河北工程大学

外 文 资 料 翻 译

姓　　名　袁留威

学　　号　190440227

指导教师　吕艳芬

专　　业　数据科学与大数据技术

学　　院　数理科学与工程学院

2023 年 4 月 1 日

一种利用机器学习和神经网络预测中风的预测分析方法

摘　要：中风对社会的负面影响导致了共同的努力来改善中风的管理和诊断。随着技术和医疗诊断之间的协同作用的增强，护理人员通过系统地挖掘和归档患者的医疗记录，为更好的患者管理创造了机会。因此，研究这些危险因素在患者健康记录中的相互依赖性，了解它们对卒中预测的相对贡献至关重要。本文系统地分析了电子健康记录中的各种因素，以有效地预测中风。利用各种统计技术和主成分分析，我们确定了中风预测的最重要的因素。我们的结论是，年龄、心脏病、平均血糖水平和高血压是检测患者中风的最重要的因素。此外，与使用所有可用的输入特征和其他基准测试算法相比，使用这四个属性的感知器神经网络提供了最高的准确率和最低的错过率。由于数据集在中风发生方面高度不平衡，我们在通过子采样技术创建的平衡数据集上报告了我们的结果。

关键词：预测分析　机器学习　神经网络　电子健康记录　中风

1 介绍

在技术的帮助下，我们见证了医学领域的惊人发展[1].随着医疗记录注释数据集的出现，我们现在可以使用数据挖掘技术来识别数据集中的趋势。这种分析有助于医生对任何医疗状况作出准确的预后。它改善了医疗保健条件，并降低了治疗成本。在医疗记录中使用数据挖掘技术对医疗保健和生物医学领域有很大的影响[2,3].这有助于医生在早期阶段确定疾病的发病情况。我们对中风特别感兴趣，并确定与中风发生相关的关键因素。

几项研究[4－7]分析了生活方式类型的重要性，患者的医疗记录对患者发生中风的概率的重要性。此外，机器学习模型现在也被用来预测中风的发生[8,9].然而，有没有研究试图分析所有与患者相关的情况，并确定卒中预测所必需的关键因素。在本文中，我们试图通过提供对各种患者记录的系统分析来弥补这一差距，以预测中风。利用 29072 名患者记录的公开数据集，我们确定了中风预测所必需的关键因素。我们使用主成分分析（PCA）将高维特征空间转换为低维子空间，并了解每个输入属性的相对重要性。我们还在患者记录数据集上对几种流行的基于机器学习的分类算法进行了基准测试。

本文的主要贡献如下-(a)，我们提供了对中风预测的各种危险因素的详细了解。我们分析了患者电子健康记录（EHR）记录中存在的各种因素，并确定了中

风预测的最重要因素；(b)我们也使用维度在特征空间的低维子空间中识别模式；和(c)，我们在一个公开的数据集中对用于中风预测的流行机器学习模型进行基准测试。我们遵循可重复性研究的精神，因此，在本文中使用的所有模拟的源代码都可以在网上获得。

本文的结构如下：第一部分提供了相关工作的概述，并描述了在我们的研究中使用的数据集；第二部分包括相关性分析和特征重要性分析；第三部分主成分分析的结果在第节中进行了解释；第四部分用于预测建模的数据挖掘算法及其在数据集上的性能详见章节；最后，第五部分总结了论文，并讨论了未来的工作。

1.1 相关工作

现有的文献工作已经对中风预测的各个方面进行了研究。Jeena 等人提供了各种危险因素的研究，以了解中风的概率[8].它使用了一种基于回归的方法来确定一个因素及其对中风的相应影响之间的关系。在 Hanifa 和 Raja [9]通过在非线性支持向量分类模型中应用径向基函数和多项式函数，提高了预测卒中风险的精度。本研究中确定的危险因素被分为四组——人口统计学组、生活方式组、医疗/临床组和功能组。同样，Luk 等人研究了 878 名中国受试者，以了解年龄是否对中风康复结果有影响[10].Min 等。在[11]开发了一种从潜在的可改变的危险因素中预测中风的算法。辛格和乔杜里在[12]已经使用了心血管健康研究（CHS）数据集上的决策树算法来预测患者的中风。并研究了一种基于前馈多层人工神经网络的深度学习模型[13]来预测中风。类似的工作也在[14–16]，以建立一个智能系统来预测中风的病人记录。Hung 等人在[17]比较了来自电子医疗索赔数据库的中风预测的深度学习模型和机器学习模型。除了传统的中风预测外，Li 等人在[18]使用机器学习方法来预测心房颤动中的缺血性卒中和血栓栓塞。

来自各种技术的结果表明，多种因素可以影响任何已进行的研究的结果。这些不同的因素包括数据收集的方式、所选择的特征、用于清理数据的方法、缺失值的计算、数据的随机性和标准化，这些都将对所进行的任何研究的结果产生影响。因此，研究人员确定电子健康记录中的不同输入因素如何相互关联，以及它们如何影响最终中风预测的准确性是很重要的。

在相关领域的研究[3,19]证明了识别重要特征会影响机器学习框架的最终性能。对我们来说，识别完美的特征组合是很重要的，而不是使用特征空间中所有可用的特征。如图所示 3 在使用分类算法之前，应该识别和删除一个类的冗余属性和/或完全无关的属性。因此，医疗保健领域的数据挖掘从业人员必须识别电

子健康记录中捕获的风险因素如何相互依赖，以及它们如何独立地影响中风预测的准确性。

1.2 电子健康记录数据集

电子健康记录（EHR）也被称为电子病历（EMR），是一个针对患者的信息存储库。它是一个自动化的，计算机可读的存储病人的医疗状态，由合格的医生输入。这些记录包括病人的生命体征、诊断结果或体检结果。未来 EHR 的最佳应用看起来很有前途。据统计，2009 年至 2014 年，美国医院对 EHR 的使用从 12.5%增加到 75.5%[20]。

在我们的研究中，我们使用了麦肯锡公司发布的电子健康记录数据集，作为他们的医疗保健黑客马拉松挑战的一部分。该数据集可以从 Kaggle 网站上获得，一个用于数据集的公共数据存储库。该数据集包含 29072 例患者的 EHR 记录。它总共有 11 个输入属性和 1 个输出特性。输出响应是一种二进制状态，指示患者是否有中风症状。EHR 中剩下的 11 个输入特征有：患者标识符、性别(G)、年龄(A)、患者是否患有高血压（HT）、患者是否患有心脏病（HD）的二进制状态、婚姻状况(M)、职业类型(W)、居住（城市/农村）类型（RT）、平均血糖水平（AG）、体重指数（BMI）和患者的吸烟状况（SS）。该数据集对于中风事件的发生情况高度不平衡；EHR 数据集中的大部分记录都属于没有患过中风的病例。数据集的发布者已经确保了与该数据相关的道德要求达到最高标准。在本文的后续讨论中，我们将排除患者标识符作为输入特征之一。在我们的研究和分析中，我们将考虑剩下的 10 个输入特征和 1 个响应变量。

2 分析电子健康记录

在本节中，我们将提供对电子健康记录数据集的分析。我们对这些特征进行了相关性分析。我们使用整个 EHR 记录的数据集来对 EHR 记录的输入特征进行这样的分析。相关分析对于特征选择的方法如下：如果两个特征具有很高的相关性，那么在预测卒中发生中可以忽略其中一个，因为它不能为预测模型提供任何额外的知识。此外，我们还评估了个体特征和群体中这些特征的行为，以观察每个个体特征在预测中风发生中的重要性。对输入特征空间的系统分析是行程预测的一个组成部分。寻找最优的和最小的预测特征集对于降低建模和有效的 EHR 记录存档的计算成本是很重要的。这为临床医生只记录 EHR 记录中那些对中风预测最有效的特征铺平了道路。

2.1 特征之间的相关性

我们使用皮尔逊相关系数来生成图.1，显示了不同患者属性之间的相关性。患者电子健康数据的任何两个特征之间的线性关系的强度将由这个相关值决定。我们已经用了一个彩色图图.1，这样蓝色表示正相关，而红色表示负相关。颜色越深，圆圈的大小越大，这两个患者属性之间的相关性就越高。

直观地看，属性与自身的相关性是统一的。患者的婚姻状况与年龄之间存在显著的相关性，相关指数为 0.5。患者年龄与工作类型呈正相关，患者是否患有高血压、心脏病与患者平均血糖水平相关指数为 0.38。患者的年龄与其他属性之间的相关性似乎是直观的，因为大多数疾病发生在老龄化人口中。患者的居住类型与任何其他任何属性无关。患者的工作类型与婚姻状况呈正相关，相关指数为 0.35。

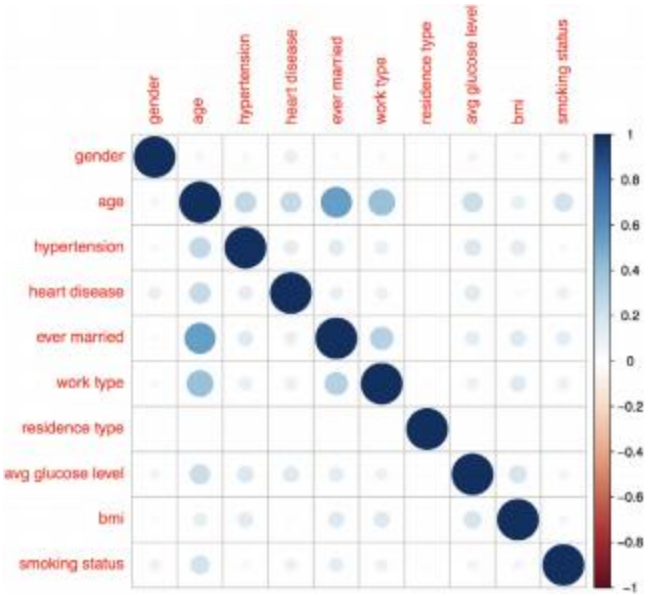


图 1 数据集中患者属性的相关矩阵。这些属性包括性别、年龄、患者患高血压为 0/1、患者患心脏病状态为 0/1、婚姻状况、工作类型、居住类型、平均血糖水平、体重指数和患者吸烟情况。

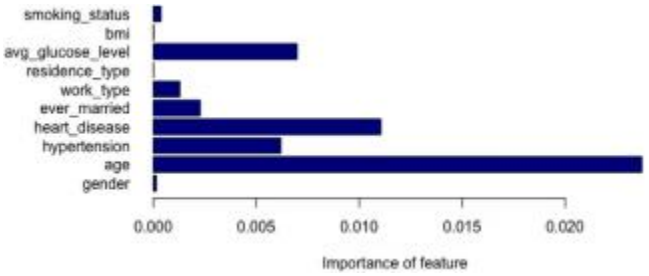


图 2 患者属性在使用线性向量量化（LVQ）模型预测卒中发生中的重要性。

综上所述，相关矩阵如图.1 告诉我们这些特征之间没有高度相关。因此，每个特征可能对中风预测都有各自的贡献。接下来的两个小节分析了单个特征对中风预测的重要性。

2.2 中风预测的个体特征

图.2 显示了每个患者的属性在使用学习向量量化（LVQ）模型预测卒中发生中的重要性。患者属性的相对重要性是通过由于该属性而导致的模型预测误差的增加来衡量的。我们使用了 R 插入符号包中的 varImp 方法 4 来计算这个相对特征的重要性。作为图.2 说明中，患者的年龄为(A)为在预测中风发生中具有最重要的特征。其他高度重要的特征是存在心脏病（HD）、患者的平均血糖水平（AG）和高血压的存在。

上述分析表明，患者的年龄(A)本身具有相对较高的重要性，但不同特征的组合可能会改善预测，因为它们彼此之间没有相关性。此外，我们还计算了这些字符 2 为 EHR 记录获得的分数。CHADS2 评分为非瓣膜性心房的卒中风险评分。S 代表纤颤，C 代表充血性心力衰竭或左心室功能障碍，H 代表患者是否有高血压，A 代表年龄，D 代表糖尿病，代表中风，或短暂性缺血发作，血栓栓塞史。

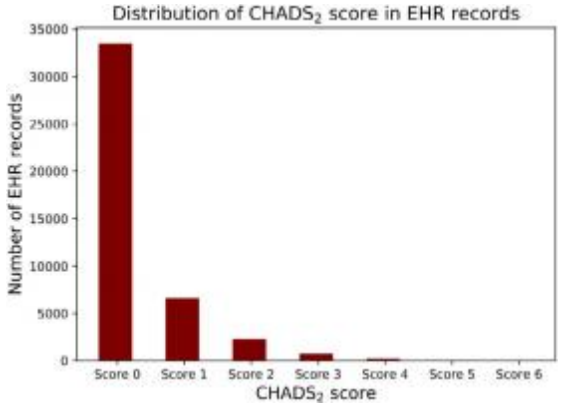


图 3 我们展示了字符的分布 2 在 EHR 记录数据集中的得分。我们观察到大部分的特征 2 分数值很低。

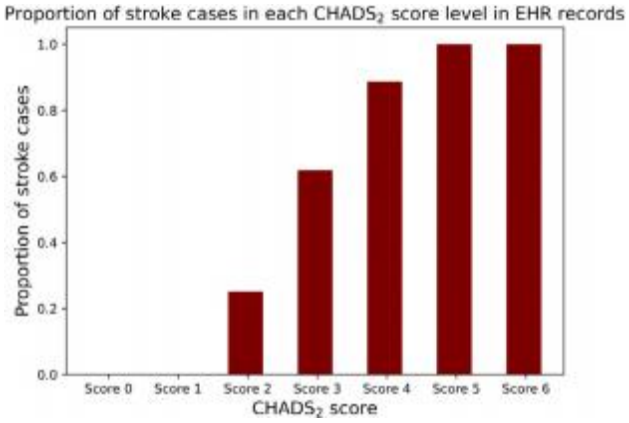
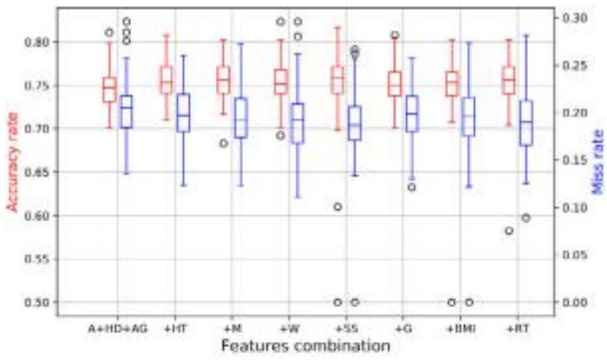


图 4 CHADS 预测的中风事件的比例 2 为每个分数水平分别。

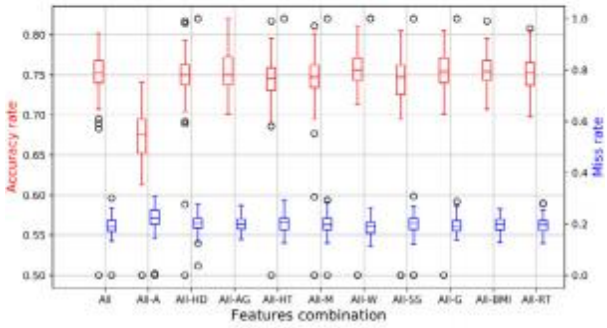
图.3 显示了我们的数据集的 CHADS2 分数的分布。我们观察到，大多数 EHR 观察结果的 CHADS2 得分较低。图.4 显示了 CHADS2 对每个评分水平所预测的发生中风事件的病例的比例。我们观察到，CHADS2 评分越大，卒中病例的发生率就越高。与图.3 和 4，我们发现大多数人都有聊天 21 分或 2 分，中风的可能性很低。我们观察到只有少数人有特征 2 得分大于 2 分，发生中风的概率越高。

2.3 对脑卒中预测的最佳特征的选择

在上一节中，我们一次使用一个特征，发现只有少数特征在中风预测中具有更高的重要性。在本节中，我们使用所有特征，然后一次删除一个特征或一次添加一个特征，并进一步分析单个特征在中风预测中的重要性。我们使用神经网络算法来产生结果。我们使用了一个感知器神经网络来进行这个实验。我们使用 EHR 记录的整个数据集来执行特征分析。



(a) Adding features



(b) Removing features

图 5 当(a)一次添加一个特征，一次删除一个(b)个特征时，我们测量了中风预测的准确率和错过率。箱形图代表了由 100 个实验计算出的指标的分布。

图.5(a)显示了随后每次添加一个特性的结果。第一个结果对应于使用特征 A、HD 和 AG。这些特性是选择从图.2，我们观察到这三个特征比其他特征具有更高的重要性。当 HT 添加到这个池中，我们看到准确率略有提高，错过率略有下降。请注意，HT 是从分析中得到的第四个重要变量图.2.随后，当其他特征被逐一添加时，正确率和遗漏率的变化非常小。所有剩余配置的精度和遗漏率几乎相同。因此，这说明 A、HD、AG、HT 四个功能可以是最优的功能，因为添加其他功能并没有提供任何改进。

图.5(b)显示了每次从所有特性池中删除一个特性的结果。在这里，我们可以观察到，当特征 A 从池中移除时，准确率会受到显著影响。这与我们之前的讨论一致，该讨论表明 A 在所有特征中得分最高(参考文献到图.2).当从池中删除其他特性时，没有多少重大变化。

3 主成分分析

在本节中，我们将使用主成分分析（PCA）来分析数据集中的方差。在这种多元分析中，数据集被转换为一组线性不相关的变量的值，称为主成分，以便从变量中提取最大的方差。这些主组件作为数据集特性的摘要。这些新的基函数没有物理解释。然而，这些新的基函数是原始特征向量的线性组合。在这项工作中，我们不限制自己在使用传统的特征消除技术的特征分析。然而，我们使用降维技

术将高维特征空间转换到二维特征空间上，以理解特征空间之间的相互关系。如果前几个成分捕获了数据中的大部分方差，那么 PCA 可以用于减少预测建模的特征空间。

我们使用主成分分析分析了低维子空间中的 10 维患者属性。我们提供了一个简短的关于主成分的入门书，用分析和数学方法来表述我们的问题陈述。假设 X 是维数为 m×n 的变量矩阵。

在本例中，m 表示 EHR 中输入属性的总数，n 表示数据集中的患者记录的总数。因此，m = 10 和 n = 29072 在本卒中预测分析中。我们对单个特征 f1−10 从矩阵 X，到相应的 j∈Rmn×1 其中 j = 1,2，……，10。最后，j 特征被堆叠在一起，以创建矩阵 X∈Rmn×10:

$$\hat{\mathbf{X}} = [\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \widetilde{\mathbf{v}}_3, \ldots, \widetilde{\mathbf{v}}_{10}] \tag{1-1}$$

我们对 X 的归一化矩阵进行主成分分析，使用相应的均值 vj 和标准差进行归一化 vj 个人特征。归一化矩阵 X…表示为：

$$\ddot{\mathbf{X}} = \left[\frac{\widetilde{\mathbf{v}}_1 - \bar{v}_1}{\sigma_{v_1}}, \frac{\widetilde{\mathbf{v}}_2 - \bar{v}_2}{\sigma_{v_2}}, \ldots, \frac{\widetilde{\mathbf{v}}_j - \bar{v}_j}{\sigma_{v_j}}, \ldots, \frac{\widetilde{\mathbf{v}_{10}} - \bar{v_{10}}}{\sigma_{v_{10}}} \right] \tag{1-2}$$

我们解释了 PCA 的结果如何与患者属性代表的预测变量或特征和医疗健康记录代表的个体观察结果相关。我们研究了前两个主成分和个体输入变量之间的关系。我们还研究了前两个主成分对一个给定的观察结果的重要性。我们使用阿卜迪和威廉姆斯的指南为这项研究。

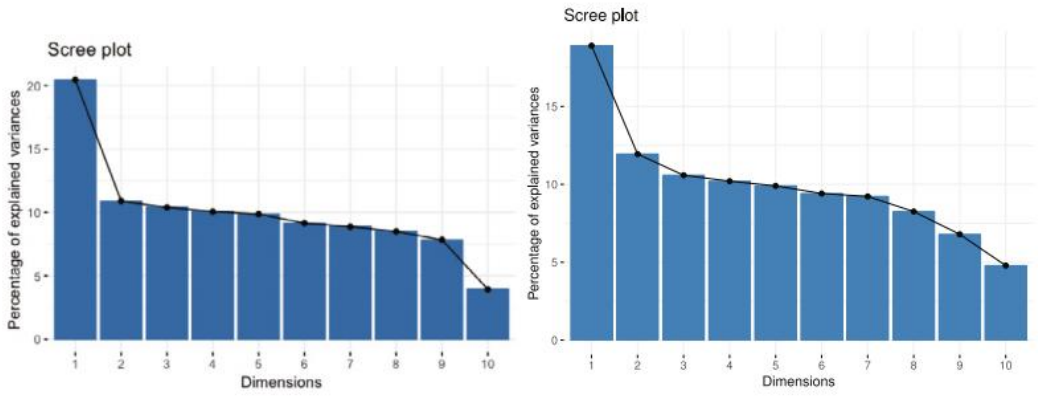
3.1 由主成分解释的方差

碎石图用于选择可以解释数据集中最大部分可变性的组件，通常是 80%或更多的方差。图.6(a)显示了由原始数据集中不同主成分解释的数据集中的方差百分比。在这里，我们可以观察到，由不同的主成分解释的方差非常低。在 10 个主成分中，需要 8 个来解释 88.2%的方差。平衡的数据集意味着数据集相对于使用随机采样的笔画标签是平衡的。原始数据集是不平衡的，因为有更多的样本具有负值中风标签，与阳性标签中风样本进行比较。我们通过考虑所有的阳性中风样本，然后从其余样本中随机选择相同数量的阴性中风样本来保持平衡。这将使一个平衡的数据集与相同数量的阳性和阴性中风样本。

所有的主成分都是相互正交的，因此也不相关。因此，每个单独的 PC 都可以用来解释一个独特的现象。的分布方差图.6 表明不同的主成分解释了不同的潜

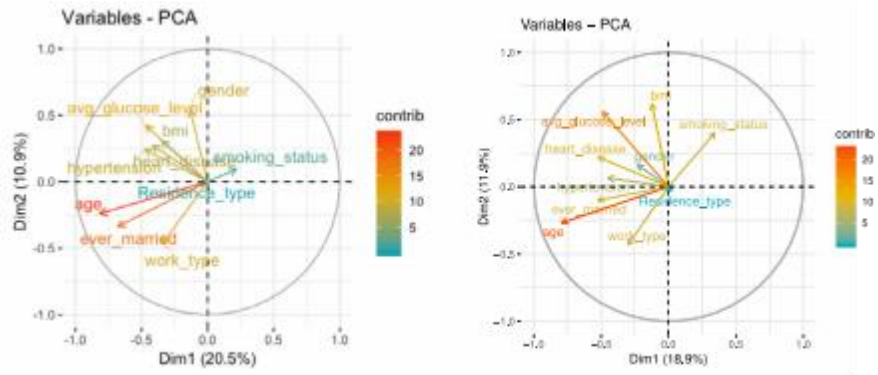
在现象。这些现象可以根据可变负载进行分析。变量负荷是指不同变量对单个主成分的贡献。我们还包括了平衡数据集的碎石图图.6(b).这对研究人员理解数据集的不平衡性质对子空间表示的影响是很有用的。表 1 显示了每个变量对前两个主成分的贡献。单个主成分的所有负荷的平方和等于单位。因此，如果变量加载超过一个阈值为 $\sqrt{1/10} = 0.31$，这表明该变量对主成分有很大的贡献。在表 1，超过阈值的变量用粗体表示。这里我们观察到像 A、AG、HT 和 M 等变量对第一个主成分有很大的贡献。变量 HD 也显示出了重要的贡献。从早期的讨论中，我们发现，从中风预测的角度来看，这些特征实际上很重要。因此，这表明第一个主成分（从这些变量中有更强的负荷）可能有助于预测中风。

在接下来的章节中，我们将评估这些主成分在中风预测中的作用。



(a) (b)

图 6 由不同主成分解释的方差百分比。我们展示了(a)原始的，(b)平衡的数据集的碎石图。



(a) (b)

图 7 在前两个主成分上的输入属性的双曲线图表示。我们展示了(a)原始，(b)平衡数据集的双图。

表 1 第一和第二个主成分中不同特征的贡献

Features	PC1	PC2
gender	0.092	0.516
age	0.571	0.230
hypertension	0.331	0.232
heart_disease	0.290	0.261
ever_married	0.475	0.322
work_type	0.236	0.442
residence_type	0.001	0.003
avg_glucose_level	0.326	0.403
bmi	0.242	0.293
smoking_status	0.152	0.090

3.2 主成分与患者属性之间的关系

图.7 描述了双曲线图的表示方式。它显示了不同的输入属性如何与其他输入属性相关联，也依赖于第一和第二主成分。x 轴和 y 轴分别表示第一和第二主成分。每个向量表示一个输入属性，其长度表示它的重要性。我们观察到平均血糖水平和心脏病是相互相关的。“年龄”在中国前两个主成分。我们还观察到，不同的特征向量在二维特征空间中的方向与不平衡的数据集是相同的。

图.7(a)说明，患者的居住类型对这两个主要成分的贡献最小的。我们还观察到，年龄和婚姻地位是相互关联的，并且对第一主成分有很大的贡献。患者的吸烟状况及其平均血糖水平与两者均呈正交关系其他，表明它们向特征空间提供了不同的信息。然而，吸烟状况和年龄点彼此相反，表明它们提供的信息相似，但有不同的特征。我们还计算了平衡数据集上的 EHR 记录的双曲线图。我们展示了相应的双图图.7(b).我们观察到平均血糖水平和心脏病是相互相关的。年龄对前两个主要成分的贡献最大。我们还观察到，不同的特征向量在二维特征空间中的方向与不平衡的数据集是相同的。

3.3 主成分与个别记录的关系

最后，我们还检查了个体患者记录观察结果对前两个主成分的关系。在图.8，每个点代表一个病人记录观察。

图.8(a)显示了主成分在每个记录上的重要性。这通常用 cos2 测度来表示，

表示距离原点的平方距离。这表明，具有高 cos2 值的观测值可以用主成分来表示，而不是具有低 cos2 值的观测值。我们观察到，大量的点聚集在原点附近，这不能完全用主成分表示。因此，主成分只能表示特征空间中整个信息的一部分。

图.8(b)颜色根据中风的状态对患者的记录进行编码。由于数据集是高度不平衡的，我们观察到大多数观察结果的颜色编码为中风的负状态。我们观察到一些中风阳性的观察（1 标签观察）。然而，这些积极的中风观察结果并不位于集群中。这表明，高维特征对于分离观测结果是必要的。

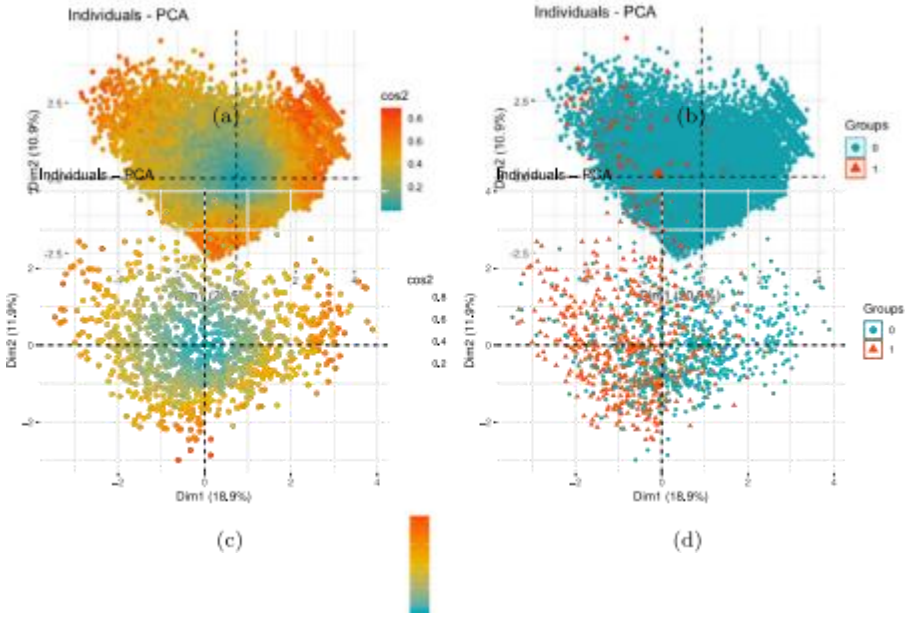


图 8 参照前两个主成分的不同患者记录的子空间表示。观察结果基于(a) cos2 测量，表明主成分对观察的重要性；以及中风的(b)状态。我们还使用平衡数据集，并观察(c)和(d).中的子空间表示

我们还计算了平衡的 EHR 数据集的不同特征的子空间表示。我们把它放在图.8(c)和(d).我们观察到，关于中风和没有中风的观察结果分散在两个主轴上。具有相似标签的观察结果并没有聚集在一起。因此，EHR 记录的特征对于有效的中风预测具有重要意义。

3.4 讨论

本节讨论了主成分分析如何帮助清晰地理解患者记录的原始特征空间。我们表明，前两个主成分只能累积捕获输入特征空间中总方差的 31.4%。更重要的是，前八个成分只能解释总方差的 88%左右。此外，我们还研究了不同患者属性对前

两个主成分的贡献。我们可以看到，这两个组件并不能完美地表示健康记录数据。我们还研究了前两个主要成分在个人健康记录表示中的贡献。我们可以看到，一些健康记录可以用这两个组成部分来表示，但有些则不能。因此，需要所有的主成分都能很好地表示医疗记录中的差异。我们不能在不显著损失的情况下显著减少预测建模的特征空间。因此，我们使用所有的主成分来预测中风的发生模型。

在下一节中，我们将比较在患者的医疗记录中预测中风发生的先进机器学习分类技术。如前所述，我们使用 10 个患者属性作为模型的输入特征。

4 中风预测

在本节中，我们提供了对中风预测中的各种基准测试算法的详细分析。我们对三种流行的分类方法进行了基准测试——神经网络（NN）、决策树（DT）和随机森林（RF），用于从患者属性中预测卒中。决策树模型是一种流行的二值分类算法。这种方法包括构建一个具有多个条件测试的类似树的决策过程，然后将该树应用到医疗记录数据集。该树中的每个节点代表一个测试，其分支对应于测试的结果。叶节点最终表示类标签。该算法的剪枝能力使其灵活、准确，这是医学诊断所必需的。我们还以随机森林的方法对数据集进行了基准测试。随机森林算法的灵活性和易用性，加上它产生良好结果的一致性，即使对超参数的最小调整，使该算法在这个应用中很有价值。过度拟合的可能性受到森林中存在的树木数量的限制。此外，随机森林还可以提供足够的指标来说明它为这些输入变量分配重要性的方式。我们还对一个两层浅层神经网络的性能进行了基准测试。人工神经网络现在非常流行，他们提供了有竞争力的结果。我们使用 nnet R 包实现了前馈多层感知器模型。

4.1 使用所有功能进行基准测试

我们的数据集总共包含 29072 份医疗记录。其中，只有 548 条记录属于卒中患者，其余 28524 条记录没有卒中患者。这是一个高度不平衡的数据集。这就给直接使用这些数据来训练任何机器学习模型带来了问题。因此，我们使用随机降采样技术来减少数据集的不平衡性质的不利影响。我们将 548 条记录作为少数类，其余的 28524 条没有中风情况的记录作为多数类。随后，我们创建了一个 1096 个观测数据集，其中包括 548 个少数样本和 548 个多数样本。这个平衡的数据集是通过考虑所有的 548 个少数民族样本而创建的，其余 548 例多数样本从 28524 例患者记录中随机抽取。所有三种机器学习模型都是在 1096 个观测的平衡数据集上训练的。

表 2 神经网络、决策树和随机森林的性能评估

Features	Way	Precision	Recall	F-score	Accuracy	Miss rate	Fall-out rate
Original features (All)	DT	0.75	0.74	0.74	0.74	0.17	0.24
	RF	0.74	0.73	0.73	0.74	0.18	0.25
	NN	0.80	0.74	0.77	0.77	0.16	0.18
	CNN	0.74	0.72	0.73	0.74	0.17	0.24
	SVM	0.67	0.68	0.68	0.68	0.23	0.32
	LASSO	0.78	0.72	0.75	0.76	0.19	0.20
	ElasticNet	0.79	0.71	0.75	0.76	0.19	0.19
Original features (A+HD+AG+HT)	DT	0.78	0.71	0.74	0.75	0.20	0.21
	RF	0.76	0.74	0.75	0.75	0.18	0.24
	NN	0.78	0.71	0.74	0.75	0.19	0.20
PCA features (PC1 and PC2)	DT	0.78	0.65	0.71	0.73	0.24	0.19
	RF	0.71	0.68	0.69	0.69	0.23	0.28
	NN	0.77	0.67	0.72	0.74	0.22	0.20
PCA features (PC1 till PC8)	DT	0.75	0.68	0.72	0.73	0.21	0.23
	RF	0.73	0.69	0.71	0.72	0.21	0.25
	NN	0.80	0.68	0.73	0.75	0.21	0.17

表 3 神经网络、SVM、套索和弹力网的精度方差

Features	Way	Precision	Accuracy variance
Original features (All)	NN	0.80	0.000377
	SVM	0.67	0.000470
	LASSO	0.78	0.000380
	ElasticNet	0.79	0.000467

表 4 CNN 模型的超参数

Layers	In channels/Out channels	Kernel, Stride, Padding	Activation functions
Gonv1	1/16	3, 1, 1	ReLu
Gonv2	16/8	2, 1, 0	ReLu
Layers	In channels/Out channels	Kernel, Stride, Padding	Activation functions
Linear1	32/16	-	ReLu
Linear2	16/1	-	Sigmoid

在我们的实验中，另一种深度学习方法，卷积神经网络（CNN）实现了预测中风。在我们的配置中，隐藏层数为 4 层，前两层为卷积层，后两层为线性层，在中给出了 CNN 模型的超参数表 4.我们在 CNN 的训练和测试过程中使用相同的训练和测试分割，将 10 个输入特征重塑为 1 * 2 * 5 作为输入。

我们还计算了基准测试方法的精度方差。表 3 给出了神经网络、SVM、套索和弹性网的精度方差实验。

4.2 使用前四大功能进行基准测试

在这里，我们给出了当使用所有特征和只使用 4 个特征（A、HD、AG 和 HT）时的中风预测结果。这些特性是根据我们之前的讨论而选择的。除了这些特征外，我们还显示了当主成分作为输入时的行程预测结果。由于我们观察到需要近 8 个主成分来解释大于 80%的方差，所以我们提出了只使用前 2 个主成分和使用所有成分时两种情况的结果。表 2 显示了所有不同配置的评估指标。为了消除抽样偏差，我们进行了 100 个随机降采样实验。训练观察数和测试观察数的比例为 70： 30。表 2 报告所有方法的平均值。

我们观察到，在不同的机器学习方法中，5 神经网络对不同的特征组合具有较好的工作精度。当我们比较使用所有特征和只使用 4 个特征的情况下的神经网络结果时，我们没有观察到所有特征超过 4 个特征的显著改善。因此，仅使用 4 个特征（A、HD、AG 和 HT），我们就可以获得高达 78%的中风预测精度，低遗漏率为 19%。这一结果可能不足以指导个人层面的治疗和预防措施，但它将有助于支持人口和/或队列层面的分配和资源。

这里，当主组件作为输入的情况，我们观察到，当所有组件只使用前两个组件时，神经网络的准确性只有轻微的提高。这与我们之前的讨论一致，其中我们说明了从中风预测的角度来看，第一个主成分可能是重要的。对第一个成分贡献最大的变量显示出了良好的中风预测可能性。因此，与使用所有组件时的情况相比，只使用前两个组件具有相似的结果。

当我们将主成分结果与使用实际特征的情况进行比较时，可以观察到，这两种情况下的神经网络的准确性是可比较的。然而，如果我们看错过率，在主成分的情况下，错过率更高。漏诊率是一个重要的评估指标，因为我们希望能够检测所有中风。我们预计错过率将尽可能低，而错过率值的轻微下降对我们的进一步考虑是很重要的。因此，我们的分析表明，以 4 个重要特征（A、HD、AG 和 HT）为输入的神经网络，可以获得脑卒中预测的最佳结果。

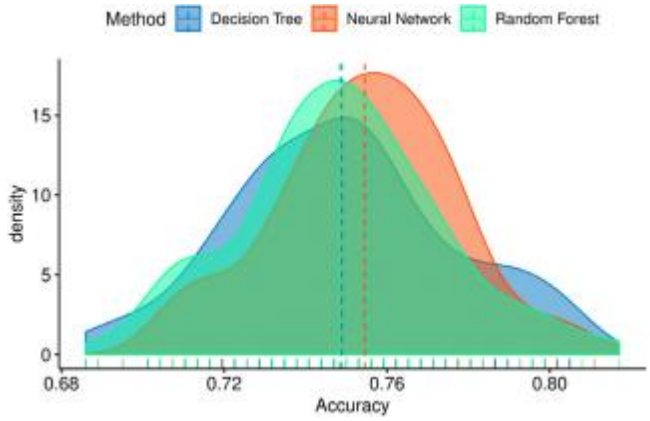


图 9 采用年龄、心脏病、平均血糖水平、高血压等前 100 个实验得到的分类精度直方图分布

最后，我们通过使用数据集中的年龄、心脏病、平均血糖水平、高血压等前 4 个特征来说明准确性值的分布。我们进行了 100 次实验，以消除训练集和测试集中的任何抽样偏差。图.9 说明这一点。我们观察到，它们中的大多数具有相似的性能，它们的平均值在相似的值附近重叠。我们还计算了 100 个随机子抽样观测值的基准测试方法的准确性的方差。决策树、神经网络和随机森林的方差分别为 0.00073, 0.00049 和 0.00061。这表明，在基准测试结果中不存在抽样偏差。

5 结论和未来的工作

在本文中，我们详细分析了患者在电子健康记录中的属性预测。我们系统地分析了不同的特征。我们进行了特征相关分析和逐步分析来选择最优的特征集。我们发现不同的特征并没有很好的相关性，只有 4 个特征（A、HD、HT 和 AG）的组合可能对脑卒中预测有很好的贡献。此外，我们还进行了主成分分析。分析表明，几乎所有的主成分都需要解释较高的方差。然而，变量负荷表明，方差最

高的第一个主成分可能解释了脑卒中预测的潜在现象。最后，在一组不同的特征和主成分配置上实现了三种机器学习算法。我们发现，神经网络对 A、HD、HT 和 AG 的特征组合效果最好。该组合的准确率和漏检率分别为 78%和 19%。

我们只使用 4 个特性就看到了有希望的结果。感知器模型的准确性不能进一步提高，主要原因有两个：缺乏额外的区分性特征集；以及缺乏额外的数据集。我们观察到，EHR 数据集中的大部分现有特征彼此之间高度相关，因此没有向原始特征空间添加任何额外的信息。此外，一个更大的数据集将使我们能够更有效地训练我们的深度神经网络。我们计划在我们计划的未来工作中收集机构数据。对电子健康记录中的不同特征进行系统的分析，将有助于临床医生有效地归档这些记录。数据管理团队不能记录和存储所有的特征，而是只能存档那些对中风预测至关重要的特征。因此，在未来，我们计划利用语义网络技术，将电子记录数据集与不同疾病和药物的背景知识集成起来[22,23].知识图技术[23,24]可用于以一种可互操作的方式向研究界发布电子健康记录。增加来自其他数据集的背景知识也可能提高脑卒中预测模型的准确性。

我们打算收集我们的机构数据集以进行进一步的基准测试这些用于中风预测的机器学习方法。我们还计划对我们提出的方法进行外部验证，作为我们即将进行的计划工作的一部分。