 1.32	1.80 ± 1.14	2.01 ± 1.20	1.84 ± 1.06	2.70
SUA (μmol/L)	314.54 ± 88.45	310.96 ± 93.73	317.44 ± 83.91	314.38 ± 77.80	326.66 ± 92.24	0.47
LDH (U/L)	152.91 ± 31.28	155.62 ± 32.72	154.29 ± 30.88	146.03 ± 28.34	148.45 ± 27.53	2.73
Na (mmol/L)	140.96 ± 3.05	140.29 ± 3.34	141.65 ± 2.67	141.32 ± 2.63	141.86 ± 2.31	0.63
Chloride (mmol/L)	102.22 ± 3.42	101.61 ± 3.79	103.05 ± 3.00	102.43 ± 2.90	102.65 ± 2.78	0.57
P (mmol/L)	1.22 ± 0.19	1.22 ± 0.20	1.22 ± 0.19	1.22 ± 0.16	1.19 ± 0.17	0.77
HDL-C (mmol/L)	1.06 ± 0.29	1.06 ± 0.31	1.08 ± 0.29	1.01 ± 0.25	1.05 ± 0.27	0.70
LDL-C (mmol/L)	2.78 ± 0.91	2.82 ± 0.95	2.76 ± 0.88	2.76 ± 0.87	2.68 ± 0.78	0.60
CK isoenzyme (U/L)	15.28 ± 4.67	15.58 ± 4.67	15.33 ± 4.60	14.43 ± 4.68	15.33 ± 4.70	15.67
GLU (mmol/L)	9.33 ± 4.06	10.53 ± 4.63	8.08 ± 3.13	8.81 ± 3.35	7.62 ± 2.11	1.50
FT3 (pmol/L)	15.36 ± 2.36	15.46 ± 2.41	15.01 ± 2.39	15.53 ± 2.25	15.43 ± 2.13	6.52
FT4 (pmol/L)	4.43 ± 0.61	4.33 ± 0.66	4.49 ± 0.58	4.57 ± 0.52	4.52 ± 0.51	6.36
UCr (mmol/L)	54.81 ± 101.22	71.85 ± 118.00	50.32 ± 95.99	31.22 ± 63.93	25.16 ± 51.60	19.07
Complication (n, %)						
Diabetic macroangiopathy						
No (0)	2 136(71.1)	1 035(71.9)	541(70.4)	398(72.1)	162(66.1)	0.00
Yes (1)	869(28.9)	404(28.1)	228(29.6)	154(27.9)	83(33.9)	0.00
Diabetic nephropathy						
No (0)	2 282(75.9)	1 033(71.8)	562(73.1)	472(85.5)	215(87.8)	0.00
Yes (1)	723(24.1)	406(28.2)	207(26.9)	80(14.5)	30(12.2)	0.00
Diabetic retinopathy						
No (0)	2 290(76.2)	1 041(72.3)	567(73.7)	469(85.0)	213(86.9)	0.00
Yes (1)	715(23.8)	398(27.7)	202(26.3)	83(15.0)	32(13.1)	0.00
Diabetic peripheral neuropathy						
No (0)	2 104(70.0)	993(69.0)	514(66.8)	425(77.0)	172(70.2)	0.00
Yes (1)	901(30.0)	446(31.0)	255(33.2)	127(23.0)	73(29.8)	0.00
Hypertensive						
No (0)	1 486(49.5)	734(51.0)	349(45.4)	291(52.7)	112(45.7)	0.00
Yes (1)	1 519(50.5)	705(49.0)	420(54.6)	261(47.3)	133(54.3)	0.00
Hyperlipidemia						
No (0)	1 656(55.1)	785(54.6)	440(57.2)	298(54.0)	133(54.3)	0.00
Yes (1)	1 349(44.9)	654(45.4)	329(42.8)	254(46.0)	112(45.7)	0.00

表 4 建模样本构成  
Tab. 4 Composition of samples

Type of data set	Class 1	Class 2	Class 3	Class 4	Total
Training dataset	575	461	441	220	1 697
Testing dataset	846	308	111	25	1 308

表 5 单一 CART 决策树模型的混淆矩阵  
Tab. 5 Prediction result of CART model

Predict	True				Total
	1	2	3	4	
1	620	101	31	10	763
2	81	136	20	4	243
3	161	70	59	7	300
4	2	1	1	4	12
Total	864	308	111	25	1 308

表 6 Adaboost 模型的混淆矩阵  
Tab. 6 Prediction result of Adaboost model

Predict	True				Total
	1	2	3	4	
1	625	86	25	8	745
2	81	138	14	5	240
3	156	83	72	7	321
4	2	1	0	5	12
Total	864	308	111	25	1 308

## 讨 论

本研究针对 2 型糖尿病患者临床用药选择多而复杂的问题, 在样本分布不均衡的情况下, 创新性地提出将 Adaboost 算法应用于降糖药用药方案的分类预测中。通过建立 Adaboost 实现了对四种用药方案的分类, 预测准确率达到 64.2%, 优于单一 CART 模型, 证明了 Adaboost 算法对降糖药用药方案分类识别的有效性, 可以为糖尿病的临床用药提供参考。

Adaboost 算法最早是由 Freund 等在 20 世纪 90 年代提出的集成学习算法, 被广泛应用于人脸检测、车辆识别、疾病预测、故障诊断等问题中<sup>[14-18]</sup>。Adaboost 的核心思想是将多个弱分类器, 通过加权组合的方式组合成强分类器。Adaboost 算法的建模过程不同于传统的分类器, 它打破了已有的样本分布, 通过调整样本的采样权重, 将易错分的样本权重提高, 使分类器更多地关注难学习的样本。尤其在多分类问题中, Adaboost 算法能提高决策树对每个类别的识别能力, 从而在一定程度上提高模型整体的准确率和一致性。因此, Adaboost 算法更适用于临床的多分类预测问题, 如用药方案预测、几种相似疾病的鉴别诊断等。

基于 AdaBoost 算法获得了与降糖药用药决策相关的影响因素排序, 本文从 51 项指标中选出重要性评分较高的前 20 个影响因素。其中, 血糖相关的指标如糖化血红蛋白、空腹血糖水平的重要性评分都比较高, 表明血糖水平对用药方案的决策起着关键性作用。其次, 肾功能指标如尿肌酐、尿素和肝功能指标  $\gamma$ -谷氨酰基转移酶重要性排名也比较靠前, 说明肾功能和肝功能对降糖药的选择也具有重要影响。另外, 心功能检查指标如肌酸激酶同工酶、肌酸激酶、乳酸脱氢酶等, 也对降糖药用药方案的分类预测起到相对重要的作用, 说明一些心脏疾病对降糖药的用药决策也具有重要价值。除此之外, 年龄、BMI、三酰甘油、甲状腺功能指标游离 T3、游离 T4 等对模型分类也有一定贡献。上述指标都与 2 型糖尿病的血糖控制情况密切相关, 与《中国 2 型糖尿病防治指南(2017 年版)》中提出的临床用药策略基本一致<sup>[13]</sup>。且本文得到的影响降糖药用药决策的重要因素, 也与国内外一些相关研究的结果相符, 侧面证明了 Adaboost 算法的有效性, 进一步证明了该模型在临床上的使用价值。郭立新<sup>[19]</sup>提出肝肾功能减退的老年患者, 药物的代谢和排泄能力减慢, 应慎用磺酰脲类药物, 且应用双胍类药物的患者应定期检查肝肾功能, 这说明肝肾功能会影响降糖药的选择; 蔡林江和沈菊<sup>[20]</sup>指出二甲双胍是为数不多的能够在对抗糖尿病的同时又不会给心脏带来负担的药物, 且二甲双胍和胰岛素联合作用可以有效治疗和预防 2 型糖尿病并发心血管疾病的发生, 说明在选用降糖药物时也要充分考虑患者的心脏功能。

本研究尚存在以下不足: 由于数据本身的不完整性, 在进行数据清洗和预处理过程中难免会引入噪声数据, 使得模型在拟合时存在一定误差, 最终影响模型的精准度。此外, Adaboost 虽然在一定程度上提高了模型准确率, 但模型整体的准确率和 Kappa 系数并不具有明显优势, 分析其主要原因可能是临床用药选择不具有唯一性, 同一个患者的药物治疗方案选择可能不止一种, 而本研究的分类预测模型输出结果仅有一个, 即经模型计算后得到的分类概率最高的类别, 因此误判率较高。在后续的研究中, 将进一步优化用药方案的分类, 改进分类模型算法, 并输出每种用药方案的预测概率供医生参考, 进一步评估模型预测结果的有效性。

## 参考文献

- 1 王丽敏. 2型糖尿病治疗现状调查分析 [D]. 郑州: 郑州大学, 2012.
- 2 Donsa K, Spat S, Beck P, et al. Towards Personalization of Diabetes Therapy Using Computerized Decision Support and Machine Learning: Some Open Problems and Challenges [M] // Smart Health. Springer International Publishing, 2015: 237-260.
- 3 Kasemthaweasab P, Kurutach W. Association analysis of diabetes mellitus (DM) with complication states based on association rules [C]. Conference: Industrial Electronics and Applications, 2012.
- 4 Li P, Chen K, Nie Y, et al. Association of obesity with glucose, blood pressure, and lipid goals attainment in patients with concomitant diabetes and hypertension [J]. *Curr Med Res Opin*, 2015, 31 (9): 1623-1631.
- 5 Jelinek HF, Yatsko A, Stranieri A, et al. Novel data mining techniques for incomplete clinical data in diabetes management [J]. *British Journal of Applied Science & Technology*, 2014, 4 (33): 4591-460.
- 6 Habibi S, Ahmadi M, Alizadeh S. Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining [J]. *Glob J Health Sci*, 2015, 7 (5): 304-310.
- 7 Gregori D, Petrinco M, Bo S, et al. Using data mining techniques in monitoring diabetes care. The simpler the better? [J]. *J Med Syst*, 2011, 35 (2): 277-281.
- 8 Liu H, Xie G, Mei J, et al. An efficacy driven approach for medication recommendation in type 2 diabetes treatment using data mining techniques [J]. *Stud Health Technol Inform*, 2013, 192: 1071.
- 9 曹莹, 苗启广, 刘家辰, 等. AdaBoost 算法研究进展与展望 [J]. *自动化学报*, 2013, 39 (6): 745-758.
- 10 代晓彤, 谢学勤, 康晓平, 等. 基于 AdaBoost 和分类树的北京市高血压患者就诊机构选择的影响因素分析 [J]. *中国卫生统计*, 2017, 34 (1): 23-26.
- 11 王莉莉, 付忠良, 陶攀, 等. 基于主动学习不平衡多分类 AdaBoost 算法的心脏病分类 [J]. *计算机应用*, 2017, 37 (7): 1994-1998.
- 12 古万荣, 谢贤芬, 何亦琛, 等. 基于 AdaBoost 算法的药物-靶向蛋白作用预测算法 [J]. *生物医学工程学杂志*, 2018, 35 (6): 113-120.
- 13 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南 (2017 年版) [J]. *中国实用内科杂志*, 2018, 38 (4): 292-344.
- 14 Bo W, Ai H, Chang H, et al. Fast rotation invariant multi-view face detection based on real Adaboost [C]. *IEEE International Conference on Automatic Face & Gesture Recognition*, 2004.
- 15 Yan G, Yu M, Yu Y, et al. Real-time vehicle detection using histograms of oriented gradients and AdaBoost classification [J]. *Optik-International Journal for Light and Electron Optics*, 2016, 127 (19): 7941-7951.
- 16 Nayak DR, Dash R, Majhi B. Brain MR image classification using two-dimensional discrete wavelet transform and AdaBoost with random forests [J]. *Neurocomputing*, 2015, 177 (C): 188-197.
- 17 Islam A, Reza S, Iftekharuddin K. Multifractal texture estimation for detection and segmentation of brain tumors [J]. *IEEE Trans Biomed Eng*, 2013, 60 (11): 3204-3215.
- 18 Fu Q, Jing B, He P, et al. Fault Feature Selection and Diagnosis of Rolling Bearings Based on EEMD and Optimized Elman\_AdaBoost Algorithm [J]. *IEEE Sensors Journal*, 2018, 18 (12): 5024-5034.
- 19 郭立新. 老年糖尿病降糖药物选择及安全性评估 [J]. *中国实用内科杂志*, 2008, 28 (4): 243-245.
- 20 蔡林江, 沈菊. 二甲双胍和胰岛素联合药物治疗和预防 2 型糖尿病并发心血管疾病的临床疗效 [J]. *中国医疗前沿*, 2012, 7 (16): 17-18.