·综述·

机器学习及其流行病学应用

林慧君 王小磊 田梦圆 李杏莉 谭红专
中南大学湘雅公共卫生学院/临床流行病学湖南省重点实验室，长沙 410078
通信作者：谭红专，Email：tanhz99@qq.com

【摘要】 机器学习作为人工智能的重要分支之一，广泛运用于各个领域。机器学习与经典统计学方法有相似之处，同时又能解决许多传统统计学难以处理的难题，因此是流行病学研究的重要工具之一。本文介绍了几种机器学习的常用算法，并对其特点和在流行病学研究中的应用进行了总结。读者可根据研究目的选择适当的机器学习方法，让机器学习技术更好地为流行病学研究服务。

【关键词】 机器学习； 算法； 流行病学

Machine learning and its epidemiological applications

Lin Huijun, Wang Xiaolei, Tian Mengyuan, Li Xingli, Tan Hongzhuan
Xiangya School of Public Health, Central South University, Hunan Key Laboratory of Clinical Epidemiology, Changsha 410078, China

【**Abstract**】 As an important branch of artificial intelligence, machine learning is widely used in various fields. Machine learning has similarity to classical statistical methods, but can solve many problems which are difficult for traditional statistics, so it is one of the important tools in epidemiological research. This paper introduced 9 common algorithms of machine learning and summarized their characteristics and applications in epidemiological research. Readers could choose appropriate machine learning method according to the research purpose for the better application of machine learning in epidemiological research.

【**Key words**】 Machine learning; Algorithms; Epidemiology

机器学习（machine learning）是人工智能的重要组成部分，通过程序赋予计算机“学习”能力。实际应用中，计算机通过数据拟合来提高“学习”能力，与传统的统计学方法相比，机器学习更强调预测结果的准确性，能在多维的数据集中发现规律[1-2]。随着大数据在流行病学中的应用，机器学习为流行病学研究提供了重要工具。本文将对机器学习的几种常用算法及其在流行病学中的应用进行概述。

一、机器学习的基本算法

1. 监督学习（supervised learning）：与流行病学中标准模型拟合相似。在机器学习中，输出值（即因变量）被称为“标记”，运用于监督学习的训练数据的输出值都有一个确切值，称为“标记数据”。而对数据（尤其是大数据）进行“标记”费时且昂贵，限制了它的运用。监督学习在运用中将样本数据与训练的“标签数据”进行比较，并不断调整模型至输出结果稳定准确。回归和分类问题是最常见的2类监督学习任务。常用的监督学习算法有人工神经网络、决策树等。

（1）人工神经网络（artificial neural networks）：生物神经网络是由一群神经元通过复杂的信号通路连接组成，而人工神经网络模拟生物神经元结构，将“神经元”分为3层：输入层（对应于流行病学中的自变量）、隐藏层、输出层（对应于流行病学中的应变量）。相邻层别间通过激活函数（即信号通路）相互连接，激活函数将输入层的数据进行加权转换至输出层[3]。根据激活函数类型的不同，输出可分为二分类数据和连续型数据。其中，在输入层进行的加权相当于统计学中回归分析的系数计算。

在进行数据拟合时，神经元的数量一旦确定，人工神经网络的连接权重将通过迭代的方式拟合到带有标签的训练数据集中，直至达到预设目标或错误阈值。初始连接权重通常是随机产生的，拟合过程中使用损失函数比较网络输

DOI：10.3760/cma.j.cn112338-20200722-00970
收稿日期 2020-07-22 **本文编辑** 李银鸽
引用本文：林慧君, 王小磊, 田梦圆, 等. 机器学习及其流行病学应用[J]. 中华流行病学杂志, 2021, 42(9): 1689-1694. DOI: 10.3760/cma.j.cn112338-20200722-00970.

出值与真实值的差异，并使其尽可能接近零，同时防止过度拟合。二者的误差值则通过网络结构从输出层向输入层传播，称为反向传播[4]，从而计算出隐藏层和输入层中每个神经元的误差权重值。

人工神经网络具备自学、联想储存及快速找到最优解决方案的能力，可应用于模式识别、智能控制、组合优化及预测等领域[3]。在流行病学研究中人工神经网络通常用于非线性关系及变量间交互作用的研究，其模型具有良好的容错能力，且不易受异常值与多重共线性的影响。但应用于大型数据中时，由于隐藏层的复杂性可能使得输出结果不好解释，同时数据拟合要求大量的训练数据集以实现模型的最优化，限制其流行病学的应用[5]。医学上人工神经网络目前在放射、泌尿、检验和心血管等方向应用广泛，利用其计算机处理技术可辅助诊断各类疾病同时还可指导临床用药。该算法可在R软件、SAS、Python中实现。

（2）决策树（decision trees）：决策树是根据输入变量的决策规则来预测结果的。分类决策树（用于分类变量）和回归决策树（用于连续变量）是流行病学研究中最常用的2类决策树。决策树是一种树型的分类预测模型，推导决策树时，决策树算法对决策树上每个分支的数据均进行分割，直至达到预设目标。分割完成后对初始决策树进行简化合并，使得最终生成的决策树简洁明了。

决策树结构简单，便于操作，结果直观易于理解，且具有较高的分类精确度，可用于非线性关系及变量间高阶交互作用的研究[6]。但在实施过程中，决策树易发生过度拟合现象，基于集成方法的决策树能较好地解决此类问题，但其输出结果也随之变得更为复杂，不易解释。同时，由于输入变量的微小改变也可能会导致最终决策树的改变，故应用时应注意各输入变量的稳定性[7]。决策树主要用于解释性建模、结果预测和亚组识别，可筛选出与结果相关程度最大的协变量。此外，临床医生及流行病学工作者可通过构建决策树进行决策分析，有助于解决各种复杂问题和制定相关方案。该算法可在R软件、Stata、SAS、Python中实现。

（3）支持向量机（support vector machines）：支持向量机构造了一个超平面概念对观测值进行分类，分类前可使用核函数将观测值投射至变量可分离的高维空间中进行转换[1]。常用的核函数有多项式核函数、高斯核函数与S型核函数3种，可在常用核函数中通过交叉验证选择最优核函数进行转换。经核函数转换后，超平面能最大化实现变量之间的分类。

支持向量机对小样本、高维数据有良好的适应性且错分率较低[8]，可用于分类和回归问题的处理，在处理多分类问题时可采用多重比较的方法进行[9]。在大型训练数据集中运行时，支持向量机可能会出现效率低下、算法复杂等问题。此外，最优核函数的选择是分析的难点，选择过程耗时，且不能保证标准函数中包括最优核函数，因此在一些复杂问题中需自行设计核函数进行分析。支持向量机广泛用于模式识别（如基因、蛋白质结构识别等）和疾病预测模型的建立，且可直接使用结构化数据。该算法可在R软件、Stata、SAS、Python中实现。

（4）朴素贝叶斯算法（Naïve Bayes algorithms）：朴素贝叶斯算法是基于贝叶斯定理，对变量间的独立性进行简单概率分类的算法。该算法根据一组协变量（先验概率与似然函数的乘积）计算所有可能类别的概率，概率最大的类别即为正确分类。

朴素贝叶斯算法不受缺失数据的影响且要求的训练数据集较少[10]，因此广泛应用于流行病各领域，是数据挖掘中最有效的归纳学习算法之一，临床上与数据挖掘相结合可预测临床事件的发生、诊断肿瘤的分期。但当变量间存在相关性时可能会导致预测偏倚，且先验概率的选择需能代表总体，否则结果也将产生偏倚。此外，朴素贝叶斯算法的输出概率不能代表各类别内变量的真实概率。该算法可在R软件、Stata、SAS、Python中实现。

（5）隐马尔可夫模型（hidden Markov model）：当训练数据中包括状态序列时，隐马尔可夫模型属于监督学习。它是一种以马尔可夫链为理论基础的机器学习算法，描述了观察变量和隐藏变量之间的关系，其中隐藏变量可以代表一个潜在和不可测量的疾病状态。该算法由2个随机过程组成，通过观测过程所对应的观测值来推断马尔可夫过程的转移状态及其特征[11]。隐马尔可夫模型是评价顺序和时间数据最流行的模型之一，在语音识别、基因等生物信息的识别中应用广泛，临床上也可用于疾病预测模型的建立，并可与其他机器学习方法共同作用以达到最佳效果。该算法可在R软件、Python中实现。

2. 无监督学习（unsupervised learning）：数据集通常为“未标记数据”，是根据其内部的自然属性进行分类而不参考某确切值的一种学习方法，可利用统计学方法识别出具有相似特征的子集并将其归类。聚类、关联是最常见的无监督学习方法。

（1）期望最大化算法（expectation maximization algorithm）：EM算法广泛应用于缺失数据概率模型的最大似然估计。该算法通过以下步骤的迭代实现模型的参数估计：①计算期望值：根据观察变量的初始值与参数的估计值计算缺失数据的期望值；②实现最大化：用期望值替代缺失数据进行参数的估计。当模型得到的参数估计值收敛平稳时迭代停止[12]。

EM算法图像处理技术成熟，在计算机、影像学等领域应用广泛，也是流行病学参数估计的常用方法。且该算法易于实现、数值稳定性高[13]，尤其适用于缺失数据的参数估计。同时EM算法可作为辅助工具实现其他算法的功能最大化。但收敛速度慢，且对观察变量初始值的准确性要求高。该算法可在R软件、SAS、Python中实现。

（2）k-means聚类（k-means clustering）：k-means聚类算法是EM算法思想的体现，该算法将观测值划分至预先指定的k种类别中，同时使各类别内差别最小。k-means聚类实现步骤：①在观测值中随机指定k个凝聚点作为重心，每个

凝聚点即为1个类别；②计算各观测值与重心的距离，将其划分至距离最近的重心所属类别中；③初步分类确定后重新计算各类别的重心；④重复步骤②、③至相邻两次分类结果一致。

k-means 聚类算法适用于类内团聚状分布的样本数据[14]。多应用于市场调研方向，临床上也可用于人群中疾病的识别。该算法操作简单，结果易于解释，计算效率高。但类别数量需事先指定，且类别数的微小变化可能会产生完全不同的分类结果[15-16]，同时聚类中心的选取也是影响分类结果的重要因素，若选取不当，则不会得到全局最优解。此外，当类别间数量相差过大时，也可能会导致非直观结果[13]。该算法可在R软件、Stata、SAS、Python中实现。

（3）Apriori 算法（Apriori algorithm）：Apriori 算法是一种关联规则算法，通过寻找变量关系的规则，来确定多元数据集间有用的关联规则。Apriori 算法的基本思想是在频繁k-项集的基础上生成候选(k+1)-项集。记录每个项集的出现频率，并将其与预先设定的阈值比较，大于阈值的项集即为频繁项集。然后在选定的所有频繁项集中筛选出具有强关联规则的项集[17]。Apriori 算法是文本挖掘的常用工具之一，与其联合可用于临床决策的生成，且速度较快。但其扫描数据库过于频繁，且生成大量候选集，导致筛选过程繁琐耗时。该算法可在R软件、JAVA、SAS、Python中实现。

3. 集成方法（ensemble methods）：集成方法是通过特定的方式将若干基分类器的预测结果进行综合[18]。与单一模型相比，集成方法可通过在不同数据集中使用相同的底层算法或在相同的数据集中使用不同的定性模型，利用来自多个模型的信息以提高预测性能。集成方法同时拥有多种算法的优点，可防止模型过度拟合，提高分类效果，通常是组合运用于各领域。以下是3种常见的集成方法：

（1）装袋算法（Bagging）：Bagging 为原始训练数据的每个子集匹配了相同的底层算法，然后根据结果参数化模型的输出创建最终的预测。定量结果的最终预测是通过对预测结果进行平均得到的，而定性结果则取决于分类器中投票的数量或拟合数量概率的平均值。

Bagging可以显著提高不稳定的基分类器的泛化能力，还可处理数据不平衡的问题。该算法在不影响偏差的情况下显著降低了模型的方差同时减少了过度拟合，但模型之间的偏倚较大[18]。Bagging与决策树算法结合产生的随机森林是其经典的模型之一，随机森林是一种非参数分类和回归方法，可用于预测变量数量大于观察数量的情况[19]。在随机森林中，决策树使用随机选择的训练数据集和预测变量的随机子集来构建模型结果[20]。因此与单一决策树模型比，随机森林拥有更高的准确性，同时可较好地处理小样本和高维数据。随机森林不依赖于数据分布假设，适用于非线性关系、交互作用及数据缺失的研究，同时，在遗传流行病学和肿瘤学上应用广泛。另一方面，随机森林给予了弱相关变量进入模型的机会，同时可输出各变量对结果的重要性排名[21-23]。该算法可在R软件、Stata、SAS、Python中实现。

（2）提升算法（Boosting）：Boosting按照顺序在数据子集上进行模型演练，并通过分析预测误差来改进分类器。初始时赋予观测值相同的权重，若分类器在迭代过程中发生错误，权重则相应增加，随后再进行迭代。可根据最终的输出分类器中权重的变化情况判断分类器的预测精度，权重越大，预测精度越高。Boosting常用于决策树与神经网络中以提高二者的泛化能力[18]，其分类精确度较高，不易发生过度拟合。但当数据样本不大时，Boosting执行效果较差。该算法可在R软件、Stata、SAS、Python中实现。

（3）贝叶斯模型平均（Bayesian model averaging）：贝叶斯模型平均通过计算特定模型估计的加权平均值来估计预测值的后验分布，其中权重是由每个竞争模型的数据量多少决定的。贝叶斯模型平均已被广泛用于统计模型中，包括线性模型、Cox比例风险模型等，减少了模型的不确定性，较单一模型而言具有更强的预测能力[24]。目前贝叶斯模型平均被用于空间流行病学、交互作用及基因相关研究中。该算法可在R软件、SAS、Python中实现。

二、流行病学应用

1. 因果推论：机器学习可较好地解决非线性及变量间交互的问题，同时也可进行多维数据的处理，而传统的统计学方法难以处理此类问题。在观察性研究中，机器学习作为估计因果关联的方法之一，可降低偏倚、控制混杂[25]。观察性研究中通常使用倾向评分法来控制混杂以估计因果效应，倾向评分一般是利用logistic回归进行估计，若假设不明，此方法可能会导致偏倚。而最大似然估计是倾向评分的一种替代方法，机器学习技术可对该方法进行改进增加其稳定性[26-27]。另外，贝叶斯网络也可用于探讨药物与不良反应之间的因果关系[28]。因果结构学习是机器学习的一个独特分支。线性、非高斯、非循环模型方法是因果结构学习中的基本方法之一[29]，在睡眠障碍与抑郁的因果关系研究中，Rosenström等[30]引入此方法，从一队列资料建立已知变量间的因果关系模型来测试线性、非高斯、非循环模型方法的准确性，再对睡眠障碍和抑郁的因果关系进行不同的假设，根据横断面资料使用由DirectLiNGAM算法建立的3个线性模型分别对假设进行检验，并最终得出睡眠障碍是导致抑郁的原因之一。

2. 疾病的预测、诊断、预后及临床决策：机器学习已被应用到传染病的预测模型中，如使用光谱聚类算法检测婴儿呼吸道微生物群的变化，为婴儿呼吸道感染的诊断及预防提供了依据[31]。贝叶斯模型可用于结核病的管理中，以确定结核病患者的临床属性[32]。

在新冠肺炎的防治中，机器学习也发挥了巨大的作用。利用机器学习技术建立易感暴露感染消除的流行病学模型，结合疾病相关流行病学数据来预测疫情的进展[33-34]。使用深度学习技术根据时间序列数据建立一个长短期记忆网络模型评估加拿大疫情严重程度[35]。研究人员利用随机森林、人工神经网络技术根据入院病人的全血细胞计数结

果建立新冠肺炎的预测模型，在不了解病人症状及病史的情况下能识别出85%的社区新冠病毒阳性患者[36]。

此外，机器学习技术广泛应用于疾病的诊断，尤其是涉及图像及高维数据的问题。k-means聚类算法可用于确定脓毒血症临床表型的分类，改善治疗效果[37]。Pandey等[38]使用RNA测序技术分析了轻度至中度哮喘患者的基因表达，建立了一个高维的基因数据库，将机器学习与经典统计学方法结合经过变量的选择、分类模型的形成及优化、模型的确定三步骤建立了一个基于鼻刷的分类器，用于哮喘的诊断及分型，且经外部验证模型精确性和特异性均较高。机器学习技术可作为辅助手段用于各项疾病的筛查，有助于疾病的早期诊断[39-40]。深度学习是机器学习领域的新兴技术，其图像处理技术实现了高危糖尿病视网膜病变及糖尿病黄斑水肿的快速检测[41-42]。机器学习的发展提高了临床预测的准确性，如利用机器学习技术改进Cox比例风险模型从而建立一种靶向实时预警评分系统，并将其应用于预测ICU脓毒血症患者发病后28 h内感染性休克的概率[43]。根据电子健康记录，使用支持向量机、朴素贝叶斯模型、随机森林等机器学习技术建立自动化识别感染患者的模型以辅助急诊科的诊断[44]。

3. 全基因关联研究：通常用来探究遗传因素对疾病的影响，此类数据包含大量待提取的遗传物质且样本量有限，随机森林等集成方法能较好地处理这些高维数据。Krishnan等[45]构建了一个大脑特定功能区的基因交互网络，在全基因组分析数据的基础上利用贝叶斯方法、支持向量机建立了一个证据加权网络的分类器，可筛选自闭症新的致病基因并按照基因与疾病的关联程度进行排序。还可根据全基因关联性研究结果量化个体疾病风险，建立疾病的个性化治疗方案。如在哮喘病人中运用随机森林根据基因信息预测疾病的严重程度，制定针对性治疗方案[46]。研究人员采用一种可扩展的机器学习技术来识别人群基因组中DNA拷贝数的变化以达到癌症早期诊断[47]。机器学习还可用于识别基因-基因和基因-环境的交互作用，提供疾病的病因线索[48-49]。

4. 空间流行病学研究：将随机森林与遥感、空间地理数据结合，可绘制城市人口密度图[50]。在数据有限的地区，机器学习可以用来预测和绘制疾病发生与健康指标的分布。使用增强回归树可建立人畜共患病宿主的预测模型并绘制疾病可能发生的地区分布地图[51]。机器学习结合卫星遥感技术可建立城市高温预测模型，有助于高温的预防[52]。研究者基于机器学习技术将泰国某地区登革热的疾病监测数据、气象监测数据、社会经济数据及地理信息结合，使用广义相加模型来拟合数据集并预测该地区未来一个月之内登革热的传播[53]。另一方面，将机器学习、文本挖掘和空间地理信息结合可为环境流行病学的暴露评估建模提供重要工具[54]。

5. 文本挖掘：在对具有高维属性的海量数据进行信息转化提取的过程中，传统的人工分析方法已不太适合，文本挖掘技术应运而生。文本挖掘是指在大量非结构化数据文本中发现、提取出有价值的数据信息，并对其进行分析[55]。文本挖掘一般包括信息检索、信息抽取、数据挖掘3个部分[56]，而机器学习技术将经典统计学知识与计算机技术结合，是实现文本挖掘的常用方法。

目前利用机器学习技术实现的文本挖掘广泛应用于医疗领域。如：利用自然语言处理从医院的死亡证明信息中提取癌症相关信息，再使用支持向量机对常见肿瘤进行分类，规则算法则用于处理罕见癌症的信息。最后将两种方法融合可实现准确及时地报告各类癌症的死亡率[57]。自然语言处理也可用于从电子医疗记录中提取数据，在临床上用于从放射学报告中准确识别脂肪肝[58]。另一方面，自然语言处理也可在医疗记录中识别疾病危险因素，有助于疾病的早期预防及治疗[59-61]。精准医疗是目前的医疗大趋势，需要大量数据库支持，文本挖掘工具可在全网大量文献中提取所需信息，提高效率[62]。使用Apriori算法的关联规则分析进行文本挖掘，有助于在各类文献中寻找疾病的治疗方法[63]。此外，文本挖掘还可用于癌症存活率的预测，研究表明使用决策树建立的乳腺癌存活率预测模型准确性可高达93.6%[64]。

三、小结

本文介绍了机器学习的几种常见算法，读者可根据研究目的选择不同的类型。值得注意的是，集成算法虽然预测准确性高，但由于其复杂性使得结果难以解释，故在临床上应用受限。机器学习等新兴技术的发展开启了流行病学研究的新纪元，为经典统计学方法难以解决的问题提供了新方向，但流行病学家目前还没有熟练使用此类新兴技术的技能。因此，应鼓励流行病学家加强对机器学习的使用，让机器学习更好地为流行病学研究服务。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] Rasmussen CE, Williams CKI. Gaussian processes for machine learning[M]. Cambridge:The MIT Press, 2006.

[2] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author) [J]. Statal Sci, 2001, 16(3):199-231. DOI:10.1214/ss/1009213726.

[3] Wu YC, Feng JW. Development and application of artificial neural network[J]. Wireless Personal Commun, 2018, 102(2):1645-1656.

[4] Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application[J]. J Microbiol Methods, 2000, 43(1): 3-31. DOI: 10.1016/S0167-7012(00)00201-3.

[5] Hershey S, Chaudhuri S, Ellis DPW, et al. CNN architectures for large-scale audio classification[C]// 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans, LA, USA: IEEE, 2016.

[6] Boulesteix AL, Janitza S, Hapfelmeier A, et al. Letter to the editor: on the term "interaction" and related phrases in

the literature on Random Forests[J]. Brief Bioinform, 2015, 16(2):338-345. DOI:10.1093/bib/bbu012.

[7] Aluja-Banet T, Nafria E. Stability and scalability in decision trees[J]. Computat Stats, 2003, 18(3): 505-520. DOI:10.1007/BF03354613.

[8] Belousov AI, Verzakov SA, von Frese J. A flexible classification approach with optimal generalisation performance: support vector machines[J]. Chemomet Intellig Lab Syst, 2002, 64(1): 15-25. DOI: 10.1016/S0169-7439(02)00046-1.

[9] Noble WS. What is a support vector machine? [J]. Nat Biotechnol, 2006, 24(12): 1565-1567. DOI: 10.1038/nbt1206-1565.

[10] Rish I. An empirical study of the naive Bayes classifier[J]. J Univers Comput Sci, 2001, 1(2): 41-46. DOI: 10.1002/9781118721957.ch4.

[11] 朱明, 郭春生. 隐马尔可夫模型及其最新应用与发展[J]. 计算机系统应用, 2010, 19(7):255-259, 216. DOI:10.3969/j.issn.1003-3254.2010.07.061.
Zhu M, Guo CS. Hidden markov model and its latest application and progress[J]. Computer Syst Applicat, 2010, 19(7): 255-259, 216. DOI: 10.3969/j.issn.1003-3254.2010.07.061.

[12] Huang WH, Chen YG. The multiset EM algorithm[J]. Stats Probabil Lett, 2017, 126: 41-48. DOI: 10.1016/j.spl.2017.02.021.

[13] Dwidayati N, Zaenuri. Convergence properties of the EM algorithm in the mixture model with missing data[J]. J Phys Conf Ser, 2018, 983(1): 012091. DOI: 10.1088/1742-6596/983/1/012091.

[14] 周世兵, 徐振源, 唐旭清. K-means算法最佳聚类数确定方法[J]. 计算机应用, 2010, 30(8):1995-1998.
Zhou SB, Xu ZY, Tang XQ. Method for determining optimal number of clusters in K-means clustering algorithm[J]. J Comput Appl, 2010, 30(8):1995-1998.

[15] Tibshirani R, Walther G, Hastie WT. Estimating the number of clusters in a data set via the gap statistic[J]. J Roy Statal Soc B, 2001, 63(2): 411-423. DOI: 10.1111/1467-9868.00293.

[16] Raykov YP, Boukouvalas A, Baig F, et al. What to do when K-means clustering fails: a simple yet principled alternative algorithm[J]. PLoS One, 2016, 11(9): e0162259. DOI:10.1371/journal.pone.0162259.

[17] 饶正婵, 范年柏. 关联规则挖掘Apriori算法研究综述[J]. 计算机时代, 2012(9): 11-13. DOI: 10.3969/j.issn.1006-8228.2012.09.005.
Rao ZC, Fan NB. A review of associative rule mining Apriori algorithm[J]. Computer Era, 2012(9): 11-13. DOI: 10.3969/j.issn.1006-8228.2012.09.005.

[18] 杨剑锋, 乔佩蕊, 李永梅, 等. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019, 35(6):36-40. DOI:10.13546/j.cnki.tjyjc.2019.06.008.
Yang JF, Qiao PR, Li YM, et al. A review of machine-learning classification and algorithms[J]. Statist Decis, 2019, 35(6):36-40. DOI:10.13546/j.cnki.tjyjc.2019.06.008.

[19] Wang Y, Li Y, Pu WL, et al. Random bits forest: a strong classifier/regressor for big data[J]. Sci Rep, 2016, 6: 30086. DOI:10.1038/srep30086.

[20] Speiser JL, Miller ME, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modeling[J]. Exp Syst Appl, 2019, 134:93-101. DOI:10.1016/j.eswa.2019.05.028.

[21] Breiman L. Random forests[J]. Mach Learn, 2001, 45(1): 5-32. DOI:10.1023/A:1010933404324.

[22] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction [M]. 2nd ed. New York:Springer-Verlag, 2009.

[23] van der Laan MJ. Statistical inference for variable importance[J]. Int J Biost, 2006, 2(1): 2. DOI: 10.2202/1557-4679.1008.

[24] Kaplan D, Lee C. Optimizing prediction using Bayesian model averaging: examples using large-scale educational assessments[J]. Eval Rev, 2018, 42(4): 423-457. DOI: 10.1177/0193841X18761421.

[25] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available[J]. Am J Epidemiol, 2016, 183(8): 758-764. DOI: 10.1093/aje/kwv254.

[26] Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies [J]. Am J Epidemiol, 2017, 185(1): 65-73. DOI: 10.1093/aje/kww165.

[27] Herrera R, Berger U, von Ehrenstein OS, et al. Estimating the causal impact of proximity to gold and copper mines on respiratory diseases in Chilean children: an application of targeted maximum likelihood estimation[J]. Int J Environ Res Public Health, 2017, 15(1): 39. DOI: 10.3390/ijerph15010039.

[28] Rodrigues PP, Ferreira-Santos D, Silva A, et al. Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network[J]. Artif Intell Med, 2018, 91:12-22. DOI:10.1016/j.artmed.2018.07.005.

[29] Shimizu S. Non-gaussian methods for causal structure learning[J]. Prev Sci, 2019, 20(3): 431-441. DOI: 10.1007/s11121-018-0901-x.

[30] Rosenström T, Jokela M, Puttonen S, et al. Pairwise measures of causal direction in the epidemiology of sleep problems and depression[J]. PLoS One, 2012, 7(11): e50841. DOI:10.1371/JOURNAL.PONE.0050841.

[31] Biesbroek G, Tsivtsivadze E, Sanders EA, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children[J]. Am J Respir Crit Care Med, 2014, 190(11): 1283-1292. DOI:10.1164/rccm.201407-1240OC.

[32] Getoor L, Rhee JT, Koller D, et al. Understanding tuberculosis epidemiology using structured statistical models[J]. Artif Intell Med, 2004, 30(3): 233-256. DOI: 10.1016/j.artmed.2003.11.003.

[33] Yang ZF, Zeng ZQ, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. J Thorac Dis, 2020, 12(3):165-174. DOI:10.21037/jtd.2020.02.64.

[34] Marfak A, Achak D, Azizi A, et al. The hidden Markov chain modelling of the COVID-19 spreading using Moroccan dataset[J]. Data Brief, 2020, 32: 106067. DOI: 10.1016/j.dib.2020.106067.

[35] Chimmula VKR, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks [J]. Chaos Solit Fract, 2020, 135: 109864. DOI: 10.1016/j.chaos.2020.109864.

[36] Banerjee A, Ray S, Vorselaars B, et al. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population[J]. Int Immunopharmacol, 2020, 86: 106705. DOI: 10.1016/j.intimp.2020.106705.

[37] Seymour CW, Kennedy JN, Wang S, et al. Derivation,

validation, and potential treatment implications of novel clinical phenotypes for sepsis[J]. JAMA, 2019, 321(20): 2003-2017. DOI:10.1001/jama.2019.5791.

[38] Pandey G, Pandey OP, Rogers AJ, et al. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal RNA sequence data[J]. Sci Rep, 2018, 8(1):8826. DOI:10.1038/s41598-018-27189-4.

[39] Polak S, Mendyk A. Artificial intelligence technology as a tool for initial GDM screening[J]. Expert Syst Appl, 2004, 26(4):455-460. DOI:10.1016/j.eswa.2003.10.005.

[40] Abbas H, Garberson F, Glover E, et al. Machine learning approach for early detection of autism by combining questionnaire and home video screening[J]. J Am Med Inform Assoc, 2018, 25(8): 1000-1007. DOI: 10.1093/jamia/ocy039.

[41] Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning[J]. Ophthalmology, 2017, 124(7):962-969. DOI:10.1016/j.ophtha.2017.02.008.

[42] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. JAMA, 2016, 316(22): 2402-2410. DOI: 10.1001/jama.2016.17216.

[43] Henry KE, Hager DN, Pronovost PJ, et al. A targeted real-time early warning score (TREWScore) for septic shock[J]. Sci Transl Med, 2015, 7(299): 299ra122. DOI: 10.1126/scitranslmed.aab3719.

[44] Horng S, Sontag DA, Halpern Y, et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning[J]. PLoS One, 2017, 12(4):e0174708. DOI:10.1371/JOURNAL.PONE.0174708.

[45] Krishnan A, Zhang R, Yao V, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder[J]. Nat Neurosci, 2016, 19(11): 1454-1462. DOI:10.1038/nn.4353.

[46] Xu MS, Tantisira KG, Wu A, et al. Genome wide association study to predict severe asthma exacerbations in children using random forests classifiers[J]. BMC Med Genet, 2011, 12:90. DOI:10.1186/1471-2350-12-90.

[47] Manogaran G, Vijayakumar V, Varatharajan R, et al. Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering[J]. Wirel Pers Commun, 2018, 102(3): 2099-2116. DOI:10.1007/s11277-017-5044-z.

[48] Hu T, Sinnott-Armstrong NA, Kiralis JW, et al. Characterizing genetic interactions in human disease association studies using statistical epistasis networks[J]. BMC Bioinformatics, 2011, 12: 364. DOI: 10.1186/1471-2105-12-364.

[49] Roetker NS, Page CD, Yonker JA, et al. Assessment of genetic and nongenetic interactions for the prediction of depressive symptomatology: an analysis of the Wisconsin Longitudinal Study using machine learning algorithms[J]. Am J Public Health, 2013, 103 Suppl 1: S136-144. DOI: 10.2105/AJPH.2012.301141.

[50] Steele JE, Nieves J, Tatem AJ, et al. Worldpop-fusion of earth and big data for intraurban population mapping [C]//IGARSS 2018 - 2018 IEEE international geoscience and remote sensing symposium. Valencia, Spain: IEEE, 2018.

[51] Han BA, Schmidt JP, Bowden SE, et al. Rodent reservoirs of future zoonotic diseases[J]. Proc Natl Acad Sci USA, 2015, 112(22):7039-7044. DOI:10.1073/pnas.1501598112.

[52] Dos Santos RS. Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data[J]. Int J Appl Earth Observat Geoinformat, 2020, 88: 102066. DOI: 10.1016/j.jag.2020.102066.

[53] Jain R, Sontisirikit S, Iamsirithaworn S, et al. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data[J]. BMC Infect Dis, 2019, 19(1):272. DOI:10.1186/S12879-019-3874-X.

[54] Vopham T, Hart JE, Laden F, et al. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology[J]. Environ Health, 2018, 17(1):40. DOI:10.1186/S12940-018-0386-X.

[55] Voyant C, Notton G, Kalogirou S, et al. Machine learning methods for solar radiation forecasting: A review[J]. Renew Energy, 2017, 105: 569-582. DOI: 10.1016/j.renene.2016.12.095.

[56] 王浩畅, 赵铁军. 生物医学文本挖掘技术的研究与进展[J]. 中文信息学报, 2008, 22(3): 89-98. DOI: 10.3969/j.issn.1003-0077.2008.03.012.
Wang HC, Zhao TJ. Research and development of biomedical text mining[J]. J Chin Inf Process, 2008, 22(3): 89-98. DOI:10.3969/j.issn.1003-0077.2008.03.012.

[57] Koopman B, Zuccon G, Nguyen A, et al. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers[J]. Artif Intell Med, 2018, 89: 1-9. DOI: 10.1016/j.artmed.2018.04.011.

[58] Redman JS, Natarajan Y, Hou JK, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing[J]. Dig Dis Sci, 2017, 62(10):2713-2718. DOI:10.1007/s10620-017-4721-9.

[59] Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes[J]. J Biomed Inform, 2015, 58 Suppl 1:S128-132. DOI:10.1016/J.JBI.2015.08.002.

[60] Torii M, Fan JW, Yang WL, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records[J]. J Biomed Inform, 2015, 58 Suppl 1: S164-170. DOI:10.1016/J.JBI.2015.08.011.

[61] Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease[J]. J Biomed Inform, 2015, 58 Suppl 1: S171-182. DOI: 10.1016/J.JBI.2015.09.006.

[62] Singhal A, Simmons M, Lu ZY. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine [J]. PLoS Comput Biol, 2016, 12(11): e1005017. DOI: 10.1371/journal.pcbi.1005017.

[63] Hsieh PC, Cheng CF, Wu CW, et al. Combination of acupoints in treating patients with chronic obstructive pulmonary disease: an apriori algorithm-based association rule analysis[J]. Evid Based Complement Alternat Med, 2020, 2020: 8165296. DOI: 10.1155/2020/8165296.

[64] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods [J]. Artif Intell Med, 2005, 34(2): 113-127. DOI: 10.1016/j.artmed.2004.07.002.