血液疾病辅助诊断方法研究

2019 年 7 月 3 日

血液样本数据指标项

表: 血常规指标项（正负样本字段）

指标项	统一字段名	指标项	统一字段名
平均血红蛋白含量	MCH	中性粒细胞比率	NEUT_P
平均血红蛋白浓度	MCHC	中性粒细胞计数	NEUT
红细胞平均体积	MCV	血小板压积	PCT
平均血小板体积	MPV	红细胞分布宽度	RDW
嗜碱性粒细胞计数	BAS	淋巴细胞计数	LY
嗜碱性粒细胞比率	BAS_P	淋巴细胞比率	LY_P
嗜酸性粒细胞计数	EOS	单核细胞计数	MONO
嗜酸性粒细胞比率	EOS_P	单核细胞比率	MONO_P
血红蛋白	HB	红细胞计数	RBC
血小板分布宽度	PDW	红细胞压积	HCT
血小板计数	PLT	性别	Sex
白细胞计数	WBC	年龄	Age
诊断结果	Result		

正样本数据

表: 疾病分类汇总数据

疾病分类	英文对照	总数
淋巴瘤	Lymphoma	5202
多发性骨髓瘤 [卡勒病] (M97320/3)		2183
急性髓样白血病		2051
血小板减少性紫癜		1155
急性淋巴细胞白血病		1154
过敏性紫癜 [亨诺克 (－舍恩莱因) 紫癜]		1105
再生障碍性贫血 NOS		458
噬血细胞综合症		374
骨髓增生异常综合征		229
合计		13911

负样本数据

合计：728,599

实验流程

1. 样本比例
   - 1 : 1
   - 1 : 2
   - 1 : 5
   - 1 : 10
2. 训练模型
   - 逻辑回归
   - 随机森林
   - Light GBM
   - XGBoost
   - 神经网络
3. 记录结果参数与过程

Imbalanced Class Distribution(aka. Skewed Class)

- Challenges with Standard Machine Learning Techniques
- Resampling Techniques
- Algorithmic Ensemble Techniques

Resampling Techniques

- Random Under-Sampling: randomly **eliminate majority** class examples
  - 👍 improve run time and storage problems
  - 👎 information loss & biased sample
- Random Over-Sampling: randomly **replicate** instances in the **minority** class
  - 👍 no information loss & Outperforms under sampling
  - 👎 overfitting
- Cluster-Based Over Sampling: K-means clustering
- Informed Over Sampling: Synthetic Minority Over-sampling Technique
  - 👍 Mitigates overfitting & No loss of useful information
  - 👎 SMOTE can introduce additional noise & is not very effective for high dimensional data
- Modified synthetic minority oversampling technique (MSMOTE)
  - Security/Safe samples: k nearest neighbors
  - Border samples: nearest neighbor
  - Latent nose samples: nothing

P vs NP，二分类，Train 1:1 , Test 1:1000, Random Forest

1:1 建模 1:1000 验证结果

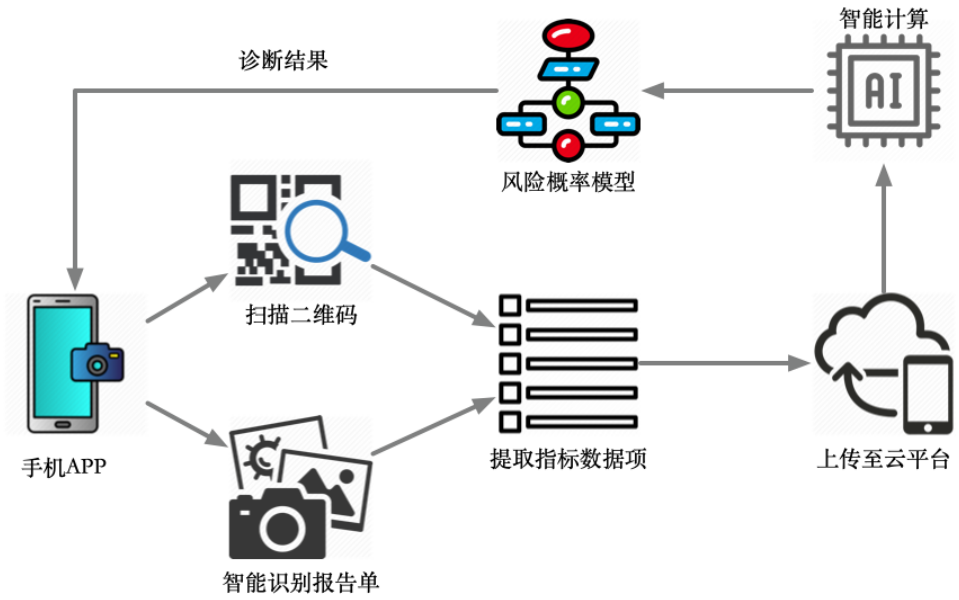
Disease	Cross_val_score	Training_score	Validation_accuracy	AUC	Recall
再生障碍性贫血 NOS	0.984	0.994	0.996	1.000	0.971
噬血细胞综合症	0.935	0.958	0.955	0.992	0.938
多发性骨髓瘤[卡勒病] (M97320/3)	0.929	0.932	0.915	0.983	0.935
急性淋巴细胞白血病	0.941	0.952	0.945	0.995	0.967
急性髓样白血病	0.946	0.949	0.954	0.985	0.945
淋巴瘤	0.922	0.926	0.930	0.980	0.932
血小板减少性紫癜	0.986	0.989	0.984	0.998	0.979
过敏性紫癜[亨诺克(－舍恩莱因)紫癜]	0.971	0.972	0.984	0.996	0.948
骨髓增生异常综合征	0.981	0.984	0.987	0.999	1.000

P vs NP，二分类，Train 1:1，Test 1:1000, XGBoost

xgboost 1:1000 模型				
疾病名称	scores2	test_auc	recall	specificity
淋巴瘤	0.969515	0.994046	0.956438	0.976957
过敏性紫癜[亨诺克(－舍恩莱因)紫癜]	0.97868	0.997832	0.96988	0.987961
血小板减少性紫癜	0.991952	0.999278	0.991354	0.989118
骨髓增生异常综合征	0.978125	0.997625	0.956522	0.989681
再生障碍性贫血 NOS	0.985938	0.999825	0.992754	0.993348
噬血细胞综合症	0.982692	0.99565	0.973451	0.970097
多发性骨髓瘤[卡勒病]	0.97219	0.995496	0.963359	0.972397
急性髓样白血病	0.980121	0.996792	0.980519	0.991218
急性淋巴细胞白血病	0.976998	0.996922	0.988473	0.986277

患病率

疾病	患病数	统计人群数	患病率 (10 万)
噬血细胞综合症	77	5156988	1.493119627
再生障碍性贫血	3316	5156982	64.3011746
血小板减少性紫癜	256	5156990	4.964136056
过敏性紫癜	1075	5156988	20.84550129
骨髓增生异常综合症	315	5156988	6.108216657
急性淋巴细胞白血病	134	5156988	2.598415975
急性髓样白血病	29	5156988	0.562343756
多发性骨髓瘤	727	5156988	14.09737622
淋巴瘤	89	5156988	1.725813595
合计	6018		116.6960978


智能计算
AI
诊断结果
风险概率模型
扫描二维码
手机APP
提取指标数据项
上传至云平台
智能识别报告单