案例使用说明

|  |  |  |
| --- | --- | --- |
| **序号** | **具体事项** | **内容填写基本要求** |
| 0 | 案例库名称 | 工业大数据分析与工程应用案例库 |
| 1 | 案例名称 | 药物的血脑屏障通过率预测分析 |
| 2 | 教学目的与用途 | 适用的课程：《工业大数据处理与分析》  适应专业：控制科学与工程学科各专业方向  教学目标：通过此案例的学习，使学生了解药物分子通过血脑屏障的重要性；了解药物分子血脑屏障通过率预测对工业药物研发的重要性；了解血脑屏障的定义；了解基于机器学习在工业药物研发中的重要作用。 |
| 3 | 涉及知识点 | 本案例涉及基于机器学习的特征提取、机器学习模型、预测性能评价指标等知识点。 |
| 4 | 配套教材 | 无 |
| 5 | 启发思考题 | 1. 常规的多元统计方法和机器学习方法有哪些？ 2. 为什么要进行特征提取？常规的特征提取方法和机器学习分类模型分别有哪些？ |
| 6 | 分析思路 | 本案例分析的逻辑路径为：首先对血脑屏障问题进行介绍，分析大数据时代利用机器学习进行提前预测的好处和一般步骤；然后从药物分子本身出发，考虑输入变量预测的不同影响，基于分子指纹和特征提取，建立各种机器学习模型进行拟合学习；最后，基于auccary对预测结果进行分析。 |
| 7 | 理论依据 | 本案例设计的相关理论包括：血脑屏障、分子指纹、特征提取、机器学习模型。 |
| 8 | 背景信息 | 选用本案例的教师，要求对血脑屏障、机器学习特征提取，机器学习模型等知识有深刻了解。 |
| 9 | 关键要点 | 本案例分析中涉及的关键知识点是：血脑屏障、特征提取、机器学习模型等。 |
| 10 | 建议课堂计划 | 本案例用于研究生课程教学，建议讲述时间：1个课时；  本案例用于本科生课程教学，建议讲述时间：2个课时 |

案例名称：药物的血脑屏障通过率预测分析

作者姓名：杨森

作者单位：中南大学湘雅药学院

案例版权：中南大学

目录

[首页注释 1](#_Toc82415999)

[摘要 2](#_Toc82416000)

[**1.** **引言** 3](#_Toc82416001)

[**2.** **案例背景** 3](#_Toc82416002)

[**3.** **基于不同特征的药物分子血脑屏障通过预测** 5](#_Toc82416003)

[3.1 特征提取 5](#_Toc82416004)

[3.2 机器学习模型 7](#_Toc82416005)

[3.3 训练流程 9](#_Toc82416006)

[3.4 预测结果分析 11](#_Toc82416007)

[**4.** **结束语** 14](#_Toc82416008)

[参考文献 14](#_Toc82416009)

# 

# 首页注释

本案例来源于工业药物研发中的关键指标预测问题，具体涉及基于不同特征的药物分子血脑屏障通过预测，主要包括药物分子的特征选择和提取，三种机器学习分类模型的建立，机器学习模型的训练和测试等内容。

本案例内容未涉及任何保密技术和商业机密，无知识产权纠纷问题。

与本学科相关的主要知识点包括：药物分子的特征选择和提取、机器学习分类模型、预测性能评价指标等知识点。

本案例适应于控制科学与工程各专业方向：应用于《工业大数据处理与分析》、《人工智能与模式识别》等研究生课程。

本案例由中南大学湘雅药学院杨森负责整理与编写。

**药物的血脑屏障通过率预测分析**

摘要：在工业药物研发中，血脑屏障（BBB）的通过率是否合格是新药开发的关键环节之一。血脑屏障是人脑内部保护中枢神经系统的重要关卡，因其对药物分子的选择性极高，所以能否穿过血脑屏障直接决定了药物是否可以发挥针对脑部疾病的疗效。然而，传统的实验室实验方法通常耗时长、成本高、人力成本较高，难以满足现代药物研发的高效率要求。机器学习模型通过分析分子特征与血脑屏障通透性之间的复杂关系，可以实现快速、高效且可靠的预测，对候选化合物的筛选有一定的指导意义，有助于降低研发成本并加速新药上市。本案例结合不同的特征提取方法和三种机器学习分类方法对预测过程进行简单分析。

**关键字：**血脑屏障；特征提取；机器学习

**Prediction and analysis of blood brain barrier passing rate of drugs**

**Abstract:** In the research and development of industrial drugs, the prediction of blood-brain barrier (BBB) passing rate is one of the key links in the development of new drugs. As an important barrier to protect the central nervous system, blood-brain barrier has high selectivity for drug molecules. Whether it can cross the blood-brain barrier directly determines whether drugs can play a role in the treatment of brain diseases. However, the traditional experimental methods are usually time-consuming and costly, which are difficult to meet the high efficiency needs of modern drug research and development. By analyzing the complex relationship between molecular characteristics and blood-brain barrier permeability, machine learning model can achieve rapid, efficient and reliable prediction, and provide an important reference for the screening of candidate compounds, so as to significantly reduce the cost of research and development and accelerate the process of new drug listing. This case simply analyzes the prediction process by combining different feature extraction methods and various machine learning methods.

**Keywords:** Blood-Brain Barrier; Feature Selection; Machine Learning

1. **引言**

现代工业药物研发一直是耗时耗力，大投入高风险的产业，据统计，一款药物的研发平均需要花费10亿美元到20亿美元不等，且时间上最长需要15年[1]。工业药物研发过程耗时耗力的主要原因是复杂的生物学和化学因素、高昂的实验成本和安全性和有效性验证。其中安全性验证需要大量实验室实验，如果能加速这一过程有望大幅降低人力成本和经济成本。

目前的工业界药物研发追求的是高效率和低风险低成本，与药物研发的内在性质相差较多，而相关计算机预测方法则能一定程度地趋利避害，提高新药进入市场的成功率，在药物研发的早期阶段将不符合要求的药物提前排除出去，从而降低成本，加快研发进度。传统的药物研发流程通常需要大量的时间和财力投入，并且具有较高的失败率。随着计算能力和数据科学的快速发展，机器学习在药物研发中的应用为该领域带来了前所未有的变革。在药物研发领域，计算机的应用呈现出一定的规律性，最初的计算机辅助药物设计，到现代人工智能方法的引入，为生物科学提供了高度可靠的计算方法[2]。机器学习作为人工智能领域的一个分支，经过了多年研究，目前发展较为成熟，更多的突破则出现在深度学习阶段。而在本案例中，我们将通过结合不同的特征选择和机器学习算法来预测药物血脑屏障通过率，在工业研发过程中能够显著加速项目进程并降低相关成本。

1. **案例背景**

血脑屏障是位于人类大脑中的一种屏障，由于它存在的最大作用是保护人体的重要中枢神经系统，因此它的选择性非常高，包括药物和毒素在内的大多数物质难以成功进入大脑。与此相对的是，由于选择性较高的血脑屏障的存在，神经系统疾病的药物治疗也因此存在着巨大的挑战，因为它选择地过滤了许多药物能够顺利地进入到大脑中发挥疗效。这一特性使得许多潜在的治疗药物无法通过血脑屏障，降低了治疗效果和药物的临床应用潜力[3]。

血脑屏障的特殊性主要体现在其严密的细胞间连接、选择性通透性以及对不同化学性质物质的排斥。传统的药物研发方法往往依赖于实验室动物模型来评估药物是否能够通过血脑屏障，这一过程不仅费时费力，而且具有一定的局限性。随着高通量数据的积累和计算技术的进步，基于计算模型的血脑屏障预测成为一种越来越重要的手段。通过构建精准的预测模型，研究人员能够快速评估分子是否具备通过血脑屏障的潜力，减少实验中的不必要的重复性测试[4]。

为利用机器学习对药物血脑屏障通过率进行预测，需要获取输入模型的特征。本案例中药物分子的输入特征采用分子指纹和其他相关分子属性，并使用不同先进的机器学习分类模型对输入特征进行处理，并计算准确率对结果进行分析。

1. **基于不同特征的药物分子血脑屏障通过预测**

机器学习是人工智能领域发展较为成熟的一个分支，它通过构建不同具有先验知识的模型，在各自适合的领域拟合研究人员输入的数据，并对从未见过的测试数据进行相关的预测。在药物分子的血脑屏障穿透性预测中，机器学习提供了强大的工具来处理复杂的结构数据，并从中发现潜在的规律。先进的机器学习分类模型包括：RF[5]、SVM[6]、XGB[7]。本案例中，对不同的特征利用上面三种模型进行预测。

## 3.1 特征提取

本文采用的特征为分子指纹和其他相关属性。其中，药物分子的Morgan分子指纹是一种用于表示化学分子结构的二进制向量，它的计算方式是利用递归编码分子的原子及其邻域的结构特征来捕捉分子的局部信息。该方法从每个原子出发，经过多轮迭代，将原子的邻域信息哈希并合并为一个固定长度的位向量，通常为1024位或2048位。Morgan指纹广泛应用于化学相似性搜索、分子聚类、定量构效关系建模（QSAR）等领域，并且也能够计算指纹之间的相似度，从而非常高效地比较分子结构的相似性，以验证分子的创新性，从而避免某些专利问题。因其灵活的结构捕捉能力和较低的计算消耗使其成为化学信息学中的重要工具[8]。另一个特征提取方法为计算分子的相关属性，包括分子量，分子极性，氢键受体和供体数量，分子的极性表面积与水和辛醇中的分配系数，本案例选取了上面提高的各个属性将它们分别计算后作为另一种特征提取方法与分子指纹特征进行对比，将他们两种分子特征输入机器学习分类模型中计算预测准确度。

## 3.2 机器学习模型

随机森林是一种十分先进的机器学习分类算法，它的基础模型是决策树模型，基于和而不同的思想，随机森林利用多个模型的结果极大地提高预测的准确性和稳定性。每棵决策树都是在不同的训练集上独立训练的，训练集通过自助采样（bootstrap sampling）从原始数据集中随机抽取，同时在树的构建过程中随机选择特征，避免了单一决策树可能出现的过拟合问题。最终，随机森林通过投票（对于分类问题）或平均（对于回归问题）的方法来结合各棵树的预测结果，从而得到更为准确和鲁棒的模型。虽然随机森林在许多应用中表现出色，能够处理高维数据和复杂模式，但它的计算资源消耗较大，且模型本身的可解释性较差。不过，它仍因其高准确性、良好的泛化能力和在实际应用中的高效性，广泛应用于分类、回归、特征选择等任务，成为机器学习中的常用工具[5]。

支持向量机是十分早期提出的一种监督学习算法，到目前为止仍然是分类算法中性能最好的一批模型。这个算法的思路是通过在高维度的线性空间中找到一个分类超平面，从而尽可能地将不同类别的数据样例分开，以此实现高效鲁棒的分类目标。它的优化目标是使得超平面与两类数据点之间的间隔最大化，这样可以提高模型的泛化能力，减少过拟合的风险。对于线性不可分的数据，SVM引入了核函数（Kernel Function）将数据映射到高维空间，在高维空间中寻找一个能够分开不同类别数据的超平面。此外，SVM还通过引入“软间隔”概念，允许一定的误分类，从而使得模型在面对噪声数据时更加鲁棒。尽管SVM在处理高维数据和复杂边界时表现优异，但它的训练过程可能较为缓慢，尤其是在样本数量较大时，且模型的选择（如核函数的选择）和调参过程可能较为复杂。尽管如此，SVM凭借其在许多实际应用中展现出的高精度和稳健性，仍然是数据科学和机器学习领域的常用算法之一[6]。

XGBoost是另一种十分先进的机器学习算法，也属于集成学习的一个高效模型，它也被广泛应用于分类和回归任务。它是梯度提升树（GBDT）的一种优化实现，旨在通过集成多个弱学习器（通常是决策树）来提升模型的准确性。XGBoost的核心思想是通过迭代训练一系列决策树，每次训练时重点学习前一棵树未能正确预测的数据点。与传统的梯度提升方法相比，XGBoost在性能和计算效率上进行了诸多改进，包括采用正则化项来控制模型复杂度，从而减少过拟合的风险；同时，利用了更高效的分裂查找算法和缓存优化，使得训练速度大幅提升。此外，XGBoost支持处理缺失值，并能自动优化树的结构，使得其在大规模数据集和高维特征空间中表现尤为出色。XGBoost不仅具有较高的准确性和鲁棒性，还因其可调的超参数、良好的可扩展性和适应性，被广泛应用于金融风控、广告点击率预测、医学影像分析等领域。然而，XGBoost的训练过程需要一定的计算资源，且模型的调参较为复杂，需要针对不同任务进行仔细的调整。尽管如此，XGBoost凭借其卓越的性能和在许多比赛中取得的优异成绩，已成为机器学习领域中的一项重要工具[7]。

## 3.3 训练流程

将药物分子数据和对应能否通过血脑屏障标签收集之后将数据集分为训练集和测试集，利用不同的特征提取方法提取到特征，每种特征分别输入三种机器学习模型中得到结果并进行预测，对比各种特征提取和机器学习方法的准确度结果，具体训练流程如下：

（1）数据收集。假定样本数据既包含输入也包含输出，则数据集可表示为，其中为输入，为标签输出。对于数据集，基于上文提到的两种特征提取方法构造分别构造两组输入变量的特征数据集。

（2）将完整数据集按照8：2的比例分割为训练集和测试集。

（3）建立三种机器学习模型，将训练集输入机器学习模型进行参数拟合。

（4）训练结束后，将测试集输入三种机器学习模型中进行预测，并计算准确度。

三种机器学习模型训练完成之后根据准确度结果对模型性能进行评估。本案例中，我们使用准确度auccary评价预测性能。一般地，auccary越大，预测性能越好。

## 3.4 预测结果分析

本案例中的模拟数据来源自https://github.com/ardigen/MAT/tree/master/data/bbbp/bbbp.csv

在如下的表1展示了三种分类模型的分类准确性，从结果中可以看出，分子指纹相较于人工特征选择的分子属性在预测结果上表现更好，在三个模型上分子指纹作为特征都优于手动提取的特征，其中SVM上表现差距较大。另外从模型选择的角度观察，XGB在三种模型中表现最好，在处理本案例的结构化数据时更加适配。。

表 1 三种分类模型在测试集的预测准确度

|  |  |  |  |
| --- | --- | --- | --- |
| 特征提取 | RF | SVM | XGB |
| 分子指纹 | 0.85 | 0.85 | 0.87 |
| 相关属性 | 0.83 | 0.77 | 0.83 |

1. **结束语**

在新药研发的工业界，血脑屏障通过率的检测是新药研发中能否成功发挥疗效的关键性实验。在本案例中，通过利用三种机器学习分类模型即随机森林，支持向量机和梯度提升树，对两种特征提取方法即分子指纹和人工构造分子属性指纹来作为模型输入，对药物分子的血脑屏障通过率进行预测分析，准确率都相对较高，并且案例分享了特征选择和模型构建的基本流程，并通过对比得到了分子指纹和XGB作为相对优秀的特征提取方法和机器学习分类模型。

# 参考文献

[1] R. C. Mohs and N. H. Greig, “Drug discovery and development: Role of basic biological research,” *Alzheimers Dement. Transl. Res. Clin. Interv.*, vol. 3, no. 4, pp. 651–657, Nov. 2017, doi: 10.1016/j.trci.2017.10.005.

[2] C. Selvaraj, I. Chandra, and S. K. Singh, “Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries,” *Mol. Divers.*, vol. 26, no. 3, pp. 1893–1913, Jun. 2022, doi: 10.1007/s11030-021-10326-z.

[3] N. J. Abbott, A. A. K. Patabendige, D. E. M. Dolman, S. R. Yusof, and D. J. Begley, “Structure and function of the blood–brain barrier,” *Neurobiol. Dis.*, vol. 37, no. 1, pp. 13–25, Jan. 2010, doi: 10.1016/j.nbd.2009.07.030.

[4] D. E. Clark, “In silico prediction of blood–brain barrier permeation,” *Drug Discov. Today*, vol. 8, no. 20, pp. 927–933, Oct. 2003, doi: 10.1016/S1359-6446(03)02827-7.

[5] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

[6] V. Jakkula, “Tutorial on Support Vector Machine (SVM)”.

[7] T. Chen *et al.*, “xgboost: Extreme Gradient Boosting.” p. 1.7.8.1, Sep. 01, 2014. doi: 10.32614/CRAN.package.xgboost.

[8] L. Pattanaik and C. W. Coley, “Molecular Representation: Going Long on Fingerprints,” *Chem*, vol. 6, no. 6, pp. 1204–1207, Jun. 2020, doi: 10.1016/j.chempr.2020.05.002.