DOI:10.13718/j.cnki.xsxb.2021.05.022

基于格拉姆矩阵和随机森林的疾病预测方法①

邹劲松

重庆水利电力职业技术学院 普天大数据产业学院,重庆 402160

摘要:针对现有预测方法中存在预测精度低、预测时间长及存储空间大等问题,提出一种基于格拉姆矩阵和随机森林的疾病预测方法,该方法首先从数据集中收集大量数据,其次使用格拉姆对称矩阵对采集数据进行存储和归类.然后引入随机森林二元回归和分类技术,通过二元变量相关性来衡量预测结果和数据之间的关系,并根据相关性构造决策树用于结果分类.最后,应用表决方案输出最终预测结果.实验结果表明:与其他方法相比,本文提出的方法提高了预测准确度,降低了预测的时间开销和空间复杂度.

关 键 词:大数据;格拉姆矩阵;随机森林;预测分析

中图分类号:TP393 **文献标志码**:A **文章编号**:1000-5471(2021)05-0147-06

在大数据中,预测分析是指从大型数据集中提取相关信息用于总结规律并对未来做预测分析的过程[1].大数据分析大致分为数据收集、存储、分析和预测 4 个阶段,研究人员根据收集到的大数据信息,通过执行数据聚合进行数据分析来完成预测[2].预测分析可用于各种应用领域,如医疗保健、保险业务及天气预报等[3-4].

近年来,许多研究学者从不同角度提出了用于大数据预测分析的方法.Chen 等[5]提出一种基于卷积神经网络的多模态疾病风险预测算法,并成功应用于有序和非结构化数据中.但是该算法无法以较少的时间开销实现有效的预测.Gu 等[6]采用贝叶斯和神经网络相结合的方式对大数据进行更广泛、更一致的预测,但是该技术的预测精度不高.Nair 等[7]在开源大数据处理引擎 Apache Spark 中引入可扩展的机器学习方法来预测用户的健康状况,但是该模型预测疾病的种类较少.Babu 等[8]提出了一种基于灰狼优化和自编码递归神经网络的疾病预测模型,该模型利用灰狼优化对数据进行特征选择,然后采用自编码递归神经网络进行疾病预测.Mohan 等[9]利用混合随机森林线性模型建立的心脏病预测模型,通过多种机器学习分类器技术寻找显著特征,有效提高了预测精度.但是这种方法是多种分类器的组合,在预测过程中时间花费较多.Yao 等[10]在 MapReduce 框架中引入分布式并行极限学习机和层次式极限学习机用于大数据多模态过程质量预测,将高效的极限学习机算法转化为分布式并行建模形式,降低计算时间,实现在线预测.

针对上述文献中存在的预测精度低,时间开销大以及空间复杂度高等问题,提出一种基于格拉姆矩阵和随机森林的大数据预测方法,该方法首先对采集到的大数据集利用格拉姆对称矩阵进行存储,然后使用随机森林二元回归和分类技术对存储后的数据进行处理,以最小的时间和空间复杂性提高预测精度.

1 随机森林模型

随机森林(Random Forest,RF)[11] 是一种机器学习方法,该方法融合了 Bagging 思想和随机子空间思

① 收稿日期:2020-04-08
基金项目:重庆市教育科学"十三五"规划 2017 年度课题(2017-GX-181).
作者简介:邹劲松,硕士,副教授,主要从事计算机软件及理论研究.

想，利用随机重采样和节点随机分裂技术构建多个决策树，然后采用投票的方式得到最终的分类结果. 对于每个决策树模型(X，β_k)都拥有选择最终分类结果的投票权，分类决策公式为

$$H(X) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{1}$$

式(1)中：$H(X)$ 和 $h_i(x)$ 分别表示随机森林和单个决策树的分类结果，Y 表示分类目标，$I(\cdot)$ 表示示性函数.

随机森林的建模及预测步骤可以总结为：① 给定包含 N 个样本的数据集，采用 Bootstrap 方法进行 K 次有放回的随机抽样操作，得到 K 个训练样本集；② 对每个采样集，从所有特征中随机选择 m 个特征，然后从中选择具有最佳分类能力的特征作为节点进行分裂，构建 K 棵分类回归树；③ 保证每棵树最大限度地生长，不对其做任何剪裁；④ 将生成的多棵树组成随机森林，利用随机森林对新数据进行预测：分类任务使用投票法确定样本的最终分类，回归任务使用平均法确定样本的最终预测.

2 基于随机森林分类的大数据预测

大数据是海量和复杂数据的集合. 通过计算非常大的数据集形成众所周知的模式，并将其用于预测分析. 随着大数据的发展，预测是在早期阶段确定未来结果，从而最大程度地降低风险水平的一项主要任务. 目前，已经存在一些用于执行预测分析的研究工作. 但是，准确的预测仍然是一个具有挑战性的问题. 为了以最小的时间开销提高预测精度，本文提出一种基于随机森林分类的大数据预测方法.

本文提出方法包括3个主要步骤，即预处理、数据分析和预测. 首先，从数据集中采集大量数据，这些收集到的数据使用格拉姆对称矩阵进一步存储. 由于格拉姆矩阵被视为是数据特征之间的偏心协方差矩阵，根据格拉姆矩阵的差异可以度量特征之间的相关性，进而度量各个维度自身的特性及各个维度之间的关系. 本文将大量数据存储在格拉姆矩阵中，有助于在预测分析中将干扰因素降至最低. 数据存储后，采用随机决策森林学习方法进行回归和分类. 数据分析是通过双变量相关分析来确定相关数据和独立数据之间的关系. 然后，利用根节点、分支节点和叶节点3个不同的节点构造决策树，将决策树进行组合，应用投票方案得到准确的预测结果.

2.1 数据的预处理

数据的预处理首先是从现有的公开疾病数据集中收集数据，本文选择加利福尼亚大学欧文分校公开的心脏病数据集、糖尿病数据集和癌症资料集. 通过选择相关属性，获得有助于预测决策的有效信息. 假设从大型数据集中收集的数据数量可以定义为

$$D_1, D_2, \cdots, D_n \in D^l \tag{2}$$

式(2)中：D_i 表示从大数据集 D^l 收集的数据. 为了进一步分析数据，在收集数据后将进行数据存储.

数据存储也是数据分析的步骤之一. 本文使用格拉姆对称矩阵存储数据，获取的数据不会被修改并存储在矩阵中. 任意 n 个向量之间两两的内积所组成的矩阵，称为 n 个向量的格拉姆矩阵. 格拉姆矩阵的构造为

$$\boldsymbol{g}_{ij} = \begin{bmatrix} \langle D_{1,1} \rangle & \langle D_{1,2} \rangle & \cdots & \langle D_{1,n} \rangle \\ \vdots & \vdots & & \vdots \\ \langle D_{n,1} \rangle & \langle D_{n,2} \rangle & \cdots & \langle D_{n,n} \rangle \end{bmatrix} \tag{3}$$

式(3)中：$\boldsymbol{g}_{ij}$ 表示按行和列排列的存储数据 D_1，D_2，…，D_n 的格拉姆矩阵. $\langle D_{1,1} \rangle$ 表示存储在矩阵第一行第一列中的数据. 如果在矩阵中添加了任何附加信息，则会生成新的列和行. 这种存储方法为数据提供了简单的访问方式，有助于算法在处理数据时最大程度地减少计算时间.

2.2 随机森林回归和分类

存储数据后，使用随机森林二元回归和分类(Random Forest Bivariate Regression and Classification，RFBRC)模型进行预测分析，它是一种通过构造多个决策树来执行回归和分类的集成学习方法. 回归是一种数学过程，用于测量两个变量之间的关系，其中将被预测的变量称为因变量，而自变量是数据的成员.

在更改一个或多个自变量时，因变量会发生变化.

RFBRC 技术最初通过测量因变量和自变量之间的关系来构造决策树对数据进行分类，RFBRC 技术包含 n 个用于分类的二元决策树，单个决策树的预测在其训练集中极易受到噪声的影响，每个决策树都是通过一种随机方法来限定的，因此该分类器被称为随机决策森林分类. 考虑一个训练集，其数据 D_1，D_2，…，D_n 取自格拉姆对称矩阵的第一行. 这些数据以文件的形式作为随机林的输入，随机森林通过二元相关技术度量因变量和自变量之间的关系进行回归分析. 二元相关技术的定义为

$$\rho = Z/\sqrt{\sum p^2 - (\sum p)^2}\left(\sum q^2 - (\sum q)^2\right) \tag{4}$$

式(4) 中：$Z = (\sum pq) - (\sum p)(\sum q)$，$\rho$ 表示相关系数，p 表示因变量，q 表示自变量. 二元相关系数是 p 和 q 之间关系的量度，在[−1，1]之间取值，1 表示正相关，−1 表示负相关. 基于关系度量，构造决策树对数据进行分类. 决策树使用 3 种类型的节点，即根节点，分支节点和叶节点. 根节点表示对数据的测试，每个分支均提供测试结果；叶节点表示类标签，根节点被分为两个子节点及一个确定的决策. 每个决策树分类器的输出可以表示为

$$y \to h_1(D),\ h_2(D),\ \cdots,\ h_n(D) \tag{5}$$

所有决策树分类器输出被合并后的投票结果定义为

$$V_{\text{final}} = \arg\max_n V\{h_i\} \tag{6}$$

式(6) 中：V 表示应用于决策树 $\{h_t\}$ 的表决. 对预测过程中的泛化误差进行测量，用于识别算法在基于历史数据预测结果值的准确性. 误差被测量为预期误差与观察误差之间的差.

$$\Delta = \varepsilon_{ex} - \varepsilon_{em} \tag{7}$$

式(7) 中：Δ，ε_{ex} 和 ε_{em} 分别表示预测误差、期望误差和观察误差. RFBRC 技术的误差计算用于最小化预测过程中的假阳性率.

3　实验与结果分析

为了验证本文算法的有效性，使用 Java 语言对提出的大数据预测技术进行实验评估，测试数据是从加利福尼亚大学欧文分校的机器学习存储库(UCI)[12] 中公开的心脏病数据集、糖尿病数据集和癌症资料集 3 个大数据集中收集的.

3.1　数据集采集和存储

心脏病数据集包含 76 个属性，在这些属性中只有 14 个属性被用于实验评估. 用于预测分析的属性包括患者识别号、年龄、性别、患者姓名、血压、胆固醇等. 数据中有 303 个病患实例用于预测其未来的健康状况. 糖尿病数据集包含 55 个属性，通过收集这些数据来构建患者文件用于预测分析. 该数据集包含 100 000 个实例. 患者数据包括种族、性别、年龄、入院类型，进行的实验室检查次数、糖化血红蛋白(HbA1c) 检测结果和糖尿病药物等均从大型数据集中收集. 乳腺癌数据集还收集用于构建患者记录文件的数据，属性的数目有身份证(Id) 号、细胞大小、形状、类属性如良性肿瘤或恶性肿瘤等.

测试数据集是从上述 3 个数据集中采集得到的，然后将数据以行和列的形式存储在矩阵中，并对存储的数据进行分析和分类. 第一行代表一个患者的采集数据，作为模型的训练集，然后对测试数据进行分类以预测疾病类型. 出于实验考虑，每个数据集中的患者数量选择在 10 ～ 200 例.

3.2　实验结果分析

实验评估采用预测准确度、预测时间、误报率和空间复杂度 4 个指标对提出的预测方法进行评估，并将测试结果与基于卷积神经网络的多模态(CNN-MDRP) 疾病风险预测方法[5]、基于加权集成神经网络的疾病风险预测(WENN-MDRP) 方法[13] 和基于朴素贝叶斯技术(BPA-NB) 的疾病预测分析模型进行比较[14].

预测准确度是指将多个病人档案预测为实验评估所取的输入档案总数. 预测精度计算公式为

$$PA = \frac{PC}{n} \times 100\% \tag{8}$$

式(8)中：PA 表示预测精度，PC 和 n 分别表示正确预测文件和输入文件数量.

预测时间定义为基于患者当前信息预测疾病类型所需的时间量. 预测时间的数学公式为

$$PT = N \times time \tag{9}$$

式(9)中：N 表示预测的疾病文件数，$time$ 表示预测疾病类型所需要的时间.

误报率是指错误预测的文件数与实验评估所用文件总数的比率. 测量方法为

$$FPR = \frac{PIC}{n} \times 100\% \tag{10}$$

式(10)中：n 表示输入文件数量，FPR 表示误报率，PIC 表示错误预测文件的数量.

空间复杂性被定义为在大型医疗数据分析中存储患者文件所需的存储空间量. 测量空间复杂度的数学公式为

$$SC = m \times space \tag{11}$$

式(11)中：m 表示病人档案的数量，$space$ 表示存储单个档案文件的空间量.

表1描述了关于患者档案数量的预测准确度的实验结果，结果表明本文提出算法的预测精度高于现有方法. 以10个病人为例，他们的数据在一个记录文件中处理. 病人的资料，如病人识别号、年龄、性别、病人姓名、血压信息、胆固醇、胸痛类型、糖化血红蛋白测试，都是从3个不同的数据集中收集的. 每个患者数据都以文件格式进行维护，以便进行简单访问，这些病人的档案作为疾病类型预测的输入. 疾病预测早期阶段在医疗社区中起着至关重要的作用，可以最大限度地降低患者的死亡率，而本文技术可以有效地预测早期的疾病类型，这是通过应用机器学习分类器来实现的. 将患者信息作为随机森林决策树分类器的输入，分类器通过患者数据和疾病类型之间的关系度量来分析输入的患者文件，用二元相关法确定两者之间的关系. 相关测度提高了疾病预测的准确性，基于相关性的决策树分类器可识别疾病类型. 集成分类器包含更多的决策树分类器，每个分类器分别提供预测结果. 将所有的分类器结果进行组合，并将其应用于表决方案，最终预测结果由决策结果的多数票决定. 因此，医生可以有效地找出影响病人的疾病类型，并将通过早期预测使疾病风险最小化.

表1 不同预测方法的准确度对比结果

患者数量/例	预测准确度/%			
	本文方法	CNN-MDRP	WENN-MDRP	BPA-NB
10	82.3	62.2	75.6	80.9
50	88.5	73.3	80.7	85.6
100	95.7	90.0	91.9	93.2
150	96.9	92.5	93.6	94.4
200	97.8	93.4	94.5	96.1

表2给出了4种方法的误报率. 在大型医疗数据分析中，误报率是获得无风险疾病预测的主要参数. 从表2中可以清楚地看出，本文方法的误报率明显低于其他方法，其中集成分类器对分类后的泛化误差进行度量是降低误报率的一个重要方法.

表2 不同预测方法的误报率对比结果

患者数量/例	误报率/%			
	本文方法	CNN-MDRP	WENN-MDRP	BPA-NB
10	19.6	40.8	28.3	23.4
50	26.2	47.6	38.8	31.3
100	34.5	53.2	47.6	40.3
150	38.9	58.9	52.8	44.4
200	43.7	62.1	56.9	49.5

图1显示了预测时间与患者文件数量的性能结果. 从图1中可以清楚地看出，随着病人档案数量的增

加，疾病预测时间也相应上升. 与其他算法相对，本文提出的算法疾病预测时间最少. 这是因为本文算法在收集患者数据时使用格拉姆对称矩阵存储文件，并通过随机森林回归和分类预测疾病.

图 2 显示了空间复杂度与多个患者文件的性能结果. 图 2 清楚地表明，在大型医疗数据分析中，本文算法的空间复杂度大大低于其他算法.

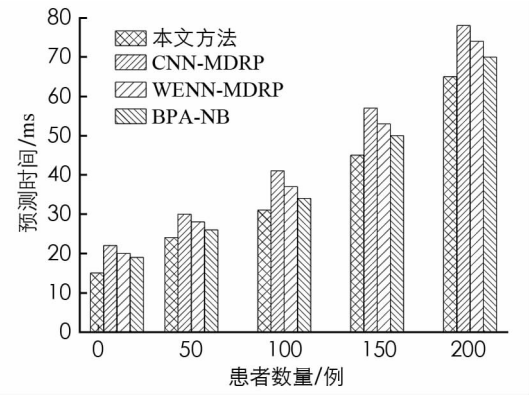

图 1　不同预测方法的预测时间对比结果

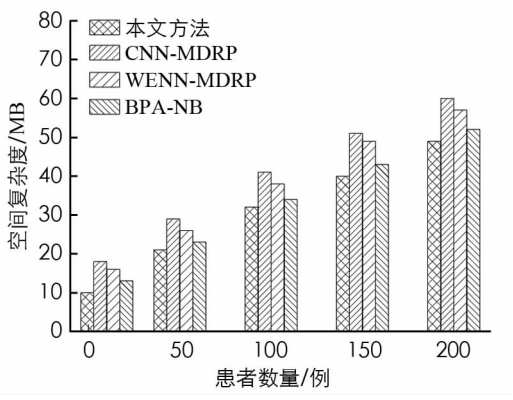

图 2　不同预测方法的空间复杂度对比结果

4　结　　语

在大数据分析中，数据分析技术存在精度和效率低，时间消耗多的问题. 为了克服这些局限性，本文提出一种基于格拉姆矩阵和随机森林的预测方法用于大型医疗数据分析，该预测方法由 4 个过程组成：①从大数据集中收集数据；②利用格拉姆矩阵对采集到的数据进行存储；③应用随机森林二元回归和分类技术，基于二元关系测度对未来结果进行预测；④利用决策树根据相关结果对数据进行分类，并提供了准确的预测结果. 实验结果表明，本文方法提高了预测精度，最大限度地减少了预测时间、误报率和空间复杂度.

参考文献：

[1] HOSSEINI M P, POMPILI D, ELISEVICH K, et al. Optimized Deep Learning for EEG Big Data and Seizure Prediction BCI via Internet of Things [J]. IEEE Transactions on Big Data, 2017, 3(4): 392-404.

[2] JINDAL A, DUA A, KUMAR N, et al. Providing Healthcare-as-a-Service Using Fuzzy Rule Based Big Data Analytics in Cloud Computing [J]. IEEE Journal of Biomedical and Health Informatics, 2018, 22(5): 1605-1618.

[3] ULLAH F, HABIB M A, FARHAN M, et al. Semantic Interoperability for Big-Data in Heterogeneous IoT Infrastructure for Healthcare [J]. Sustainable Cities and Society, 2017, 34: 90-96.

[4] DUBEY R, GUNASEKARAN A, CHILDE S J, et al. Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture [J]. British Journal of Management, 2019, 30(2): 341-361.

[5] CHEN M, HAO Y X, HWANG K, et al. Disease Prediction by Machine Learning over Big Data from Healthcare Communities [J]. IEEE Access, 2017, 5: 8869-8879.

[6] GU Y L, LU W Q, XU X Y, et al. An Improved Bayesian Combination Model for Short-Term Traffic Prediction with Deep Learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(3): 1332-1342.

[7] NAIR L R, SHETTY S D, SHETTY S D. Applying Spark Based Machine Learning Model on Streaming Big Data for Health Status Prediction [J]. Computers & Electrical Engineering, 2018, 65: 393-399.

[8] BABU S B, SUNEETHA A, BABU G C, et al. Medical Disease Prediction Using Grey Wolf Optimization and Auto Encoder Based Recurrent Neural Network [J]. Periodicals of Engineering and Natural Sciences (PEN), 2018, 6(1): 229.

[9] MOHAN S, THIRUMALAI C, SRIVASTAVA G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques [J]. IEEE Access, 2019, 7: 81542-81554.

[10] YAO L, GE Z Q. Distributed Parallel Deep Learning of Hierarchical Extreme Learning Machine for Multimode Quality

Prediction with Big Process Data [J]. Engineering Applications of Artificial Intelligence, 2019, 81: 450-465.
[11] 周传华, 柳智才, 丁敬安, 等. 基于filter+wrapper模式的特征选择算法 [J]. 计算机应用研究, 2019, 36(7): 1975-1979, 2010.
[12] 张 俐, 袁玉宇, 王 枞. 基于最大相关信息系数的FCBF特征选择算法 [J]. 北京邮电大学学报, 2018, 41(4): 86-90.
[13] NKUNDIMANA JOEL G, MANJU PRIYA S. Improved Ant Colony on Feature Selection and Weighted Ensemble to Neural Network Based Multimodal Disease Risk Prediction (WENN-MDRP) Classifier for Disease Prediction over Big Data [J]. International Journal of Engineering & Technology, 2018, 7(3): 56-61.
[14] VENKATESH R, BALASUBRAMANIAN C, KALIAPPAN M. Development of Big Data Predictive Analytics Model for Disease Prediction Using Machine Learning Technique [J]. Journal of Medical Systems, 2019, 43(8): 1-8.

Disease Prediction Method Based on Gram Matrix and Random Forest

ZOU Jin-song

Putian Big Data Industry School, Chongqing College of Water Resources & Electric Engineering, Chongqing 402160, China

Abstract: Aiming at the problems of low prediction accuracy, long prediction time and large storage space in the existing prediction methods, a disease prediction method based on Gram matrix and random forest has been proposed. In this method, at first, a large amount of data are collected from the data set, and then Gram symmetry matrix is used to store and classify the collected data. Then, the random forest binary regression and classification technology has been introduced to measure the relationship between the prediction results and the data through the correlation of binary variables, and a decision tree has been constructed to classify the results according to the correlation. At last, the voting scheme is used to output the final prediction result. The experimental results show that compared with other methods, the proposed method improves the prediction accuracy while it reduces the time and space complexity of prediction.

Key words: big data; Gram matrix; random forest; prediction analysis

责任编辑 夏 娟