计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(1):219−224 [doi: 10.15888/j.cnki.csa.006150]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于大规模不平衡数据集的糖尿病诊断研究①

魏　勋, 蒋　凡

(中国科学技术大学 计算机学院, 合肥 230022)

摘　要: 随着发病率的逐年上升, 糖尿病正日益成为严峻的世界健康难题, 尤其是在发展中国家, 其中大部分的糖尿病患者是 2 型糖尿病. 经过科学验证: 通过及时有效的诊断, 大约 80% 的 2 型糖尿病并发症能被阻止或者延缓. 基于大规模不平衡数据集, 提出一种集成模型用于精准地诊断糖尿病患者. 数据集包含了中国某省从 2009 年到 2015 年数百万人的医疗记录. 实验结果证明该方法具有良好的性能, 并取得了 91.00% 的敏感度, 58.24% 的 F_3 值以及 86.69% 的 G-mean 值.

关键词: 糖尿病诊断; 大规模数据集; 不平衡数据集; 集成模型

引用格式: 魏勋,蒋凡.基于大规模不平衡数据集的糖尿病诊断研究.计算机系统应用,2018,27(1):219–224. http://www.c-s-a.org.cn/1003-3254/6150.html

Diabetes Diagnosis Research Based on Large-Scale Imbalanced Dataset

WEI Xun, JIANG Fan

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230022, China)

Abstract: Diabetes is becoming a more and more serious health challenge worldwide with the yearly rising prevalence, especially in developing countries, where the vast majority of diabetes are type 2 diabetes. Scientific research has proved that about 80% of type 2 diabetes complications can be prevented or delayed by timely detection. In this study, we propose an ensemble model to precisely diagnose the diabetes in a large-scale and imbalance dataset. The dataset used in our work covers millions of people from one province in China ranging from 2009 to 2015, which is highly skew. Results on the real-world dataset prove that our method is promising for diabetes diagnosis with a high sensitivity, F_3 and G-mean, i.e., 91.00%, 58.24%, 86.69%, respectively.

Key words: diabetes diagnosis; large-scale dataset; imbalanced dataset; ensemble model

糖尿病是一种慢性非传染性疾病, 主要包括 1 型糖尿病, 2 型糖尿病和妊娠型糖尿病. 其中超过 90% 的患者为 2 型糖尿病. 如果缺乏良好的干预和治疗, 糖尿病患者有一定风险患上一系列并发症, 进而影响健康甚至危及生命. 并发症主要有致盲, 肾衰竭, 心脑血管疾病, 中风以及截肢等. 正是由于这些严重的并发症, 糖尿病已经成为全球第四大致死疾病.

在过去几十年中, 糖尿病发病率逐渐上升[1]. 根据 WHO 估计, 2014 年全球约有 4.22 亿糖尿病患者, 而在 1980 年这个数字仅为 1.08 亿. 在过去十年中, 相比高收入国家, 糖尿病在低收入和中等收入国家的发病率上升更加迅速. 例如, 在 2015 年中国拥有全世界最庞大的糖尿病患者群体, 高达 1.1 亿人之多. 绝大多数患者是 2 型糖尿病, 主要是由肥胖 (特别是腹部肥胖), 缺乏锻炼以及不健康饮食导致[2]. 在某些国家, 大约 50% 到 80% 的糖尿病患者从不关心他们的身体状况, 除非出现严重的并发症. 考虑到这种情况, 早期的诊断显得十分必要且有意义[3].

最近研究指出, 通过及时的筛查诊断, 大约 80% 的 2 型糖尿病并发症能够避免或者延缓[2,3]. 然而单一

① 收稿时间: 2017-04-09; 修改时间: 2017-04-26; 采用时间: 2017-05-02; csa 在线出版时间: 2017-12-22

的临床指标, 如空腹血糖检查, 不具备较高的敏感度, 接近 30% 的糖尿病患者不会被查出[4]. 因此, 智能的数据分析方法, 比如数据挖掘和机器学习技术, 对于精准地诊断糖尿病患者无疑具有很高的价值. 近些年, 已有研究人员应用了一些数据挖掘和机器学习的方法用于糖尿病诊断并取得较好的效果[5-13].

在过去, 收集真实的医疗数据是比较困难的而且相当耗时. 因此, 之前的很多研究中用的数据集主要是来源于规模较小的公开数据集和调查问卷. 随着信息技术的发展和大数据时代的来临, 目前医疗数据的规模变得十分庞大, 能够更好地反映真实情况. 然而, 真实的医疗数据往往存在类别不平衡的问题. 在糖尿病诊断过程中, 由于较低的发病率, 数据集通常是不平衡的, 即健康人群占据大多数, 而糖尿病患者通常只占据很小的比例. 在这种不平衡数据集中, 传统的分类算法往往倾向于忽略少数类样本, 难以有效地诊断出糖尿病患者.

本文提出一种集成模型: xEnsemble, 能够解决类别不平衡问题并精准地诊断糖尿病患者. 该方法基于 EasyEnsemble[14]和 XGBoost[15], 相比其他类似技术, 能够取得更高的敏感度 (Sensitivity), F 值和 G-mean 值. 本文后续内容如下: 首先, 简单介绍类别不平衡问题和常用的解决方法; 然后, 介绍 xEnsemble 方法的基本原理; 接着详细阐述实验过程, 包括数据集介绍、数据预处理过程、性能评估标准、实验设置、实验结果与讨论; 最后, 总结本文并指出进一步的研究方向.

1 类别不平衡问题

类别不平衡, 也就是某些类的样本数量大于其他类别. 在实际生活中, 尤其是在医疗领域, 类别不平衡问题十分常见. 这种情形通常是由较低的发病率导致的. 在某些情况下, 不平衡比例 (多数类样本数量与少数类样本数量之比) 甚至高达 10^6. 在诊断过程中, 如果不平衡数据没有经过适当的处理, 分类器的性能将会受到严重的影响. 例如: 在一个不平衡比例为 99 的数据集中, 即使分类器将所有样本都分类成多数类, 分类器的准确率也能高达 99%, 然而所有少数类样本都被错分. 特别地, 在糖尿病诊断过程中, 类别不平衡会使传统分类算法将大多数的糖尿病患者错误分类成健康人群, 很可能会贻误良好的治疗机会.

目前存在许多方法解决类别不平衡问题. 本文主要集中于两类方法: 代价敏感学习方法与采样方法. 代价敏感学习的一种常用实现方法是权重缩放法 (rescaling), 即通过提高少数类样本的权重来增加少数类被错分的代价. 采样方法是一系列重构样本空间的方法. 采样法有两种基本的实现方法: 欠采样 (under-sampling) 和过采样 (over-sampling). 欠采样通过减少多数类样本来创造一个规模更小的训练集; 过采样则是增加少数类样本, 形成一个规模更大的训练集. 很明显, 这两种方法都能降低不平衡比例, 构建一个更加平衡的训练集. 这两种方式都被证明能够有效地解决类别不平衡问题[16,17]. 欠采样能够缩短训练时间, 然而会忽略潜在有用的信息; 过采样通常需要更长的训练时间, 并且有过拟合的风险[18,19]. 基于欠采样和过采样, 研究者还提出了混合采样[20]和集成采样[14]的方法. 混合采样即同时应用欠采样和过采样的方法; 集成采样则是通过重复的欠采样, 构建若干个平衡训练子集.

本文使用的数据集包含了数百万条记录, 相对于常用的 Pima 公开数据集 (768 条记录), 规模可以算是十分庞大. 考虑到庞大的规模和有限的计算资源, 本文主要关注基于代价敏感学习和欠采样的方法.

2 xEnsemble 方法

为了构建一个高效的糖尿病诊断系统, 首先需要采取适当的措施来解决类别不平衡问题. 欠采样是一种有效的方法, 然而这种方法会丢失大量潜在的有用数据. 而且一次随机选取小规模的多数类样本将会增加样本方差. 众所周知, 一个优秀的分类模型需要同时具备较低的方差和较低的偏差. 所以采样之后, 我们需要一个强力的分类器去尽量拟合新样本来减少偏差. 为了同时满足这两个要求, 我们提出了一种集成模型: xEnsemble. 此方法基于 EasyEnsemble[14]和 XGBoost[15], 伪代码如算法 1 所示. 为方便表示, 本文将少数类样本视为正例, 多数类样本视为负例.

算法1. xEnsemble

1. 输入:
2. P: 正例样本集
3. N: 负例样本集
4. n: 采样子集数量
5. s_i: 每次训练XGBoost模型H_i的迭代次数
6. 步骤:
7. for i=1 to n do
8. 随机从N中采样一个子集N_i, 且$|N_i|=|P|$
9. 使用Ni和P训练Hi, 迭代si次
10. end for
11. 输出:
12. $H(x)=\mathrm{sgn}(\sum H_i-\theta)$

xEnsemble 的主要思想为: 通过重复有放回地对负例样本集采样, 然后与正例样本集合并, 生成 n 个平衡

的训练子集; 在每个训练子集上使用 XGBoost 算法拟合得到一个基分类器 H_i, 这样能够尽量学习负例样本集 N 的各个方面; 最后将所有的基分类器集成起来, 使用投票平均法构成最终的集成分类器 $H(x)$. 明显可以看出, xEnsemble 在上层使用了 Bagging 策略, 此策略被证明能够有效地降低模型方差[21]; 在下层, xEnsemble 使用了基于 Boosting 的方法来尽量拟合训练集, 能够有效地减少偏差. 与 EasyEnsemble 不同的是, xEnsemble 使用投票法来决定类别, 算法 1 中的θ表示集成模型的阈值, 即需要多少票数可以判定某样本为正例. 一般地, 本文将θ设置为 $n/2$. 还有一点明显不同, xEnsemble 采用 XGBoost 代替 EasyEnsemble 中的 AdaBoost 作为集成模型的基分类器. XGBoost 可以并行操作, 而 AdaBoost 只能串行处理, 时间开销相对较大, 不适合用来训练本文规模较庞大的数据集.

XGBoost 是最近非常流行的一种基于树提升 (tree boosting) 的高效机器学习模型. 它的算法实现是基于梯度提升框架 (Gradient Boosting Framework). 它提供了一种在特征粒度上的并行方法, 能够迅速准确地解决许多数据科学问题[1]. 正是由于 XGBoost 的种种优点, 我们将它作为 xEnsemble 的基分类器. xEnsemble 的流程图如图 1 所示.

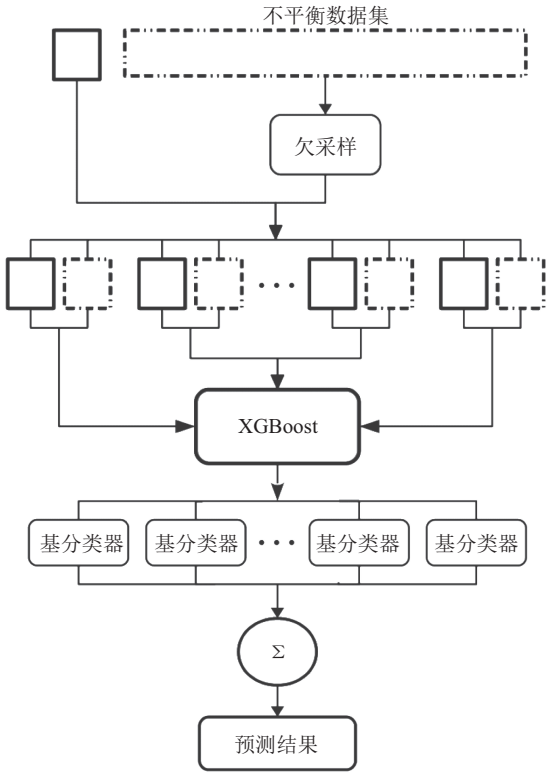

图 1 xEnsemble 示意图

3 实验

3.1 数据集

本文使用的数据集来源于中国某省的卫生部门, 包含了数百万人从 2009 年到 2015 的医疗信息. 原始数据包含三张表: 个人基本信息表, 体检信息表和糖尿病管理信息表. 个人信息表包含了个体的一些基本信息, 比如性别, 出生年月, 家族病史等; 体检信息表包含了个体的一系列医学临床指标, 如身高体重, 血常规, 尿常规, 肾功能检查, 肝功能检查等; 糖尿病管理信息表包含了糖尿病患者每次的随访记录. 其中体检信息表是本文主要使用的数据. 根据医学知识, 我们初步从体检信息表中摘选了 24 项与糖尿病有关的属性. 这 24 项属性的详细信息参见表 1. 而糖尿病管理信息表此处只用来标记某个个体是否患有糖尿病.

表 1 体检信息表中 24 个属性的详细信息

属性	缺失率 (%)	最小值	最大值	平均值	方差
心率	0.25	0	995	74.8	13.5
舒张压	17.41	0	998	80.9	18.4
收缩压	17.38	0	995	127.5	21.4
呼吸频率	0.18	0	858	18.8	5.0
体重	0.70	–1	960	58.4	13.5
腰围	4.17	–1	9955	79.3	34.9
BMI	2.37	–30.1	991.7	22.9	4.9
吸烟量	88.20	0	720	15.9	8.3
饮酒量	87.62	0	650	2.4	3.5
空腹血糖	64.21	–5.5	906	5.6	9.3
随机血糖	99.93	0	502	6.0	8.9
糖化血红蛋白	92.58	–1	990	126.3	30.1
尿微量蛋白	99.93	–1	150	2.1	16.7
谷丙转氨酶	77.19	–1	966	23.4	17.4
谷草转氨酶	80.19	–1	917	26.0	16.9
白蛋白	90.09	0	99	43.8	7.6
总胆红素	78.85	–1.3	9091	14.3	62.3
结合胆红素	93.22	–1	6051	6.5	52.6
血清肌酐	78.67	–14	9808	77.4	89.8
血尿素氮	79.23	–1	902	6.9	18.6
总胆固醇	76.17	–1	836	5.3	12.0
甘油三酯	76.29	–1	5051	1.3	22.9
低密度脂蛋白	88.96	–1	953	2.6	7.5
高密度脂蛋白	88.36	-2.4	444	1.7	5.2

3.2 数据预处理

如表 1 所示, 体检信息表中存在很多“脏数据”, 而且有些属性有较高的缺失率. 在训练模型之前, 我们必须对这些数据进行预处理.

首先, 清洗异常值. 通过查阅相关资料, 确定某个属性的参考范围, 比如收缩压的参考范围为: 90~180, 此后通过两种途径来确定最后的合理范围:

(1) 某些临床指标理论上符合正态分布, 因此在统计意义上, [–3σ, 3σ]区间能覆盖超过 99.7% 的值, 即此区间外的值均视为异常值;

(2) 将初始合理范围外的数据进行分箱操作, 根据每个区域的占比情况确定合理范围.

然后, 对缺失值进行处理. 如表 1 所示, 24 个属性均有不同程度的缺失. 针对这种情况, 缺失率超过 90% 的属性直接忽略, 小于 20% 的属性直接用均值填充, 20%~90% 之间的属性用 SPSS 分析其缺失类型, 发现其缺失相关性很小, 可以认为是完全随机缺失. 一般地, 我们用所有非缺失样本的均值进行填充.

经过预处理之后, 我们最后保留了 24 个特征, 其中 6 个特征来自个人信息表, 分别为: 性别, 年龄, 家族病史 (父亲, 母亲, 兄弟姐妹, 子女); 另外 18 个特征来自体检信息表, 分别为: 心率, 舒张压, 收缩压, 呼吸频率, 腰围, BMI, 吸烟量, 饮酒量, 空腹血糖, 谷丙转氨酶, 谷草转氨酶, 总胆红素, 血清肌酐, 血尿素氮, 总胆固醇, 甘油三酯, 低密度脂蛋白, 高密度脂蛋白. 考虑到疾病之间复杂的联系, 对于家族病史这方面, 我们从简处理: 比如只有当父亲曾经患过糖尿病, 父亲病史才被标记为 1.

我们最初从体检信息表中检索某个个体时间最近的体检记录, 再加上个人基本信息表的 6 个特征, 总共 24 个特征构成样本. 考虑到某些个体在 2009~2015 年之间具有多条体检记录, 如果只是提取其最近的一条体检记录, 无疑会损失大量的信息. 尤其是某些临床指标通常具有较大的波动性, 比如空腹血糖. 因此, 我们针对某个特征额外提取了 3 个相应的新特征: 最大值, 最小值和平均值. 最终我们对 12 个临床指标采用这个操作: 舒张压, 收缩压, 空腹血糖, 谷丙转氨酶, 谷草转氨酶, 总胆红素, 血清肌酐, 血尿素氮, 总胆固醇, 甘油三酯, 低密度脂蛋白, 高密度脂蛋白. 这新增的 3×12=36 个特征, 缺失值也用所有非缺失样本的均值填充. 最后特征数量为: 6+18+36=60.

我们使用 70% 的样本作为训练集, 剩下的 30% 作为测试集. 在所有样本中, 正例只有 56 444 个, 占比 2.9%, 其余为负例. 明显可以看出, 样本存在严重的类别不平衡问题, 不平衡比例为 34.5. 详细情况参见表 2.

表 2 样本情况

样本	数量	比例 (%)
全部	1 908 383	100
训练集	1 335 868	70
测试集	572 515	30
正例	56 444	2.9
负例	1 851 939	97.1

3.3 评价标准

如前所述, 当数据存在类别不平衡问题或者错分代价不一致的时候, 对分类器而言, 错误率并非一个合适的评价标准. 因此, 本文使用 *F* 值和 *G-mean* 值作为分类器性能的评价标准. *F* 值和 *G-mean* 值的计算均基于表 3 所示的混淆矩阵.

表 3 混淆矩阵

	预测正例	预测反例
真实正例	TP	FN
真实反例	FP	TN

考虑到本文中召回率 (recall) 相对精确率 (precision) 更加重要, 我们进一步使用 F_β 来评估性能. 其中 β 值用来衡量召回率相对精确率的重要度. 当 β=1 时, F_β 退化成标准的 F_1 值; 当 β>1 时, 召回率影响更大; 当 β<1 时, 精确率影响更大. 为了尽可能的降低 *FN* 的值, 本文将 β 设置为 3. F_β 和 *G-mean* 的定义如下所示:

$$Sensitivity = \frac{TP}{TP+FN} = recall_+$$

$$Specificity = \frac{TN}{TN+FP} = recall_-$$

$$precision_+ = \frac{TP}{TP+FP}$$

$$precision_- = \frac{TN}{TN+FN}$$

$$F_\beta = \frac{\left(1+\beta^2\right)\times recall_+ \times precision_+}{\beta^2 \times precision_+ + recall_+}$$

$$G-mean = \sqrt{Sensitivity \times Specificity}$$

3.4 实验设置

我们在训练集上使用 5-折交叉验证和网格寻优方法来获得最佳参数. 然后在测试集上运行, 得到最终的 *Sensitivity*, F_β 和 *G-mean* 值. 实验主要分成两个步骤, 第一步解决类别不平衡问题, 第二步为分类. 第一步使用 5 种策略, 第二步使用 6 种分类器, 总共 30 种模型. 5 种用于解决类别不平衡问题的策略如下所述.

1) Original: 原始情况, 不对负例样本进行任何操作, 直接用来训练. 此策略用来作为实验对比.

2) Cost-Sensitive(简称 Cost): 假设不平衡比例为 |N|/|P|, 那么负例与正例的权重比值为|P|/|N|. 通过此设置, 能够显著地提高正例错分代价.

3) Random Under-Sampling(简称 Random): 随机无放回地从负例样本集中采样一个子集, 子集大小和正例样本集大小相同.

4) Edited Nearest Neighbours(简称 ENN): 如果一

个样本的标记同它的 K 个邻居相异, 则将这个样本删除.

5) Ensemble Sampling(简称 Ensemble): 类似 Random Under-Sampling, 此方法随机有放回地从负例样本集中采样 M 次, 生成 M 个和正例样本集大小相同的子集. 考虑到本文所用数据集的不平衡比例为 34.5, 特将 M 设置为 30.

对于第二个步骤, 我们使用 3 个单分类器和 3 个集成分类器. 3 个单分类器分别为: LR, CART, Linear SVC(简称 LSVC); 3 个集成分类器分别为: Adaboost (简称 Ada), Random Forest(简称 RF), XGBoost(简称 XGB). Ada, RF 和 XGB 都是基于 CART 并且弱分类器的数量都设置成 500 个. 在这 6 个分类器中, LR, RF, XGB 能够并行操作而另外 3 个只能串行操作. 除了 XGB 之外, 我们使用 scikit-learn API[22]实现这些分类器. 另外, Ada 和 XGB 不支持设置类别权重, 因此这两个分类器无法在 Cost 策略下运行, 后面用-表示缺失的结果. 在 Ensemble 策略下, EasyEnsemble 使用 Ada 作为基分类器并采用线性加权求和的方法, 而 xEnsemble 使用 XGB 作为基分类器并使用简单的投票法, 另外 4 个分类器也同样使用投票.

我们的实验运行在一台有 24 核 CPU, 主频为 3.0GHz, 内存为 64GB 的服务器上. 整个实验的流程图如图 2 所示.

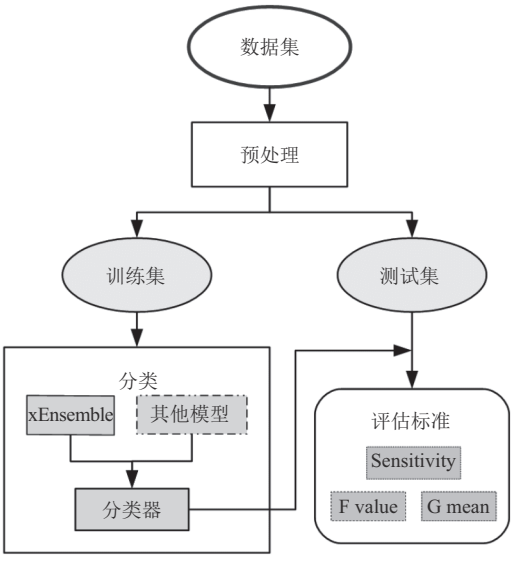

图 2 实验示意图

3.5 实验结果与讨论

表 4、表 5、表 6 分别表示这 30 个模型在测试集上的 *Sensitivity*, F_3 和 *G-mean* 值. 如表 4 所示, 在 Original 策略下, 所有分类器的 *Sensitivity* 指标都有大幅退化. 其中, XGB 取得最高的分数, 证明了其卓越的性能. 由于减少了一些边界上的负例样本, ENN 策略相比 Original 有了一些提高. 更进一步, Cost、Random 和 Ensemble 策略都有大幅度的提高. Random 比 Cost 表现稍强, 尤其是在 LSVC 分类器上. 如前所述, Ada 和 XGB 在 Cost 策略上结果是缺失的因为它们不支持 Cost 策略. 另外, 相比 Random 策略, Ada、RF 和 XGB 在 Ensemble 下表现稍好, 而 LR 和 LSVC 则反之. 从分类器层面来看, RF 和 XGB 的性能几乎是并驾齐驱, 均取得优异的表现.

表 4 所有模型的 *Sensitivity* 值

	Original	Cost	Random	ENN	Ensemble
LR	.2230	.8516	.8547	.3535	.8526
CART	.3037	.8513	.8677	.4062	.8970
LSVC	.1896	.6779	.8566	.3161	.8556
Ada	.2991	-	.8664	.4056	.8688
RF	.3048	**.8663**	**.9141**	.3755	**.9135**
XGB	**.3383**	-	.8992	**.4435**	.9100

表 5 所有模型的 F_3 值

	Original	Cost	Random	ENN	Ensemble
LR	.2389	.5693	.5657	.3671	.5687
CART	.3193	.5628	.5546	.4146	.5616
LSVC	.2079	.5623	.5678	.3226	.5681
Ada	.3124	-	.5673	.4160	.5750
RF	.3223	**.5765**	.5554	.3903	.5752
XGB	**.3506**	-	**.5795**	**.4451**	**.5824**

表 6 所有模型的 *G*-mean 值

	Original	Cost	Random	ENN	Ensemble
LR	.4714	8480	.8472	.5920	.8480
CART	.5495	8453	.8455	.6339	.8556
LSVC	.4350	.7930	.8486	.5606	.8485
Ada	.5454	-	.8507	.6335	.8544
RF	.5511	**.8544**	.8555	.6105	.8645
XGB	**.5787**	-	**.8633**	**.6607**	**.8669**

在表 5 中, Ensemble 在 F_3 上的表现优于 Random 除了 CART 分类器. 另外 Cost 的性能也比 Random 要强, 和 Ensemble 不相上下. 在分类器层面, 尽管 RF 对于 *Sensitivity* 在 Random 和 Ensemble 策略上比 XGB 表现要好, 此处 XGB 对于 F_3 却比 RF 表现更佳.

表 6 的情况更加简洁明了. 很明显, Ensemble 相比其他策略表现更加优秀, XGB 在所有分类器中取得最高的分数. 值得一提的是, xEnsemble 对于 *Sensitivity*, F_3 和 *G-mean* 均比 EasyEnsemble 效果要好.

总之, 集成分类器, 特别是 XGB, 相比单分类器, 性能表现更好. 同时, Ensemble 策略相比其他策略, 取

得更优秀的结果. 因此, 本文提出的方法: xEnsemble, 相比其他方法表现出更良好的性能.

4 结语

本文主要将研究重点放在应用不平衡学习方法来解决数据集中的类别不平衡问题, 然后对糖尿病进行分类诊断. 由于数据集的高度不平衡性, 相比之前的研究, 我们面临一个更加严峻的挑战. 本文提出的 xEnsemble 方法类似于“bagging of boosting”, 能够同时降低模型的方差和偏差. 通过采用该方法, 我们获得了一个较优的结果, 这将协助医务工作人员更高效便捷地对糖尿病诊断做出决策.

提取影响糖尿病发病的关键因素将是本文进一步的研究方向. 明确这些关键发病因素能够起到很好的预警作用, 做到“未雨绸缪”, 帮助那些潜在风险的糖尿病人群更好地管理健康和预防糖尿病的发生.

参考文献

1 World Health Organization. Global report on diabetes. Geneva: World Health Organization, 2016.

2 Tuomilehto J, Lindström J, Eriksson JG, *et al*. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. New England Journal of Medicine, 2001, 344(18): 1343–1350. [doi: 10.1056/NEJM200105033441801]

3 Franciosi M, De Berardis G, Rossi MCE, *et al*. Use of the diabetes risk score for opportunistic screening of undiagnosed diabetes and impaired glucose tolerance. Diabetes Care, 2005, 28(5): 1187–1194. [doi: 10.2337/diacare.28.5.1187]

4 World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: Report of a WHO/IDF consultation. Geneva: World Health Organization, 2006.

5 Huang Y, McCullagh P, Black N, *et al*. Feature selection and classification model construction on type 2 diabetic patients' data. Artificial Intelligence in Medicine, 2007, 41(3): 251–262. [doi: 10.1016/j.artmed.2007.07.002]

6 Goel R, Misra A, Kondal D, *et al*. Identification of insulin resistance in Asian Indian adolescents: Classification and regression tree (CART) and logistic regression based classification rules. Clinical Endocrinology, 2009, 70(5): 717–724. [doi: 10.1111/cen.2009.70.issue-5]

7 Heikes KE, Eddy DM, Arondekar B, *et al*. Diabetes risk calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care, 2008, 31(5): 1040–1045. [doi: 10.2337/dc07-1150]

8 Li L. Diagnosis of diabetes using a weight-adjusted voting approach. Proceedings of 2014 IEEE International Conference on Bioinformatics and Bioengineering. Boca Raton, FL, USA. 2014. 320–324.

9 Dogantekin E, Dogantekin A, Avci D, *et al*. An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing, 2010, 20(4): 1248–1255. [doi: 10.1016/j.dsp.2009.10.021]

10 Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Transactions on Information Technology in Biomedicine, 2010, 14(4): 1114–1120. [doi: 10.1109/TITB.2009.2039485]

11 罗森林, 成华, 顾毓清, 等. 数据挖掘在 2 型糖尿病数据处理中的应用. 计算机工程与设计, 2004, 25(11): 1888–1892. [doi: 10.3969/j.issn.1000-7024.2004.11.007]

12 罗森林, 郭伟东, 张笈, 等, 陈松景. 基于 Markov 的Ⅱ型糖尿病预测技术研究. 北京理工大学学报, 2011, 31(12): 1414–1418.

13 蒋琳, 彭黎. 基于支持向量机的Ⅱ型糖尿病判别与特征筛选. 科学技术与工程, 2007, 7(5): 721–726.

14 Liu XY, Wu JX, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 539–550. [doi: 10.1109/TSMCB.2008.2007853]

15 Chen TQ, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA. 2016. 785–794.

16 Weiss GM. Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 7–19. [doi: 10.1145/1007730]

17 Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63–77. [doi: 10.1109/TKDE.2006.17]

18 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

19 Drummond C, Holte RC. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Datasets II. Washington, DC, USA. 2003.

20 Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, 2004, 6(1): 20–29. [doi: 10.1145/1007730]

21 Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123–140.

22 Buitinck L, Louppe G, Blondel M, *et al*. API design for machine learning software: Experiences from the scikit-learn project. arXiv:1309.0238, 2013.