**基于深度学习的蛋白质残基接触位点预测系统设计与实现**

**摘 要**

伴随着人类基因组计划的顺利实施，每天都会产生大量的基因和蛋白序列等数据。这些生物数据量随着社会的发展爆炸式的增长，最终产生了数以万计的生物分子序列数据。蛋白质残基与配体蛋白结合是完成特定功能的关键，因此对蛋白质残基接触位点进行研究不仅有助于揭示生命活动的本质,而且对疾病发生机制的了解及有效药物的开发均起到推动性的作用。而生物信息和计算机科学的紧密融合为深入了解蛋白质相互作用机理提供了基础的研究条件。本文尝试设计了多特征融合方法，将氨基酸的离散编码特征、位置特异性矩阵特征以及蛋白质二级结构特征组合在一起；利用采样方法对数据进行预处理，改善数据不平衡的程度；分别构建密集连接网络模型和支持向量机模型对预处理后的样本数据进行训练预测，并对模型的性能进行评估。密集连接网络模型和支持向量机模型对金属离子锌结合蛋白质相互作用位点预测的准确度分别达到了90.07%和92.03%；通过将密集连接网络模型和支持向量机模型进行模型融合，密集连接网络模型负责特征提取任务，支持向量机模型负责分类任务，将金属离子锌结合蛋白质相互作用位点预测的准确度提升至92.18%。

**关键词** 蛋白质残基；金属离子锌；接触位点；预测；卷积神经网络；支持向量机

**Design and implementation of prediction system of protein residue contact site based on deep learning**

**Abstract**

Thousands of gene and protein sequence data are generated every moment with the successful implementation of the Human Genome Project, The amount of these biological data has exploded with the development of society, and finally generated tens of thousands of biomolecule sequence data. The interaction between proteins and ligand substances is the key to complete specific functions. Therefore, the study of protein interactions not only helps to reveal the essence of life activities, but also promotes the understanding of disease occurrence mechanism and the development of effective drugs. The basis for the reasoning condition understanding the role of protein interactions is closely linked to the biological information and computer technology. In this course, we tried to design a multi-feature fusion method, combining discrete coding features of amino acids, position-specific matrix features, and protein secondary structure features; Sampling method is used to preprocess the data to improve the degree of data imbalance and we build a densely connected network model and a support vector machine model to train and predict the preprocessed sample data, and evaluate the performance of the model. Densenet model and SVM model protein interaction sites on the metal ion zinc binding prediction accuracy are up to 90.07% and 92.03%, respectively; Up to 92.18% by the fusion of Densenet network model and model SVM model, Densenet model charging of the feature extraction and SVM model charging of the classification, the zinc metal ion binding site of protein interaction site prediction accuracy .

**Keywords** protein residue; metal ion zinc; interaction site; prediction; convolutional neural network; support vector machine

目 录

[第1章 绪论 1](#_Toc41506164)

[1.1 研究背景 1](#_Toc41506165)

[1.2 研究目的及意义 2](#_Toc41506166)

[1.3 国内外研究现状 3](#_Toc41506167)

[1.3.1 蛋白质-蛋白质相互作用位点预测研究现状 3](#_Toc41506168)

[1.3.2 锌离子结合蛋白质作用位点预测研究现状 4](#_Toc41506169)

[1.4 论文主要研究工作 5](#_Toc41506170)

[第2章 蛋白质功能位点预测相关知识 6](#_Toc41506171)

[2.1 蛋白质 6](#_Toc41506172)

[2.1.1 蛋白质的定义 6](#_Toc41506173)

[2.1.2 蛋白质的结构 6](#_Toc41506174)

[2.1.3 蛋白质功能位点 7](#_Toc41506175)

[2.2 金属离子锌结合蛋白质功能位点预测流程 8](#_Toc41506176)

[2.3 主要预测方法 9](#_Toc41506177)

[2.3.1 人工神经网络 9](#_Toc41506178)

[2.3.2 卷积神经网络 11](#_Toc41506179)

[2.3.3 密集连接网络 11](#_Toc41506180)

[2.3.4 支持向量机 12](#_Toc41506181)

[2.3.5 深度学习中的关键技术 12](#_Toc41506182)

[2.4 预测算法的评价指标 14](#_Toc41506183)

[第3章 数据集的构造 15](#_Toc41506184)

[3.1 数据集的选取 15](#_Toc41506185)

[3.2 特征的选择 15](#_Toc41506186)

[3.2.1 氨基酸的离散型编码 15](#_Toc41506187)

[3.2.2 蛋白质位置特异性得分矩阵 16](#_Toc41506188)

[3.2.3 蛋白质二级结构 17](#_Toc41506189)

[3.3 数据平衡化处理 17](#_Toc41506190)

[3.3.1 欠采样 18](#_Toc41506191)

[3.3.2 过采样 18](#_Toc41506192)

[第4章 基于密集连接网络的金属离子锌结合蛋白质作用位点预测方法 19](#_Toc41506193)

[4.1 数据集预处理 19](#_Toc41506194)

[4.2 网络模型结构 20](#_Toc41506195)

[4.3 性能评估 21](#_Toc41506196)

[4.4 本章小结 23](#_Toc41506197)

[第5章 基于支持向量机的金属离子锌结合蛋白质作用位点预测方法 24](#_Toc41506198)

[5.1 数据预处理 24](#_Toc41506199)

[5.2 SVM模型参数 24](#_Toc41506200)

[5.3 性能评估 24](#_Toc41506201)

[5.4 本章小结 25](#_Toc41506202)

[第6章 基于密集连接网络与支持向量机模型融合的金属离子锌结合蛋白质作用位点预测方法 26](#_Toc41506203)

[6.1 数据预处理 26](#_Toc41506204)

[6.2 模型融合 26](#_Toc41506205)

[6.3 性能评估 26](#_Toc41506206)

[6.4 本章小结 28](#_Toc41506207)

[结 论 29](#_Toc41506208)

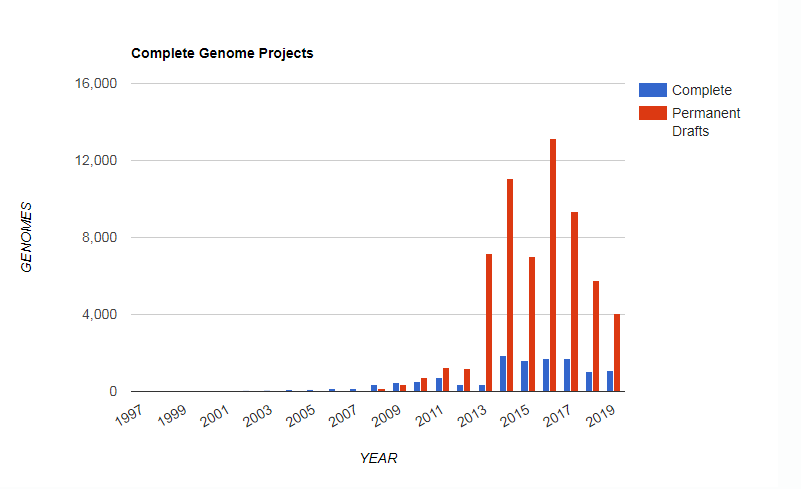
[致 谢 31](#_Toc41506209)

[参 考 文 献 32](#_Toc41506210)

# 绪论

## 研究背景

伴随着人类基因组计划的顺利实施，每天都会产生大量的基因和蛋白序列等数据。这些生物数据量随着社会的发展爆炸式的增长，最终产生了数以万计的生物分子序列数据。从GOLD数据库（网址：<https://gold.jgi.doe.gov/statistics> 需要用Google Chrome浏览器打开）查询可知，截止 2019 年 12 月，已有近2万个基因组完成测序，图 1-1 显示了 1997年到 2019年被完整测序的基因组数目。



**图1-1 1997 年到 2019 年被完整测序的基因组数目**

蛋白质生命的最重要的基本单元，蛋白质间的相互作用对生物体系的运作至关重要。蛋白质相互作用是生命活动的基础，对蛋白质相互作用的研究能够让我们从更深层面上认识生命活动本身，同时也能够用于蛋白质功能的预测[13]。蛋白质结构预测不仅可以有效降低可以使用的蛋白质序列数目与蛋白质结构数目之间的鸿沟，从而辅助结构基因组学的研究，也是人们认识和了解蛋白质折叠机理的一种重要途径(Dill and MacCallum, 2012)[14]。

蛋白质与配体蛋白质结相互作用的预测一直是生物信息学重要的研究课题。目前传统蛋白质相互作用的检测方法主要还是借助物理化学实验[17]，一般有传统的小规模生化试验方法和高通量试验方法。传统的生化实验方法可以检测出很少的蛋白质与配体之间的相互作用，高通量实验方法则可以大规模的鉴定多对相互作用对[15]。对于蛋白质与配体间相互作用的研究，无论是小规模的生化试验方法还高通量实验方法，都在这方面的研究中取得了一定的研究成果并且进行了一些应用，但是还是存在一些问题[15]：

1. 化学实验的完成需要一定的环境限制，对实验环境要求比较严格，造成实验结果复现难的问题。
2. 化学实验的基础是大量的生物数据，实验代价相对比较昂贵，实验耗时较长。
3. 现代化的测序工艺进步，蛋白质与配体作用对预测速度和蛋白质序列测序速度不匹配产生了矛盾。

随着生物信息学的发展，国内外越来越多的研究学者在尝试借助模式识别的研究方法来对预测蛋白质和配体的相互作用。目前生物信息学主要基于统计分析和机器学习算法进行工作，机器学习在特征学习方面相较于统计分析方法能够表现出出色的蛋白质特征捕捉。尽管目前已经产生了一系列基于机器学习的作用位点预测算法，但预测精度仍有待提高[16]。目前现有的机器学习方法大多数依赖人为干预的蛋白质一级结构序列特征提取过程，提取到的数据特征并不能很充分的表达蛋白质残基接触位点的表现特征。

## 研究目的及意义

传统的蛋白质与配体蛋白质相互作用的检测手段都不同方面的存在一些缺陷，生物信息学发展到目前已经有不少学者将模式识别的研究思想应用到了蛋白质相互作用位点的研究。目前用来研究蛋白质相互作用位点的传统的机器学习算法主要有人工神经网络、支持向量机[8]、决策树、随机森林以及聚类算法等一系列的方法。近几年来国内外学者开始将目光转向了机器学习中的一个分支——深度学习。在自然语言处理、图像处理和机器翻译等方面深度学习算法取得了巨大的成果，深度学习具有学习能力强、覆盖范围广、适应能力强、可移植性好等特点。而卷积神经网络（CNN）和长短时期神经网络（LSTM）在深度学习领域更是表现力优秀的神经网络模型，它们在图像处理和序列分析场景下常常能取得相当不错的效果。蛋白质一级结构是氨基酸序列，在深度学习的角度属于序列预测问题（seq2seq问题），国内外一大批学者尝试以LSTM为基础网络模型进行改进，将LSTM深度学习神经网络模型成功运用到了蛋白质与配体结合位点预测问题上[20]，让神经网络自己去提取序列特征，并且取得了不错的效果。而卷积神经网络CNN作为图像处理领域的表现优异的网络，也逐渐被研究学者运用到了蛋白质结合位点预测场景。卷积神经网络在可以通过不同维度的卷积核进行不同层次的特征提取，从而丰富蛋白质序列特征，更好的对蛋白质结构和理化特性进行表达[18]。卷积神经网络CNN[19]，不仅具备强大的自主学习能力，而且可以从样本数据中挖掘出隐含特征，简化了人工表征特征的过程。本课题重点研究将卷积神经网络应用于蛋白质与金属离子锌的相互作用位点预测，使用经典的CNN结构构建模型，通过在不同的数据集上的训练准确度来评价模型。

## 国内外研究现状

### 蛋白质-蛋白质相互作用位点预测研究现状

蛋白质与配体蛋白质的相互作用，在绝大多数的生物行为中至关重要。蛋白质本身具备功能条件，但不直接单独发挥作用，要与配体蛋白质进行相互作用才能发挥功能作用。因此，蛋白质复合物而不是单个组分将决定生物系统的行为。蛋白质的相互作用对于细胞功能的几乎所有方面都是至关重要的，包括信号和代谢途径的调控，蛋白质合成，DNA复制和基因翻译以及免疫学识别[1]。

迄今为止，国外学者已经提出了几种预测PPI位点的方法，并且根据所使用的特征可以将它们大致分为三类。第一类方法仅基于序列信息[2]。Ofran与Rost在研究PPI位点过程中使用了相同的数据集，该数据集包括333个复合物中的1,134条链和59,559个接触残基[3]，当时对部分蛋白质数据的相互作用位点预测的精度达到了70%。第二类方法整合了二级结构信息和序列信息[4]。Wang的研究工作[4]基于由69条蛋白质链组成的异源二聚体的非冗余数据集，其灵敏度为66.3％，特异性为49.7％，准确度为0.654，相关系数为0.297。Zhou及其同事[5]构建了615对非同源复合物形成蛋白进行训练的预测因子，并在一组不同的129对非同源复合物形成蛋白上进行了测试。该方法，可以正确预测了11004个界面残基中的70％。第三类方法使用3D结构信息或具有序列信息的集成3D结构进行预测[6]。Aytuna等人[6]提出了一种算法，该算法运行在67个接口的模板数据集和6170个蛋白质结构的顺序非冗余数据集上。预测的62,616种潜在相互作用的大多数已在PDB、Bind、DIP等数据库中得到验证。在Sikic等人的工作中，他们使用了与Ofran和Rost的研究相同的数据集[3]。在这项研究中，基于序列信息的预测成功的获得了84％的准确率和26％的召回率。与结构信息结合后，预测性能将提高到76％的精度和38％的召回率[7]。基于不同的特征，已经提出了几种机器学习方法来预测蛋白质相互作用的位点，例如神经网络[3]，支持向量机（SVM）[8]，贝叶斯网络[9]和隐马尔可夫模型（HMM）[10]。Li等人[11]在一组50种随机选择的蛋白质上测试了训练的最佳SVM模型[11]。预测核心界面残基的敏感性，特异性和MCC分别为60.6％，53.4％和0.243。Bradford 等人[9]使用贝叶斯网络预测180个蛋白质基准数据集上的蛋白质-蛋白质结合位点，成功率为82％。在Li及其同事的研究中[11]，当使用1276条非冗余的异质复合蛋白链作为训练和测试集时，最佳精度，召回率，准确性和MCC分别为0.536、0.595 0.692和0.328。

### 锌离子结合蛋白质作用位点预测研究现状

很大一部分蛋白质需要金属来稳定其结构或发挥其功能。金属离子锌是这些金属中最常见的一种，它与具有多种生物学作用的蛋白质有关。最近的一项研究估计，人类蛋白质组中40％的锌结合蛋白是转录因子，其余的60％主要是参与离子迁移的酶和蛋白质（[Andreini等，2006](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2144590/" \l "B04)）。金属离子锌具有许多化学性质，因此具有多种生物学功能（[Vallee和Auld 1990](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2144590/#B55)）。由于其D壳已充满，因此不会发生氧化或还原反应。这在氧化还原电势可能波动的生物环境中提供了一定程度的稳定性。它具有不同的配位数和几种类型的连接残基的能力可能使蛋白质调节其结合位点的亲和力和功能[12]。在李慧[15]博士的研究中，通过线性回归和贝叶斯算法对金属离子锌结合蛋白质位点预测准确度指标达到了94.5%、MCC指标达到了72.56%。

## 论文主要研究工作

本次课题任务围绕深度学习在蛋白质残基接触位点预测上的应用展开，并实现一个金属离子锌结合蛋白质残基接触位点的在线预测服务，为研究者提供不同算法模型对蛋白质结合位点进行预测。利用卷积神经网络模型对金属离子锌结合蛋白质相互作用位点进行预测。蛋白质结合位点预测问题存在数据非平衡问题，即样本中正负样本数量差距过大，正类数量远远小于负类数量。因此本课题尝试对样本数据进行欠采样与过采样，防止样本数据特征丢失过多和特征错误补偿。

第一章主要介绍了蛋白质结合位点的研究现状以及背景资料。

第二章主要介绍了蛋白质功能位点预测的相关知识，包括蛋白质、蛋白质作用位点的定义，常用的蛋白质作用数据库；本次课题利用的预测算法模型。

第三章主要介绍了本次课题的数据集来源以及数据特征的选取。

第四章主要介绍了如何构建密集连接网络对金属离子锌结合蛋白质相互作用位点预测。

第五章主要介绍了如何构建支持向量机模型对对金属离子锌结合蛋白质相互作用位点预测。

第六章主要介绍了如何将多个模型（本课题是用了两个模型）进行模型融合尝试对金属离子锌结合蛋白质相互作用位点预测。

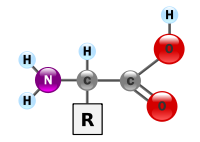
# **蛋白质功能位点预测相关知识**

## 蛋白质

蛋白质是大型的复杂分子，在体内起着许多关键作用。它们在细胞中完成大部分工作，是人体组织和器官的结构，功能和调节所必需的。

### 蛋白质的定义

蛋白质一般情况下是由多个不同的氨基酸按照某种机制组合在一起的高分子化合物[5]。蛋白质形成过程一般是通过氨基酸脱水缩合形成肽链，然后由多个肽链组合成一个大分子氨基酸[5]。下面是氨基酸的结构式如图2-1所示：



**图2-1 氨基酸的结构式**

### 蛋白质的结构

蛋白质的结构又称作蛋白质的构象，每个蛋白质都有其独特的空间构象，正因如此不同的蛋白质才能发挥不同的功能。生物化学家常常用以下三个方面来表示蛋白质的结构：

[一级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B3%AA%E4%B8%80%E7%B4%9A%E7%B5%90%E6%A7%8B)：构成蛋白质肽链的氨基酸序列排列顺序，可以表示肽链的组成元素，但是无法提现具体的肽链折叠结构[13]。

[二级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E4%BA%8C%E7%BA%A7%E7%BB%93%E6%9E%84)：若干个蛋白质肽链的连接折叠方式。蛋白质的二级结构一般分为α螺旋和β折叠这两种结构，主要通过彼此相邻的多肽单元之间的氢键结合在一起形成稳定的结构特征[13]。

[三级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B3%AA%E4%B8%89%E7%B4%9A%E7%B5%90%E6%A7%8B)：[蛋白质](https://www.thoughtco.com/protein-function-373550)多肽链的全面3-D结构。由于蛋白质折叠，在彼此紧密接触的带正电荷和带负电荷的“R”基团之间可能发生离子键合。折叠还可导致半胱氨酸氨基酸的“R”基团之间的共价键合。这种键合形成所谓的二硫键，称为[范德华力的](http://www.chemguide.co.uk/atoms/bonding/vdw.html)相互作用，也有助于稳定蛋白质结构。这些相互作用与极化分子之间产生的吸引力和排斥力有关。这些力有助于分子之间发生键合。

### 蛋白质功能位点

蛋白质的功能往往由其与不同配体相互作用决定的，在相互作用过程中，蛋白质分子中有一类残基对其行使功能起着关键作用，这类残基称为蛋白质的功能位点[15]。蛋白质功能位点包括蛋白质-蛋白质相互作用位点、蛋白质-金属离子相互作用位点等。蛋白质-金属离子相互作用位点的预测对药理研究具有重要意义。比如金属离子锌结合蛋白质作用位点，人体内含量丰富的金属离子锌与各种蛋白质结合有催化、稳定蛋白质作用。本文主要研究金属结合蛋白质作用位点的预测方法。表2-2为蛋白质作用相关数据库。

**表2-2 蛋白质作用相关数据库**

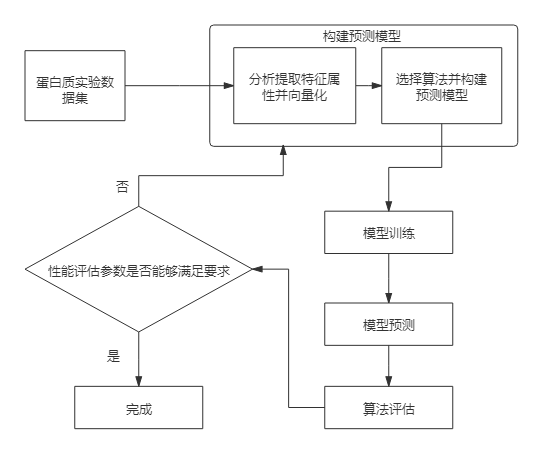
| 数据库 | 物种 | 描述 |
| --- | --- | --- |
| Bind | Many | 生物分子相互作用数据库 |
| BioGrid | Many | 蛋白质-蛋白质以及基因相互作用的数据库 |
| DIP | Many | 相互作用的蛋白质数据库 |
| HPID | Human | 人类蛋白质相互作用数据库 |
| HPRD | Human | 人类蛋白质参考数据库 |
| IntAct | Many | EBI 的蛋白质相互作用数据库 |
| MINT | Many | 分子相互作用数据库 |
| MIPS | Mammals | 哺乳动物蛋白质相互作用数据库 |
| OPHID | Human | 在线预测人类相互作用数据库 |
| STRING | Many | 实验和生物信息学方法预测的蛋白质网络数据库 |
| PDB | Many | 国际上唯一的生物大分子结构数据库，主要包括蛋白质、核酸等实验测定的结构信息 |

## 金属离子锌结合蛋白质功能位点预测流程

本次课题主要利用深度学习算法来预测蛋白质功能位点，主要原理是通过对实验测定的蛋白质功能位点数据进行分析，找到存在的内在规律，提取特征属性，利用深度学习方法建立预测模型，预测蛋白质残基是否是作用的功能位点。

蛋白质功能位点预测识别是一个典型的二类分类问题。金属离子锌结合蛋白质作用位点主要是针对四种氨基酸残基，半胱氨酸（CYS）、组氨酸（HIS)、天冬氦酸（ASP)、谷氨酸（GLU)，是否与蛋白质发生结合作用，则金属结合位点的预测问题转变成了判断这四种残基是否是结合位点和非结合位点。

首先，对选取蛋白质属性特征进行提取，并向量化作为模型的输入然后选择算法构建分类模型，并进行模型训练直至满足精度要求；最后将稳定的模型部署至服务器后端，对前端用户提供预测服务。预测流程如图2-2所示：



**图2-2 蛋白质残基接触位点预测流程**

## 主要预测方法

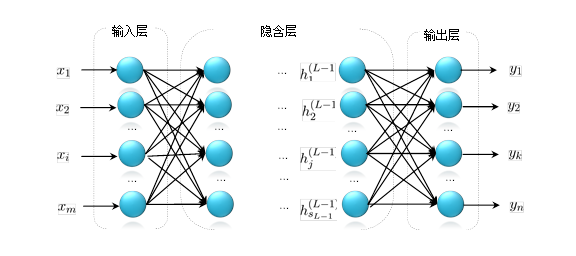
### 人工神经网络

人工神经网络(Artificial Neural Network，ANN)[24]自上世纪80年代以来就是人工智能领域的研究热点，人脑以计算机无法理解的方式来解释现实事物。人工神经网络是计算系统的一部分，旨在模拟人脑分析和处理信息的方式。它是[人工智能](https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp)（AI）的基础，解决了人类或统计标准无法证明或难以解决的问题。人工神经网络具有自学习功能，可以使它们随着更多数据可用而产生更好的结果。

一个多层神经网络具有数以万计个称为处理单元的人工神经元组成，它们通过网络节点相互连接[24]。这些处理单元由输入和输出单元组成。输入单元基于内部加权系统接收各种形式和结构的信息，并且神经网络尝试了解所呈现的信息以生成一个输出报告。就像人类需要规则和指导方针来得出结果或输出一样，人工神经网络也使用一组称为反向传播的学习规则（反向传播错误的缩写）来完善其输出结果。人工神经网络最初会经历训练阶段，在该阶段中，它将学会识别数据的模式，无论是视觉上，听觉上还是文本上。在此监督阶段，网络将其实际产生的输出与预期产生的输出（期望的输出）进行比较。两种结果之间的差异可通过反向传播进行调整。这意味着网络将向后工作，从输出单元到输入单元，以调整其在单元之间的连接权重，直到实际结果与期望结果之间的差异产生最小的可能误差为止。典型的神经网络主要分为两部分：

1. 神经元

神经元模型是神经网络中最小的功能单元。神经元包含输入和输出。一般情况下一个神经网络模型由若干层若干个神经元组成，第一层称为输入层，最后一层被称为输出层，中间所有层级被称为隐含层。典型的人工神经网络如图2-3所示：



**图2-3 典型的人工神经网络结构**

输入向量为：

|  |  |  |
| --- | --- | --- |
|  |  | (2-1) |

输出向量为：

|  |  |  |
| --- | --- | --- |
|  |  | (2-1) |

第隐含层各神经元的输出为：

|  |  |  |
| --- | --- | --- |
|  |  | (2-3) |

其中，为第层神经元的个数。

设 为从层第个神经元与层第 个神经元之间的连接权重；为第层第个神经元的偏置，那么：

|  |  |  |
| --- | --- | --- |
|  |  | (2-4) |
|  |  | (2-5) |

其中，为层第个神经元的输入，为神经元的激活函数。通常在多层神经网络中采用非线性激活函数。

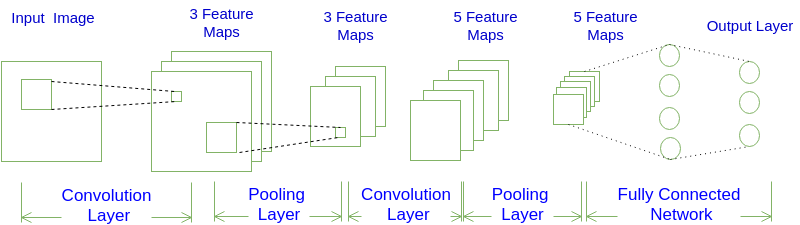
1. 激活函数

BP神经网络通常使用下面两种非线性激活函数：

|  |  |  |
| --- | --- | --- |
|  |  | (2-6) |
|  |  | (2-7) |

### 卷积神经网络

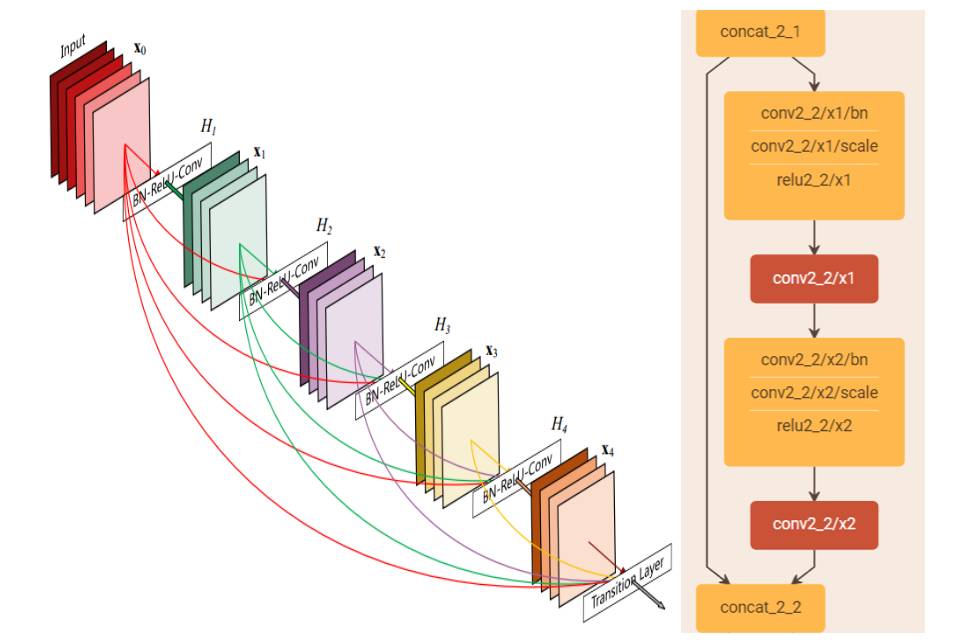
卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，被众多学者应用到图像识别场景中。卷积神经网络可以通过不同的卷积核参数可以对图像提取不同尺度的模式特征，这些模式特征与手动提取的模式特征相比更抽象有效。卷积神经网络包含若干个卷积层，卷积层对输入图像进行特征提取，过使用输入数据的小方块学习图像特征来保留像素之间的关系，使用[线性滤波器来](https://www.sciencedirect.com/topics/engineering/linear-filter)计算其输入的卷积。以与神经网络中的权重相同的方式学习每个卷积核的系数。通常，在卷积之后，将非线性（激活）函数应用于卷积结果的每个元素。因此，卷积层可以视为标准神经网络中神经元共享权重的层。卷积层的输入可以是输入图像，也可以是前一层的输出，即另一个特征图。卷积层中特定类型的权重分配不仅降低了自由度，而且还确保了平移的等方性，这意味着在忽略边界效应的情况下，如果我们对输入图像进行卷积计算，则以相同的方式对生成的特征图再次进行卷及计算。此外，附加到输入的权重对于特征图中的每个单位都是相同的（并且在各个特征图中是不同的）。典型的卷积神经网络结构图如图2-4所示：



**图2-4 典型的卷积神经网络结构**

### 密集连接网络

密集连接网络（DenseNet）是ResNet网络的连接密度的升级版、计算复杂度的优化版，当然密集连接网络也是卷积神经网络，它的特点是在神经网络的前向传递中，每一层都和前面的所有层直接连接，每层的输入来自于之前所有层的输出。一个普通的有层的神经网络会产生个层与层之间的连接，而同样是层的密集连接网络则会因为它层与层相互连接的特点，产生多达个连接。正因为如此特殊的网络结构，密集连接网络多层特征复用，模型计算量极大地减少。密集连接网络主干网络结构示意图如图2-5所示：



**图 2-5 密集连接网络主干网络结构**

### 支持向量机

支持向量机（Support Vector Machines, SVM）在无穷维空间中创建一个超平面，即分类和回归。分离是通过超平面实现的对于所有类别中最近的训练数据点而言，具有最大的分类。这个最近的数据点称为功能裕量。支持向量机分类器的泛化误差取决于功能边界的大小。支持向量机训练算法在功能裕度的基础上建立模型。功能裕度的类别构成了非概率二进制线性分类器。该技术是用于线性和非线性分类的监督学习模型。使用基于内核的函数将输入映射到高维特征空间中，从而执行非线性分类。

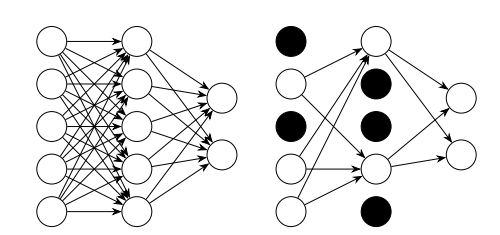
### 深度学习中的关键技术

1. 激活函数（Activation Function）

神经网络激活功能是深度学习的关键组成部分。激活函数确定深度学习模型的输出，其准确性以及训练模型的计算效率，这可以建立或破坏大规模的神经网络。激活函数还会对神经网络的收敛能力和收敛速度产生重大影响，或者在某些情况下，激活函数可能会首先阻止神经网络收敛。

1. Dropout机制

[Dropout](https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg_edit&action=edit&type=10&appmsgid=503278593&isMul=1&token=805468140&lang=zh_CN)机制是一种避免前馈神经网络中出现的过拟合现象的简单技术。神经网络可能会以很少的训练样本快速适合训练数据集。在训练过程中，按比例随机丢弃某些神经单元节点，使得训练的模型对未知蛋白质结合位点数据更具有泛化性，且能够有效地正则化网络数据。该方法示意图如图 2-6 所示：



**图2-6 标准 Dropout 的示例**

1. 损失函数（Loss Function）

损失函数是用来计算预测结果矩阵和真实标签矩阵之间距离差的尺度函数，这个尺度函数的结果导向是最优化，尺度函数一般为最小化目标函数或者是最大化目标函数。

1. 卷积块注意力模块（Convolutional Block Attention Module）

卷积块注意力模块（CBAM），这是一种用于前馈卷积神经网络的简单而有效的注意力模块[23]。给定一个中间特征图，我们的模块会沿着两个独立的维度（通道和空间）依次推断注意力图，然后将注意力图与输入特征图相乘以进行自适应特征细化[23]。由于CBAM是轻量级的通用模块，因此可以将其无缝集成到任何CNN体系结构中，而开销却可以忽略不计，并且可以与基础CNN一起进行端到端训练[23]。

## 预测算法的评价指标

衡量一个分类器预测性能的好坏，选择合适的验证方法和评价指标是非常关键的。下面介绍常用的评估验证方法。为了更好地、更公正地评价预测方法的性能，采用下面的评价指标参数对其进行评价。

1. 召回率
2. 精确度
3. 准确度
4. F1得分

其中真实值与预测值的组合，包括True Positive、True Negative、False Positive和False Negative具体含义见表2-8:。

**表2-8 真实值与预测值的组合表示含义**

|  |  |  |
| --- | --- | --- |
| 真实值与预测值的组合 | 真实值 | 预测值 |
| True Positive | TRUE | TRUE |
| True Negative | TRUE | FALSE |
| False Positive | FALSE | TRUE |
| False Negative | FALSE | FALSE |

# 数据集的构造

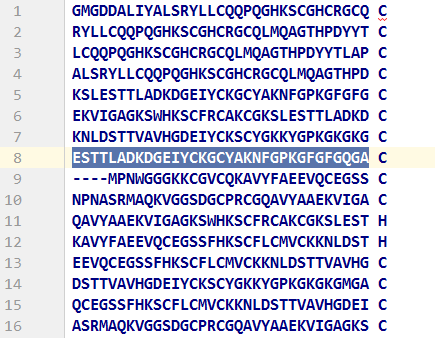
## 数据集的选取

选择 Zhao 等人[22]研究中的数据集（Zhao\_dataset）作为本次课题的训练集，Passerini 等人[21]用的数据集作为测试集。在 Zhao 数据集中去掉与 Passerini 重复的蛋白质序列，最后得到 391条蛋白质链。根据上文锌结合位点的定义，在 Zhao 数据集上得到锌结合位点 1958个，非锌结合位点 13976 个。

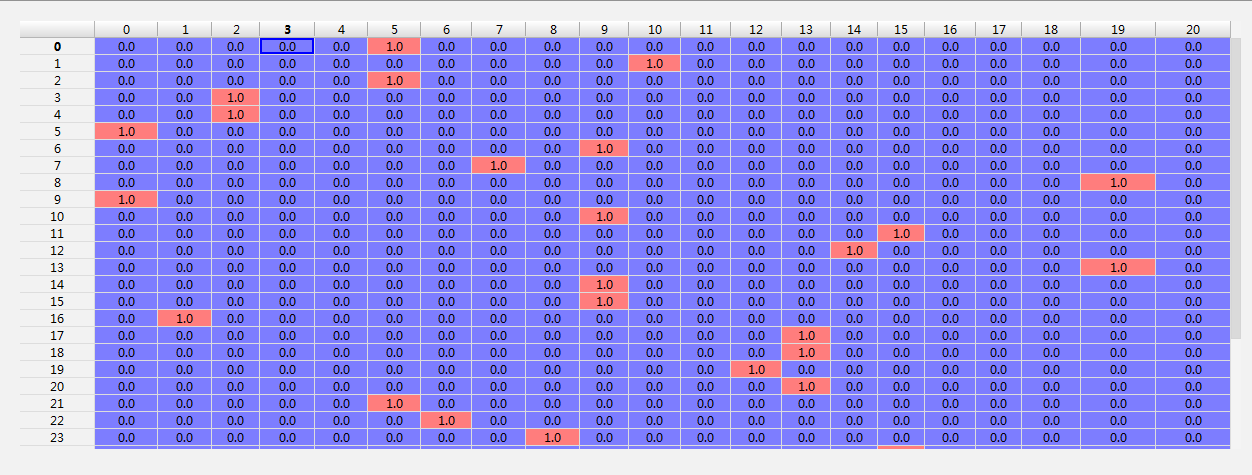
## 特征的选择

### 氨基酸的离散型编码

本课题实验对蛋白质序列作为数据输入到模型进行计算，那么就需要对原始的蛋白质序列数据进行数据特征编码。实验将蛋白质序列分割长度为L（L= 15，16,17……35）氨基酸片段。常见的氨基酸有20种，将其他非常见的氨基酸划分为X类氨基酸，那么在本次实验中氨基酸的种类为21种。我们对氨基酸片段进行特征编码时，每个氨基酸的特征编码为维的张量，张量的每个氨基酸所属编号维度数据位1，其余数据为0；若编码的氨基酸为非常见类型氨基酸，特征张量的每一维度均为0.5。最终一个蛋白质序列被转化成的二维张量。其具体的蛋白质序列编码结果如图3-1、图3-2所示。



**图3-1 长度为 33 的蛋白质片段**



**图3-2 长度为 33 的蛋白质片段离散型编码部分数据**

### 蛋白质位置特异性得分矩阵

位置特异性打分矩阵（PSSM）表示多序列比对中固有的模式一组同源序列。使用配置文件的基本思想是将数据库中的查询序列与比对表中的序列进行匹配，从而使保守位置的权重高于可变位置的权重。这些图谱是通过在比对的每个位置处每个氨基酸（或缺口）的一组概率分数获得的。位置特异性评分矩阵的进化保守性基本上表明了每个残基在序列的特定位置保守的可能性。但是，它不仅涵盖了简单的保留方式，而且还考虑了特定的替换方式，从而为表征每个位置提供了线索。在假定大多数结合位点在某种程度上是保守的情况下使用这些功能。特定的DNA结合残基的取代谱应反映特定趋势，以免失去结合能力，并且可以通过机器学习模型（例如神经网络）捕获此模式。因此，位置特异性打分矩阵可能是用于预测结合位点的序列信息的最流行表示形式。位置特异性打分矩阵可以通过PSI-BLAST在 Swiss-prot 数据库中迭代多次获得数据，长度为L的蛋白质序列可以计算得到维特征。

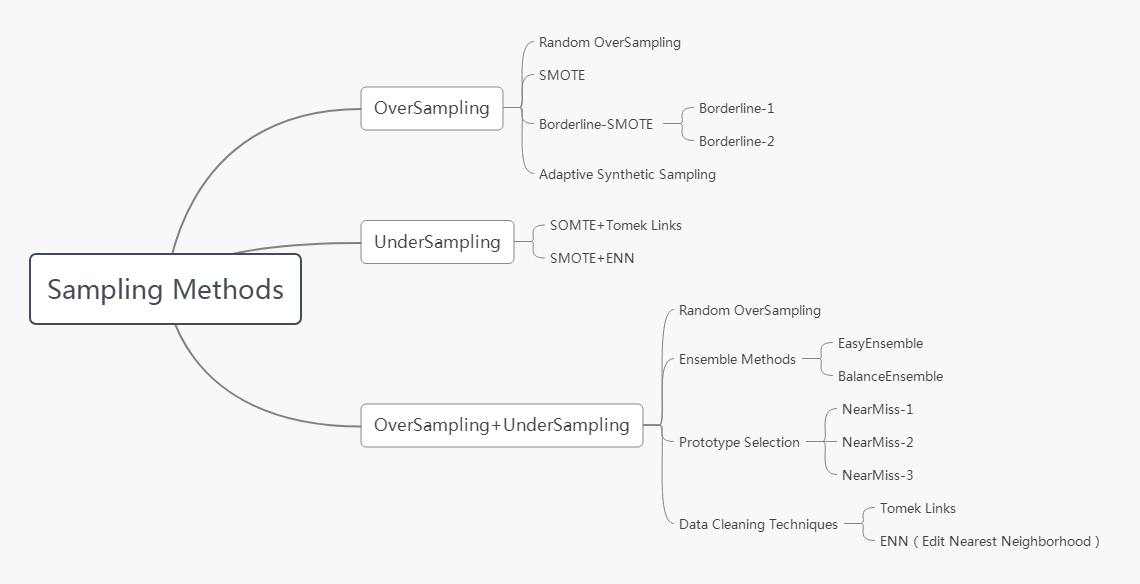
### 蛋白质二级结构

合理对蛋白质二级结构数据进行数据特征编码，有助于提升对蛋白质作用位点预测的准确度[25]。蛋白质的二级结构包括：α-螺旋，β-转角以及β-折叠和无规卷曲。蛋白质二级结构中，α-螺旋，β-折叠，分别用H和E来描述，其余的无规则卷曲部分用C来描述[26]。通过使用在线免费的蛋白质计算服务PSIPRED来计算蛋白质二级结构数据，原则上PSIPRED在线服务有上传和并行计算的限制，但是通过简单的爬虫程序可以批量计算蛋白质氨基酸的二级结构数据，长度为L的蛋白质序列利用PSIPRED计算可获得维特征。

## 数据平衡化处理

金属锌离子与蛋白质结合位点预测是一个二分类问题，实际上真正结合的作用位点相对于非结合位点数量非常少。锌离子结合蛋白质作用位点预测是典型的非平衡分类问题。经过清洗处理后的数据集非结合位点数量13976个，结合位点数量1958个。对数据进行合理的采样处理能够避免传统机器学习方法对非平衡数据集分类的偏向性，更好地提高非平衡数据分类的预测准确性。

目前采样方法（SamplingMethods）分为欠采样和过采样。具体的采样方法分类如图3-3所示：



**图3-3 采样方法分类**

### 欠采样

欠采样是从样本数量较多的类别中按照某种采样规则剔除某些样本。本次课题实验中采用的是随机欠采样来减少样本数量较多的类别（非结合位点样本）中的样本。如果剔除样本的数量过大，那么合成的训练集丢失了某些数据信息，最终会使得模型只学习到总体模型的一部分特征。

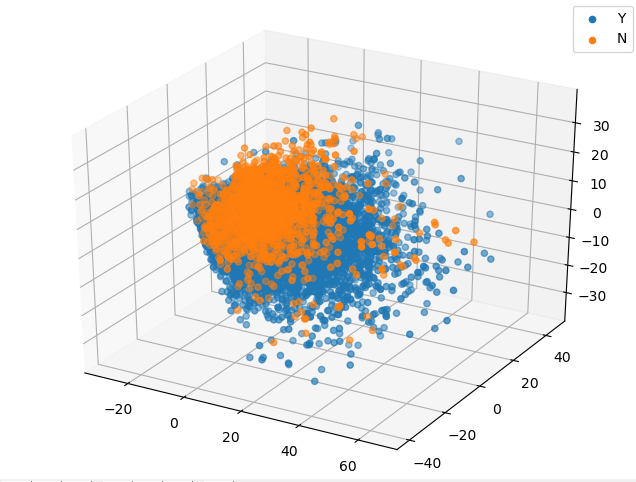
### 过采样

过采样按照某种规则对少数类的样本进行差值来增加新的样本，再合并原有多数类样本作为新的训练数据集。本次课题实验中采用的自适应综合过采样（ADASYN）来增加少数类的样本数量。自适应综合过采样在于它为“难以学习”的实例创建更多数据的自适应性，并允许您对机器学习模型进行更多负样本采样，最终可以综合平衡数据。

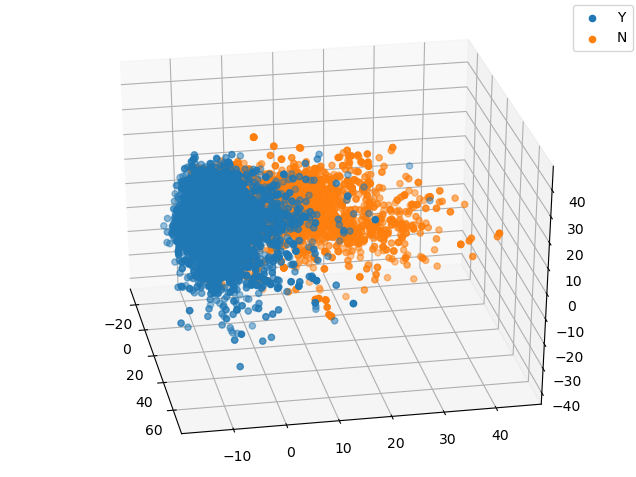
# **基于密集连接网络的金属离子锌结合蛋白质作用位点预测方法**

## 数据集预处理

目前经过清洗处理后的锌离子与蛋白质作用数据集非结合位点数量13976个，结合位点数量1958个，数据集存在严重的正负样本数据分布不平衡问题。正负样本分布如图4-1所示。实验使用随机欠采样算法（RandomUnderSampling）对正样本即非结合位点样本进行随机删除50%的样本数据，负样本数据不做任何处理，将正负样本的数量调整到6988：1958；然后使用自适应综合过采样算法（ADASYN）对负样本即结合位点样本进行自适应插值扩充，正样本数据不做任何处理，将正负样本的数量调整到6988: 6871。正负样本分布如图4-2所示：



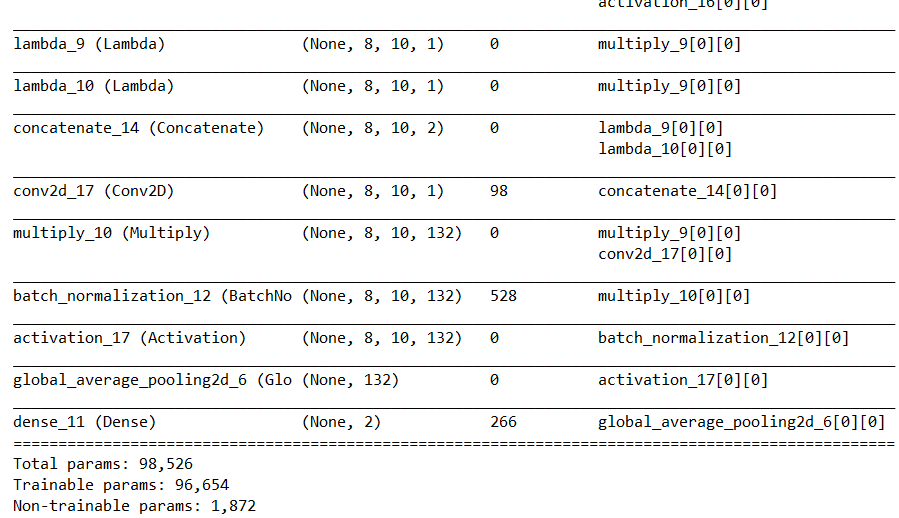
**图4-1 预处理前正负样本数据分布**



**图4-2 预处理后正负样本数据分布**

## 网络模型结构

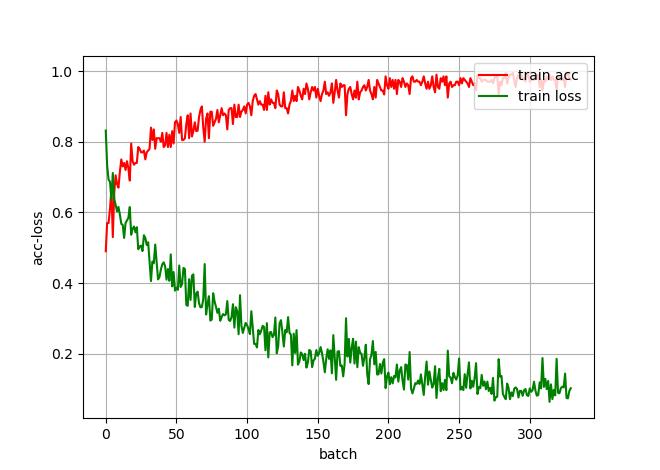
本次课题实验选用的是密集连接网络（Densenet），密集连接网络的网络深度Depth是要符合公式： 这是由密集连接网络源代码规定的计算公式。由于蛋白质结合位点数据特征为维的张量，L为蛋白质序列片段长度，L取值在10 ~ 40。这里可以将每一个蛋白质结合位点数据特征理解为一张的图像，很显然结合位点数据特征尺度比较小，根据经验本次实验分别选用了Depth= 7、Depth= 10、Depth=13、Depth=22、Depth=31等浅深度的Densenet网络模型添加注意力机制进行测试，经过测试Depth=13的浅深度密集连接网络效果最好。网络模型结构如图4-3所示：



**图4-3 Densenet网络模型部分结构图**

## 性能评估

本次课题实验尝试了不同的网络深度，最终经测试在深度为13的密集连接网络模型下，将蛋白质数据集按照0.95:0.05分割为训练集和测试集，batch size为200，epoch为5的情况下，损失和准确度如图4-4所示：



**图4-4 训练集的损失和准确度**

训练集和验证集的准确率如表 4-5 所示：

**表4-5 训练集和验证集的准确率**

|  |  |  |
| --- | --- | --- |
| **Train\_loss** | **Train\_acc** | **Val\_acc** |
| **0.225** | **0.9528** | **0.9007** |

从图4-4来看，无论是训练集的loss还是acc都在波动下降和波动上升，这是一个数据分布平衡的表现。但是在val\_acc和train\_acc差距还是比较大，可能的原因还是数据平衡处理章节里面提到的，在利用随机欠采样算法对非结合位点样本剔除样本、利用自适应综合过采样算法对结合位点样本插值，产生了数据特征丢失或者引入了无关特征，这都会导致训练集和测试集之间的样本分布有所差异，最终体现在模型的分类效果上。

对于数据不平衡的分类问题，单从val\_acc这一个指标来观察并不能合理的评估模型的好坏，我们来观察下该模型分别在训练集和测试集上的混淆矩阵，混淆矩阵如表4-6、表4-7所示：

**表4-6 Zhao\_Dataset训练集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.99 | 0.96 | 0.97 | 13976 |
| 1 | 0.75 | 0.92 | 0.83 | 1958 |
| Accuracy |  |  | 0.95 | 15934 |
| Macro avg | 0.87 | 0.94 | 0.90 | 15934 |
| Weighted avg | 0.96 | 0.95 | 0.95 | 15934 |

**表4-7 Passerni\_dataset验证集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.96 | 0.93 | 0.94 | 7133 |
| 1 | 0.57 | 0.72 | 0.64 | 986 |
| Accuracy |  |  | 0.90 | 8119 |
| Macro avg | 0.77 | 0.82 | 0.79 | 8119 |
| Weighted avg | 0.91 | 0.90 | 0.91 | 8119 |

我们对比训练集和测试集的混淆矩阵表4-6、表4-7，可以更清晰的观察到训练模型在验证集上的表现与在训练集上的表现是有所差距的，特别体现在负样本的精度上。这个差距是由于对负样本数据进行插值扩容时引入了不相关的负样本数据，导致原负样本的数据分布和平衡化后负样本的数据分布有了一定的偏差，最终由数据传导至网络模型的结果。

## 本章小结

本章首先对数据集进行了预处理操作，包括数据向量化、数据平衡化处理等，然后构建不同深度的Densenet网络模型喂入已经预处理之后的Zhao\_Dataset蛋白质结合位点数据训练网络。利用Passerni\_dataset验证集对训练好的网络模型加载进行测试，经测试得到了90.0%的准确度。

# 基于支持向量机的金属离子锌结合蛋白质作用位点预测方法

## 数据预处理

目前经过清洗处理后的锌离子与蛋白质作用数据集非结合位点数量13976个，结合位点数量1958个，数据集存在严重的正负样本数据分布不平衡问题。实验使用随机欠采样算法（RandomUnderSampling）对正样本即非结合位点样本进行随机删除50%的样本数据，负样本数据不做任何处理，将正负样本的数量调整到6988：1958；然后使用自适应综合过采样算法（ADASYN）对负样本即结合位点样本进行自适应插值扩充，正样本数据不做任何处理，将正负样本的数量调整到6988: 6871。

## SVM模型参数

SVM模型参数众多，本章节主要尝试不同核函数（Kernal）和惩罚系数（C）对蛋白质结合位点预测结果的影响。SVM模型一般使用线性核（Linear）和径向基函数核（RBF）经过测试Kernal函数选择径向基函数（rbf），惩罚系数为1.2时，在本实验的训练集上训练效果最好。

## 性能评估

当SVM模型核函数选择径向基函数核、惩罚系数为1.2时，训练集的分类效果最好，Zhao\_Dataset训练集的准确度和Passerini\_Dataset测试集的准确度如图5-1所示：

**表5-1 Zhao\_Dataset和Passerini\_Dataset的准确率**

|  |  |
| --- | --- |
| Train\_acc | Val\_acc |
| 0.9650 | 0.9203 |

从训练准确度和验证集准确度来看SVM模型的效果要优于密集连接网络。

**表5-2 Zhao\_Dataset训练集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.99 | 0.97 | 0.98 | 13976 |
| 1 | 0.82 | 0.91 | 0.86 | 1958 |
| Accuracy |  |  | 0.97 | 15934 |
| Macro avg | 0.91 | 0.94 | 0.92 | 15934 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 15934 |

**表5-3 Passerni\_dataset验证集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.95 | 0.96 | 0.95 | 7133 |
| 1 | 0.69 | 0.62 | 0.65 | 986 |
| Accuracy |  |  | 0.92 | 8119 |
| Macro avg | 0.82 | 0.79 | 0.80 | 8119 |
| Weighted avg | 0.92 | 0.92 | 0.92 | 8119 |

表 5-2 和表 5-3 分别是利用SVM分类器计算出来的训练集混淆矩阵和验证集混淆矩阵，从准确度accuracy评价指标来看，SVM模型在蛋白质结合位点预测实验中表现更好，准确度达到了92%。从召回率指标来看SVM模型在蛋白质结合位点预测实验中比密集连接网络（Densenet）表现得稍差一点。从总体来看，SVM模型分类效果优于Densenet网络模型。

## 本章小结

本章首先对Zhao\_Dataset蛋白质结合位点数据集进行了预处理操作，包括数据向量化、数据平衡化处理等，然后构建SVM预测模型，分别对不同的SVM模型的核函数、惩罚系数C进行训练测试。利用Passerni\_dataset验证集对训练好的SVM模型进行测试，得到了92.0%的准确度。

# 基于密集连接网络与支持向量机模型融合的金属离子锌结合蛋白质作用位点预测方法

## 数据预处理

目前经过清洗处理后的锌离子与蛋白质作用数据集非结合位点数量13976个，结合位点数量1958个，数据集存在严重的正负样本数据分布不平衡问题。实验使用随机欠采样算法（RandomUnderSampling）对正样本即非结合位点样本进行随机删除50%的样本数据，负样本数据不做任何处理，将正负样本的数量调整到6988：1958；然后使用自适应综合过采样算法（ADASYN）对负样本即结合位点样本进行自适应插值扩充，正样本数据不做任何处理，将正负样本的数量调整到6988: 6871。

## 模型融合

我们知道密集连接网络在数据特征提取方面遗传了卷积神经网络的天生优势，而在上一章节中支持向量机模型就表现了在二分类场景下优秀的分类效果，所以本章尝试将密集连接网络模型和支持向量机模型进行模型融合，使用密集连接网络进行蛋白质特征提取，然后将密集连接网络提取到的特征作为支持向量机模型训练的数据，即密集连接网络负责特征采集任务，支持向量机负责分类任务。本次实验将密集连接网络的倒数第二层网络层作为网络输出层，输出shape为（None，132），并将这132维特征作为输入数据喂入到SVM网络模型进行分类训练。这里我们需要思考在何种场景下密集连接网络的数据特征提取效果与何种参数场景下的支持向量机模型进行组合效果最佳。我们目前简单地将两种模型组合起来，最终得到一对最优的组合。

## 性能评估

在深度为13的密集连接网络模型下，将蛋白质数据集按照0.95:0.05分割为训练集和测试集，batch size为200，epoch为30；SVM模型的核函数依然选择径向基函数，惩罚系数C为1.2。Zhao\_Dataset训练集的准确度和Passerini\_Dataset测试集的准确度如图6-1所示：

**表6-1 Zhao\_Dataset和Passerini\_Dataset的准确率**

|  |  |
| --- | --- |
| Train\_acc | Val\_acc |
| 0.9965 | 0. 9218 |

验证集的准确度与前面两章的单一模型分类效果更好，但是训练集准确度和验证集准确度有比较大的差距。

**表6-2 Zhao\_Dataset训练集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 1.00 | 1.00 | 1.00 | 13976 |
| 1 | 1.00 | 1.00 | 1.00 | 1958 |
| Accuracy |  |  | 1.00 | 15934 |
| Macro avg | 1.00 | 1.00 | 1.00 | 15934 |
| Weighted avg | 1.00 | 1.00 | 1.00 | 15934 |

**表6-3 Passerni\_dataset验证集的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.94 | 0.98 | 0.96 | 7133 |
| 1 | 0.76 | 0.52 | 0.62 | 986 |
| Accuracy |  |  | 0.92 | 8119 |
| Macro avg | 0.85 | 0.75 | 0.79 | 8119 |
| Weighted avg | 0.91 | 0.92 | 0.92 | 8119 |

表 6-2和表 6-3分别是利用密集连接网络与支持向量机模型融合计算出来的训练集混淆矩阵和验证集混淆矩阵。通过模型融合后在训练集上各个评估指标都几乎接近100%，但是在验证集上，各个评估指标都与训练集上的结果有比较大的差距，这说明密集连接网络在Zhao\_Dataset训练集上出现了过拟合的问题，导致训练集和验证集的效果有比较大的出入。

与前面两章节的验证集结果对比，模型融合后的准确度accuracy要优于单一模型的准确率，达到了92.18%。

## 本章小结

本章节通过对前面两章节的分类模型进行组合，密集连接网络负责特征提取任务，支持向量机负责分类任务。通过实验结果观察，模型的融合确实能够给预测准确度带来提升，但是也存在特征提取模型的过拟合问题。

# 结 论

人体内含量丰富的金属离子锌与各种蛋白质结合有催化、稳定蛋白质作用，对金属离子锌结合蛋白质作用位点预测具有重要的现实意义。本课题通过对原始蛋白质数据进行数据向量化操作，将蛋白质序列信息转化为向量信息。为了使得特征数据对蛋白质序列信息更具有表达性，本课题设计了多特征融合方法，将氨基酸的离散编码特征、位置特异性矩阵特征以及蛋白质二级结构特征组合在一起，尝试用丰富的特征来描述金属离子锌结合蛋白质相互作用的奥秘；蛋白质相互作用位点预测由于正负样本数量差距过于悬殊，数据样本分布不平衡。经欠采样和过采样算法处理过后的蛋白质结合位点数据，较处理前相比，蛋白质位点数据分布达到平衡；分别构建密集连接网络模型和支持向量机模型对预处理后的样本数据进行训练预测，并对模型的性能进行评估。密集连接网络模型和支持向量机模型对金属离子锌结合蛋白质相互作用位点预测的准确度分别达到了90.07%和92.03%；通过将密集连接网络模型和支持向量机模型进行模型融合，密集连接网络模型负责特征提取任务，支持向量机模型负责分类任务，将金属离子锌结合蛋白质相互作用位点预测的准确度提升至92.18%。通过以上一系列的实验结果对比，得出：

1. 多特征融合往往比单一特征对任务本身就有更丰富的表达性；多模型融合往往比单一模型对任务预测有更好的效果；
2. 蛋白质原始数据是氨基酸的排列顺序，这是计算机无法识别运算的字符数据，需要利用特征工程的分析方法对这类无法直接参与计算机运算的数据选择合适的数据特征进行数值编码。
3. 在已有数据分布不平衡的情况下，尽可能的对数据进行采样处理，人为迫使数据回到分布平衡的轨道上，尽管这样会让原始数据丢失部分特征，也会让原始数据扩充无关样本，但这样能够很有效的改善预测分类效果；否则将会使少数类样本的预测结果完全偏向多数类结果，导致分类器会忽略样本极少的类别，这是我们极不希望看到的。在二分类问题场景下，我们看到支持向量机模型的分类效果要比密集连接网络优秀。
4. 深度学习算法是对数据极度依赖的，导致数据的质量很大程度上决定了模型的效果。传统的支持向量机算法在单一模型、且数据量不大的场景下，效果要比严重依赖数据质量的深度学习算法要好得多。

# **致 谢**

大学四年，匆匆而过。人生能有多少个四年呢？人生又能有多少个四年过得如此的充实呢？

大学光阴，转瞬即逝。但这段日子里陪伴过我的老师和同学是我终生难忘的。在这里我要感谢每一位任课老师，没有你们专业授课，就不会构建我大脑里完整的计算机体系。尤其感谢的是我的指导老师陆卫忠老师。陆老师在我的论文撰写、课题实验方面给了我相当大的帮助。陆老师传授与我的学习方法和思维对我今后的工作和生活来说是一笔宝贵的财富。在此，我向我的导师陆卫忠老师给予我的关怀和教导表示衷心地感谢。

感谢大学期间全体老师的教导，感谢我的母校为我带来精彩的四年大学生活。

# **参 考 文 献**

1. Alberts B (1989) Molecular biology of the cell: Garland Pub.
2. Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) A fast method to predict protein interaction sites from sequences. J Mol Biol 302: 917–926. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/10993732)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Mol+Biol&title=A+fast+method+to+predict+protein+interaction+sites+from+sequences&author=X+Gallet&author=B+Charloteaux&author=A+Thomas&author=R+Brasseur&volume=302&publication_year=2000&pages=917-926&pmid=10993732&)]
3. Ofran Y, Rost B (2003) Predicted protein-protein interaction sites from local sequence information. FEBS Lett 544: 236–239. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/12782323)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=FEBS+Lett&title=Predicted+protein-protein+interaction+sites+from+local+sequence+information&author=Y+Ofran&author=B+Rost&volume=544&publication_year=2003&pages=236-239&pmid=12782323&)]
4. Wang B, Chen P, Huang DS, Li JJ, Lok TM, et al. (2006) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 580: 380–384. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/16376878)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=FEBS+Lett&title=Predicting+protein+interaction+sites+from+residue+spatial+sequence+profile+and+evolution+rate&author=B+Wang&author=P+Chen&author=DS+Huang&author=JJ+Li&author=TM+Lok&volume=580&publication_year=2006&pages=380-384&pmid=16376878&)]
5. Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44: 336–343. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/11455607)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Proteins&title=Prediction+of+protein+interaction+sites+from+sequence+profile+and+residue+neighbor+list&author=HX+Zhou&author=Y+Shan&volume=44&publication_year=2001&pages=336-343&pmid=11455607&)]
6. Aytuna AS, Gursoy A, Keskin O (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21: 2850–2855. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/15855251)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Bioinformatics&title=Prediction+of+protein-protein+interactions+by+combining+structure+and+sequence+conservation+in+protein+interfaces&author=AS+Aytuna&author=A+Gursoy&author=O+Keskin&volume=21&publication_year=2005&pages=2850-2855&pmid=15855251&)]
7. Sikic M, Tomic S, Vlahovicek K (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 5: e1000278. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2621338/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19180183)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=PLoS+Comput+Biol&title=Prediction+of+protein-protein+interaction+sites+in+sequences+and+3D+structures+by+random+forests&author=M+Sikic&author=S+Tomic&author=K+Vlahovicek&volume=5&publication_year=2009&pages=e1000278&pmid=19180183&)]
8. Bradford JR, Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 21: 1487–1494. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/15613384)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Bioinformatics&title=Improved+prediction+of+protein-protein+binding+sites+using+a+support+vector+machines+approach&author=JR+Bradford&author=DR+Westhead&volume=21&publication_year=2005&pages=1487-1494&pmid=15613384&)]
9. Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein-protein interfaces using a Bayesian network prediction method. J Mol Biol 362: 365–386. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/16919296)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=J+Mol+Biol&title=Insights+into+protein-protein+interfaces+using+a+Bayesian+network+prediction+method&author=JR+Bradford&author=CJ+Needham&author=AJ+Bulpitt&author=DR+Westhead&volume=362&publication_year=2006&pages=365-386&pmid=16919296&)]
10. Bernardes JS, Fernandez JH, Vasconcelos AT (2008) Structural descriptor database: a new tool for sequence-based functional site prediction. BMC Bioinformatics 9: 492. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612011/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19032768)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Bioinformatics&title=Structural+descriptor+database:+a+new+tool+for+sequence-based+functional+site+prediction&author=JS+Bernardes&author=JH+Fernandez&author=AT+Vasconcelos&volume=9&publication_year=2008&pages=492&pmid=19032768&)]
11. Li N, Sun Z, Jiang F (2008) Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinformatics 9: 553. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2627892/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19102736)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Bioinformatics&title=Prediction+of+protein-protein+binding+site+by+using+core+interface+residue+and+support+vector+machine&author=N+Li&author=Z+Sun&author=F+Jiang&volume=9&publication_year=2008&pages=553&pmid=19102736&)]

1. [Jessica C. Ebert](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ebert%20JC%5BAuthor%5D&cauthor=true&cauthor_uid=18042678) and [Russ B. Altman](https://www.ncbi.nlm.nih.gov/pubmed/?term=Altman%20RB%5BAuthor%5D&cauthor=true&cauthor_uid=18042678) (2008) Robust recognition of zinc binding sites in proteins
2. 王菲露.基于机器学习方法的蛋白质相互作用位点及二级结构预测出[D].安徽大学. 2009.5
3. 熊大鹏.基于深度学习架构的蛋白质远程残基接触预测研究[D].生命科学学院. 2017.5
4. 李慧.蛋白质功能位点预测方法研究[D] .南京航天航空大学. 2018.10
5. 王多林.基于深度学习的蛋白质翻译后修饰位点预测研究[D].吉林大学. 2018.6.5
6. 陈震.基于序列信息的蛋白质功能位点预测的算法开发[D].中国农业大学. 2014.6
7. Haoyang Zeng, Matthew D. Edwards, Ge Liu and David K. Gifford.Convolutional neural network architectures for predicting DNAprotein binding[D].MIT.June.2016.
8. Yifeng Cui, Qiwen Dong, Daocheng Hong & Xikun Wang (2019) Predicting protein-ligand binding residues with deep convolutional neural networks.
9. Shweta Yadav, Asif Ekbal, Sriparna Saha, Ankit Kumar, Pushpak Bhattacharyya (2018) Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein–protein interaction..
10. Passerini A, Punta M, Ceroni A, et al. Identifying cysteines and histidines intransition-metal-binding sites using support vector machines and neural networks [J].Proteins-Structure Function and Bioinformatics, 2006, 65(2): 305-316.
11. Martin T. Hagan, Howard B. Demuth, Mark Beale(1995)Neural network design Zhao W, Xu M, Liang Z, et al. Structure-based de novo prediction of zinc-binding sites in proteins of unknown function [J]. Bioinformatics, 2011, 27(9): 1262-1268.
12. Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon(2018) CBAM: Convolutional Block Attention Module.

1. [M.W Gardnera1](https://www.sciencedirect.com/science/article/abs/pii/S1352231097004470" \l "!), [S.R Dorlinga1](https://www.sciencedirect.com/science/article/abs/pii/S1352231097004470#!)(1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences
2. Yang Jianyi, Roy A, Zhang Yang. Protein - ligand binding site recognition using complementary bindingspecific substructure comparison and sequence profile alignment[ J]. Bioinformatics,2013,29 ( 20): 2588 - 2595.、
3. 吴辉 . 利用序列信息预测蛋白质二级结构的深度学习模型研究[D] .天津大学. 2017