多层学习联合建模方法设计在气阴两虚型咳嗽证候的辨证诊断中的应用*

浙江中医药大学(310053)　项莎特　瞿溢谦　叶含笑△

【提　要】 目的　采用多层学习联合建模方法挖掘气阴两虚型咳嗽的辨证证候，以期为中医学习、研究临床辨证及诊断提供新的思路与方法。方法　联合采用随机森林、XGBoost 及 *logistic* 回归三种机器学习算法，对 767 例咳嗽患者病案，运用 Anaconda 3-5.2.0 软件建立算法模型进行分析。结果　运用该方法所得的证候结果与文献记载的证候表现大体一致，主要为呛咳、乏力、口干、痰少而色白，燥苔、脉弱等证候。经交叉验证得出，XGBoost 算法准确率为 86.7%，随机森林为 85.3%。结论　多层学习联合建模方法可弥补单独使用随机森林、XGBoost 或 *logistic* 回归算法所产生的缺陷，尤其对于临床病案较少的小样本数据更为有效，该方法在一定程度上降低了重要变量丢失的可能性。

【关键词】 机器学习　气阴两虚　咳嗽　证候
【中图分类号】 R256.11　　【文献标识码】 A　　DOI 10.3969/j.issn.1002－3674.2020.06.023

咳嗽为常见疾病，中医治疗该疾病的历史悠久，且对于临床检查无殊的咳嗽的疗效较为显著[1]。气阴两虚为致使原因不明、临床检查无明显病理性改变的咳嗽发作的病机之一。在当今大数据及人工智能的冲击下，探索新的中医学习、传承方式成为趋势[2]。通过机器学习提取判断事物属性特征，从而建立诊断模型的方式，成为了医学领域中蓬勃发展的一项研究[3]。此方法有助于从隐匿和重叠的证候中挖掘出其分布特点，从而预测疾病证型[4]，同时可达到挖掘和学习中医临床辨证及诊断的目的。

因此，本研究气阴两虚型咳嗽患者的基础上，联合多模型机器学习算法，组成多层学习联合建模方法，对所采集气阴两虚型咳嗽的四诊信息进行挖掘和分析，从而为机器学习辅助开发新的中医学习方式提供参考。

资料与方法

1. 数据来源

2018 年 1 月－2018 年 12 月就诊于浙江中医药大学门诊部的咳嗽患者病案。

2. 纳入标准

(1) 疾病名称为咳嗽，证型为气阴两虚；(2) 复诊资料显示，该患者咳嗽症状缓解；(3) 咳嗽的诊断标准同时参照《咳嗽的诊断与治疗指南》(2015)[1]；(4) 所采集的病案包含了完整的四诊信息。

3. 排除标准

(1) 复诊资料显示，咳嗽症状并未缓解者；(2) 未见复诊信息者；(3) 病案中四诊信息不全者；(5) 咳嗽诊断标准不符合《咳嗽的诊断与治疗指南》者。

4. 数据总量

采集咳嗽病例 767 例，其中气阴两虚型咳嗽病例为 564 例，非气阴两虚型咳嗽病例为 203 例。

5. 数据预处理

将采集到的病案按姓名、证候、是否为气阴两虚型咳嗽，建立信息标签。证候名称标准化参照《中医药学名词》[5]，录入 excel 数据表，共得到 210 个证候标签，并对各个证候进行语言规范化处理后，将信息采用“0”“1”变量赋值，是为“1”，否为“0”。对疾病类型同样采用“0”“1”变量赋值，是气阴两虚型咳嗽为“1”，非气阴两虚型咳嗽为“0”。

模型的选择、建立与运用

1. 多层学习联合建模方法

在机器学习的众多算法中，随机森林、XGBoost 算法具有较高的计算效率，且在一定程度上能有效防止模型的过拟合[6-7]，因此，与支持向量机、决策树、logistic 回归等算法相比，随机森林、XGBoost 在疾病预测中具有较高的准确度[8-10]。

随机森林采用集成学习的思想，使用决策树作为弱分类器，组合多个决策树形成具有较好效果的强分类器。该算法的准确率可与 Adaboost 相媲美[11]。采用随机森林模型可以通过计算每棵决策树的袋外数据误差，及对袋外数据所有样本的特征随机加入噪声后的袋外数据误差，得出样本特征变量的重要性[7]。

梯度提升决策树(gradient boosting decision tree，GBDT) 是 XGBoost 的基础算法，它包含一个迭代残差树的集合，每一棵树都在学习前 $N-1$ 棵树的残差，将每棵树预测的新样本输出值相加起来就是样本最终的预测值。不同于常用的梯度提升决策树在优化时仅用一阶导数信息，XGBoost 对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数，使得 XGBoost 得到良好的结果[12]，并在许多机器学习和数据挖掘挑战中得到广泛认可[13]。

在采用机器学习挖掘辨别是否为气阴两虚型咳嗽

* 基金项目：国家自然科学基金面上项目(81673672)
△通信作者：叶含笑，E-mail：yhx@zcmu.edu.cn

的重要证候时，采用随机森林结合 XGBoost 算法进行特征提取。但是，随机森林、XGBoost 均无法得出指标的方向性影响[14]，将两种算法所得的重要证候特征，再次使用 logistic 回归模型进行建模。由此，可得到由多模型组合而成，用于辅助学习中医辨证及诊断的新模型，命名为多层学习联合建模方法。

2. 多层学习联合建模方法的运用

将经过预处理的数据输入 Anaconda 3-5. 2. 0 软件，删除决定性单一输入后，可将 210 个证候特征缩减至 63 个，再经随机森林及 XGBoost 算法的运算，最终根据权重的高低排序，截取经两种算法运算后，各自所得结果的前 35 个证候特征及其相应的权重值。所得的两组证候特征结果中，具有 29 个重合变量。将此 29 个证候特征再次采用 logistic 回归算法进行建模，最终得到辨别气阴两虚型咳嗽的重要证候特征。同时采用 10 折交叉验证法，对算法进行准确性评估。（见图 1）

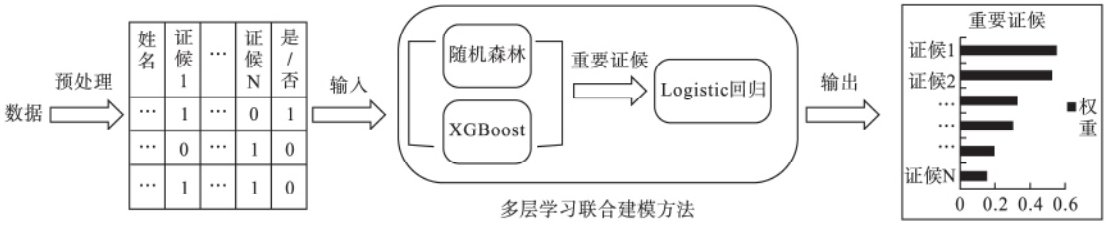

图 1　多层学习联合建模方法流程图

结　果

表 1 和表 2 分别为经随机森林、XGBoost 算法所得气阴两虚型咳嗽证候的前 35 个重要证候。表 3 为重合的 29 个证候经 logistic 回归算法后所得结果，所得权重为正值的变量，归属为气阴两虚型咳嗽的重要辨证证候；所得权重为负值的变量，归属为非气阴两虚型咳嗽的重要辨证证候。

因此，经多层学习联合建模方法运算后，最终得出判定咳嗽属气阴两虚型证型的重要证候有呛咳、乏力、口干、气急、反复咳嗽、咽中痰少、痰色白、自汗、盗汗、头晕、裂纹舌、薄苔或少苔、燥苔、脉弱、脉弦（图 2）。经 10 折交叉验证法，得出 XGBoost 准确度为 86. 7%，随机森林为 85. 3%。

表 1　基于随机森林所得气阴两虚型咳嗽证候的重要性排序

排序	证候	权重	排序	证候	权重
1	乏力	0. 086	19	胸闷	-0. 003
2	呛咳	0. 076	20	盗汗	-0. 003
3	弱脉	0. 067	21	口苦	-0. 004
4	口干	0. 045	22	咳逆倚息	-0. 004
5	咽痛	0. 035	23	少苔	-0. 005
6	黄痰	0. 032	24	入夜咳甚	-0. 005
7	反复咳嗽、	0. 028	25	咽中痰黏	-0. 006
8	鼻塞	0. 016	26	寐浅	-0. 007
9	鼻流清涕	0. 011	27	红舌	-0. 007
10	少痰	0. 011	28	便溏	-0. 007
11	头晕	0. 004	29	迟脉	-0. 008
12	气急	0. 004	30	薄苔	-0. 009
13	自汗	0. 003	31	白痰	-0. 009
14	弦脉	0. 002	32	腻苔	-0. 009
15	裂纹舌	0. 001	33	淡白舌	-0. 010
16	畏风寒	0. 000	34	便秘	-0. 010
17	燥苔	-0. 001	35	胖大舌	-0. 011
18	大便不畅不爽	-0. 001			

表 2　基于 XGBoost 所得气阴两虚型咳嗽证候的重要性排序

排序	证候	权重	排序	证候	权重
1	反复咳嗽	0. 06595	19	咳痰不畅	0. 02454
2	口干	0. 05982	20	燥苔	0. 02454
3	黄痰	0. 05215	21	气急	0. 02301
4	盗汗	0. 04294	22	自汗	0. 02147
5	大便不畅不爽	0. 04141	23	口苦	0. 02147
6	入夜咳著	0. 04141	24	纳减	0. 02147
7	呛咳	0. 03988	25	少痰	0. 01840
8	弱脉	0. 03834	26	少苔	0. 01687
9	乏力	0. 03834	27	白痰	0. 01687
10	咽痛	0. 03681	28	滑脉	0. 01380
11	弦脉	0. 03528	29	多痰	0. 01227
12	头晕	0. 03374	30	薄苔	0. 01074
13	淡白舌	0. 03374	31	腻苔	0. 01074
14	咽干	0. 02914	32	裂纹舌	0. 01074
15	鼻塞	0. 02914	33	红舌	0. 00767
16	咳逆倚息	0. 02914	34	胸闷	0. 00767
17	数脉	0. 02761	35	鼻流清涕	0. 00767
18	便溏	0. 02761			

表 3　基于 logistic 回归模型所得气阴两虚型咳嗽证候的结果

排序	证候	权重	排序	证候	权重
1	呛咳	1. 64667	16	弦脉	0. 15025
2	乏力	1. 55009	17	胸闷	-0. 00900
3	口干	1. 14555	18	腻苔	-0. 08481
4	弱脉	1. 12117	19	咳逆倚息	-0. 16509
5	气急	0. 98055	20	便溏	-0. 45661
6	自汗	0. 85687	21	鼻流清涕	-0. 46874
7	反复咳嗽	0. 83296	22	鼻塞	-0. 57360
8	燥苔	0. 65237	23	淡白舌	-0. 73343
9	少痰	0. 63197	24	大便不畅不爽	-0. 75259
10	红舌	0. 59205	25	入夜咳甚	-0. 82726
11	薄苔	0. 54686	26	口苦	-0. 88314
12	裂纹舌	0. 53108	27	咽痛	-0. 96156
13	盗汗	0. 32336	28	头晕	-1. 05955
14	少苔	0. 29216	29	黄痰	-1. 19161
15	白痰	0. 19781			

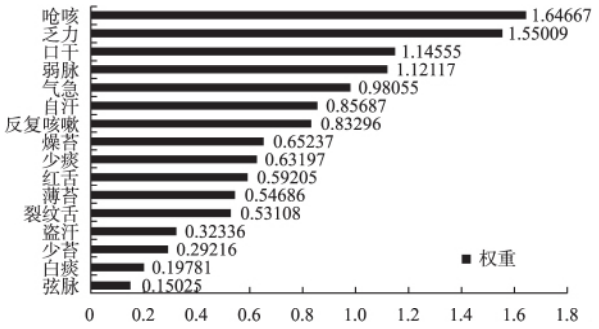

图 2　辨证气阴两虚型咳嗽的重要证候

讨　论

1. 多层学习联合建模方法所得证候与中医相关理论探讨

通过多层学习联合建模方法所得出的气阴两虚型咳嗽的重要证候结果，与临床文献报道[15-18]及《中医病证诊断疗效标准》[19]记载的证候特征大致相同，主要为乏力、口干、反复咳嗽、自汗、盗汗、痰少色白、舌红、苔薄或苔少、苔干燥、脉弱、脉弦。在此基础上，本研究尚且发现呛咳、气急、裂纹舌亦为气阴两虚型咳嗽的表现证候。

2. 多层学习联合建模方法的优势

在机器学习中，由于随机森林在每次划分时，考虑的属性较少，因而该算法在大型数据库上更为有效，但对于小样本，其准确度有所下降[11]。而 XGBoost 对于小样本的特征提取，效果明显优于随机森林、支持向量机、*logistic* 回归等算法[20]。本研究采用随机森林联合 XGBoost 算法，找其重叠证候，既能弥补随机森林对于小样本的准确率较低的缺陷，又可进一步提升 XGBoost 结果的准确性，从而得出较为满意的证候特征结果。此外，在数据维度较大的前提下，仅采用 logistic 回归算法，会使得一些对辨证影响较大的相关证候被丢失，而影响较小的证候反而得到意义[21]，故通过随机森林和 XGBoost 降维，可剔除相关性较小的证候，弥补 logistic 回归算法的不足。

综上所述，本研究所采取的多层学习联合建模方法，具有以下特征：为 3 个模型联合使用的多模型算法；可辅助中医学者学习，研究临床疾病的辨证及诊断；能在临床疾病样本较少，维度较大的情况下，确保较高的准确性。因此，该方法用于挖掘中医疾病证型的重要证候特征值得推广。

3. 多层学习联合建模方法的展望。

该方法虽能确保较好的准确度，但可能仍存在一些对辨证影响较小的证候被提取。因此，在解决该问题时，可考虑优化此模型结构，经 logisitic 回归算法后得出的证候，再经专业人士判定，得出最终较为满意的结果，即采用人机互助的模式，以减弱机器学习刻板化的缺点，增强该模型的灵活性，进一步提升结果的准确性。

参　考　文　献

[1] 中华医学会呼吸病学分会哮喘学组. 咳嗽的诊断与治疗指南(2015). 中华结核和呼吸杂志，2016，39(5)：323-354.

[2] 尚丽丽. 新医科背景下医学研究生教育的思考. 医学研究生学报，2018，31(10)：1078-1081.

[3] 陈建设，陈文垲. 聚类分析结合 logistic 回归分析在中医证候诊断量化研究中的应用探讨. 中国卫生统计，2009，26(4)：379-382.

[4] 曹云. 基于机器学习的胃食管反流病中医证候分类的应用研究. 北京中医药大学，2019.

[5] 中医药学名词审定委员会. 中医药学名词. 北京：科学出版社，2005：58-91.

[6] 李占山，刘兆赓，丁国轩，等. 基于 xgboost 的特征选择算法. 通信学报，2019，40(10)：1-8.

[7] Breiman L. Random forests. Machine Learning，2001，45(1)：5-32.

[8] Liu Z，Zhou T，Han X，et al. Mathematical models of amino acid panel for assisting diagnosis of children acute leukemia. J Transl Med，2019，17(1)：38-38.

[9] Parikh RB，Manz C，Chivers C，et al. Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA Netw Open，2019，2(10)：e1915997.

[10] Hsieh CH，Lu RH，Lee NH，et al. Novel solutions for an old disease：Diagnosis of acute appendicitis with random forest，support vector machines，and artificial neural networks. Surgery，2010，149(1)：87-93.

[11] HAN J. 数据挖掘：概念与技术. 范明等译. 北京：机械工业出版社，2012：249-249.

[12] Friedman J，Tibshirani HR. Special invited paper. Additive logistic regression：A statistical view of boosting：Rejoinder. The Annals of Statistics，2000，28(2)：400-407.

[13] Chen T，Guestrin C. Xgboost：A scalable tree boosting system. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining，2016.

[14] 王克祥，朱建林. P2p 市场借款成功率影响因素研究——基于随机森林和 logistic 回归模型. 中国物价，2019，(10)：53-56.

[15] 李建生. 肺系病辨证纲要与证候的认识. 中医学报，2019，34(1)：1-5.

[16] 冯蓓. 酸甘化阴法治疗气阴两虚型咳嗽 30 例. 广西中医药，2003，(4)：49-49.

[17] 徐超伟，郑敏宇. 郑敏宇中医药治疗咳嗽五法. 辽宁中医药大学学报，2011，13(11)：173-174.

[18] 吴峰妹，蔡敏，王学东. 感染后咳嗽 300 例中医证型观察. 中医药导报，2013，19(11)：6-8.

[19] 国家中医药管理局. 中医病症诊断疗效标准. 南京：南京大学出版社，1994：2-3.

[20] 孙琛，田晓声. 基于 xgboost 算法的变压器故障诊断. 佳木斯大学学报(自然科学版)，2019，37(3)：378-380.

[21] 阎玥，王辛秋，史琦，等. 基于经验辨证的感冒后咳嗽常见中医证候的 logistic 回归分析. 中华中医药杂志，2016，31(3)：814-817.

(责任编辑：邓　妍)