**基于深度学习的蛋白质残基接触位点预测系统设计**

**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。

蛋白质与配体物质相互作用是完成特定功能的关键，因此对蛋白质相互作用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技术的交差发展为深入了解蛋白质相互作用机理提供了基础条件。

本次课程实验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** 蛋白质残基相互作用；金属离子锌；接触位点预测；深度学习；卷积神经网络

**Prediction system of protein residue contact sites based on deep learning**

**Abstract**

With the smooth implementation of the Human Genome Project, a large amount of gene and protein sequence data are generated every day. 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 cross development of biological information and computer technology provides the basic conditions for a deep understanding of protein interaction mechanisms.

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; using sampling methods to preprocess the data to improve the degree of data imbalance; Construct a densely connected network model and a support vector machine model to train and predict the pre-processed sample data, and evaluate the performance of the model. The prediction accuracy of the densely-connected network model and the support vector machine model for metal ion zinc-binding protein interaction sites reached 90.07% and 92.03%, respectively; through the model fusion of the densely-connected network model and the support vector machine model, the densely-connected network The model is responsible for the feature extraction task, and the support vector machine model is responsible for the classification task, which improves the prediction accuracy of the metal ion zinc binding protein interaction site to 92.18%.

**Keywords** protein residue interaction; metal ion zinc; interaction site prediction; deep learning; convolutional neural network

目 录

[第1章 绪论 1](#_Toc41174114)

[1.1 课题研究背景、目的以及意义 1](#_Toc41174115)

[研究背景 1](#_Toc41174116)

[研究目的及意义 2](#_Toc41174117)

[1.2 国内外研究现状 3](#_Toc41174118)

[1.2.1 蛋白质-蛋白质相互作用位点预测研究现状 3](#_Toc41174119)

[1.2.2 锌离子结合蛋白质作用位点预测研究现状 5](#_Toc41174120)

[1.3 论文主要研究工作 5](#_Toc41174121)

[第2章 蛋白质功能位点预测相关知识 6](#_Toc41174122)

[2.1 蛋白质 6](#_Toc41174123)

[2.1.1 蛋白质的定义 6](#_Toc41174124)

[2.1.2 蛋白质的结构 6](#_Toc41174125)

[2.1.3 蛋白质功能位点 7](#_Toc41174126)

[2.2 蛋白质功能位点预测流程 8](#_Toc41174127)

[2.3 主要预测方法 9](#_Toc41174128)

[2.3.1 人工神经网络 9](#_Toc41174129)

[2.3.2 卷积神经网络 11](#_Toc41174130)

[2.3.3 密集连接网络 13](#_Toc41174131)

[2.3.4 支持向量机 13](#_Toc41174132)

[2.3.5 深度学习中的关键技术 14](#_Toc41174133)

[2.4 预测算法的评价指标 15](#_Toc41174134)

[第3章 数据集的构造 1](#_Toc41174135)

[3.1 数据集的处理 1](#_Toc41174136)

[3.1.1 数据集的选取 1](#_Toc41174137)

[3.2 特征的选择 1](#_Toc41174138)

[3.2.1 氨基酸的离散型编码 1](#_Toc41174139)

[3.2.2 蛋白质位置特异性得分矩阵 2](#_Toc41174140)

[3.2.3 蛋白质二级结构 3](#_Toc41174141)

[3.3 数据平衡化处理 3](#_Toc41174142)

[3.3.1 欠采样 4](#_Toc41174143)

[3.3.2 过采样 4](#_Toc41174144)

[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 5](#_Toc41174145)

[4.1 数据集预处理 5](#_Toc41174146)

[4.2 网络模型结构 6](#_Toc41174147)

[4.3 性能评估 12](#_Toc41174148)

[4.4 本章小结 14](#_Toc41174149)

[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 15](#_Toc41174150)

[5.1 数据预处理 15](#_Toc41174151)

[5.2 SVM模型参数 15](#_Toc41174152)

[5.3 性能评估 15](#_Toc41174153)

[5.4 本章小结 16](#_Toc41174154)

[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 16](#_Toc41174155)

[6.1 数据预处理 16](#_Toc41174156)

[6.2 模型融合 16](#_Toc41174157)

[6.3 性能评估 17](#_Toc41174158)

[6.4 本章小结 18](#_Toc41174159)

[结 论 18](#_Toc41174160)

[致 谢 20](#_Toc41174161)

[参 考 文 献 21](#_Toc41174162)

# 绪论

## 课题研究背景、目的以及意义

### 研究背景

伴随着人类基因组计划的顺利实施以及后基因组时代的到来，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。

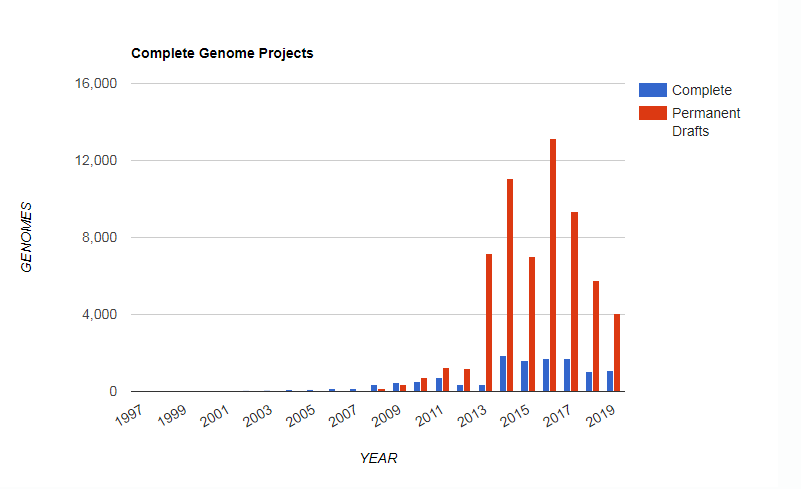


图 1.1 1997 年到 2019 年被完整测序的基因组数目

蛋白质是生命活动的体现者，它是一切生物藉以表现生命的最重要的基本单元，生物体系的运作与蛋白质间的相互作用密不可分。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]。

在生物信息学中，蛋白质结合位点的预测一直是主要的研究课题。目前传统蛋白质结合位点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。现有的生物信息学的研究方法有基于统计分析的方法与基于机器学习的方法。相较于统计分析方法，机器学习的方法可捕获更加复杂的序列特征，因此在该问题上的预测效果更佳。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序列的特性。

### 研究目的及意义

传统的蛋白质与配体相互作用检测方法或多或少存在问题，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独特的领域——深度学习。目前深度学习在自然语言处理、图像处理和机器翻译等方面以及取得了巨大的成果，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

蛋白质通过与其他蛋白质，化合物，RNA和DNA相互作用，在几乎所有生物事件中都起着至关重要的作用。了解相互作用位点的特征是了解分子识别过程的基础。蛋白质很少单独发挥作用，经常通过成为大型分子网络的一部分来发挥其功能，并通过复杂的蛋白质-蛋白质相互作用（PPI）调控网络来协调作用。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。PPI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]。特别地，鉴定两个相互作用的蛋白质之间的结合位点将为理解和确定蛋白质复合物的功能和结构，促进药理学靶标的鉴定以及最终用于药物设计提供有价值的线索。因此预测蛋白质相互作用位点具有重要意义。

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]。在他们的2003年研究[中](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3429425/#pone.0043927-Ofran1)预测精度达到了70%，正确地预测了20％（66/333）的复合物中至少一个相互作用位点。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公共数据库（包括相互作用蛋白数据库，生物分子相互作用网络数据库和PDB）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，基于序列的预测实现了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。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/#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

本课题围绕深度学习在蛋白质残基接触位点预测上的应用，主要利用卷积神经网络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

蛋白质作为生命的物质基础，承担着关键的生命活动。蛋白质是由氨基酸脱水缩合而成的大分子化合物，是构成生物细胞的基本物质之一。

### 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。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。下面是氨基酸的结构式如图2.1.1所示：

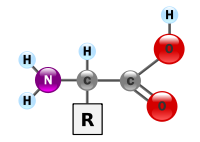


图 2.1 氨基酸的结构式

### 蛋白质的结构

蛋白质种类繁多，功能复杂。一般来说蛋白质的结构决定了蛋白质的功能，但是有些蛋白质的结构不同但是功能却很类似，例如载体蛋白，它们功能大不相同，但是都是都拥有运输物质的功能。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四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)（Primary structure）：表示氨基酸组合在一起形成蛋白质的特殊序列。

[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)（Secondary structure）：蛋白质的二级结构一般分为α螺旋和β折叠这两种结构，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。

[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)（Tertiary structure）：[蛋白质](https://www.thoughtco.com/protein-function-373550)多肽链的全面3-D结构。有几种类型的键和力将蛋白质保持在其三级结构中。在蛋白质的三级结构中，疏水相互作用极大的促进了蛋白质的折叠和成形。多肽链中和氨基酸“R”基团之间的氢键通过将蛋白质保持在疏水相互作用所建立的形状中，从而有助于稳定蛋白质结构。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。

[四级结构](https://zh.wikipedia.org/wiki/%E8%9B%8B%E7%99%BD%E8%B4%A8%E5%9B%9B%E7%BA%A7%E7%BB%93%E6%9E%84)（Quaternary structure）：是指由多个多肽链之间的相互作用形成的蛋白质大分子的结构，每个多肽链被称为一个亚基具有四级结构的蛋白质可能包含不止一种相同类型的蛋白质亚基，它们也可以由不同的亚基组成。

### 蛋白质功能位点

化学家 Anfisen 等人提出蛋白质分子的一级结构决定其二维结构，而蛋白质二级结构又决定其三级结构，而三维结构又决定了蛋白质发挥其功能。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]。常用的蛋白质功能位点类型有蛋白质-蛋白质相互作用位点、蛋白质

与金属离子相互作用位点、蛋白质翻译后修饰位点如磷酸化、糖基化和泛素化位点等、蛋白

质与核酸相互作用位点等。一般来讲，蛋白质功能位点的识别对蛋白质生物学功能预测具有重要意义。比如金属结合蛋白质作用位点，金属离子采取不同的方式结合蛋白质，蛋白质具有不同的功能，如结构维护、分子识别与催化，信号传递与存储、基因转录与表达调节等。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 |

## 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，利用该训练的模型对未知测试样本进行预测。预测流程如图2.3所示：

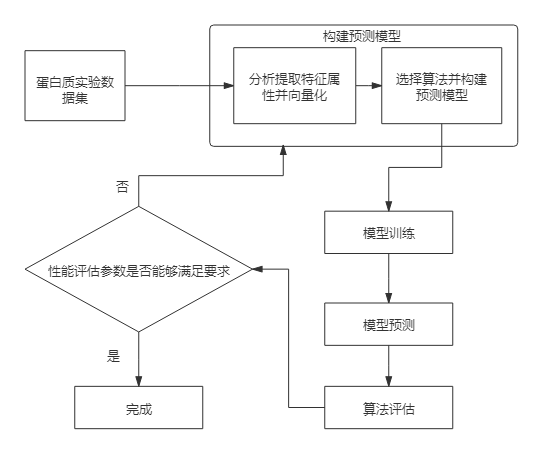


图 2.3 蛋白质残基接触位点预测流程

## 主要预测方法

### 人工神经网络

人工神经网络(Artificial Neural Network，ANN)[21]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。

一个人工神经网络具有成千上万个称为处理单元的人工神经元，它们通过节点相互连接。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。我们先忽略激活函数，神经元包含输入和输出。一般情况下一个人工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，其它各层均被称为隐含层。典型的人工神经网络如图2.4所示：

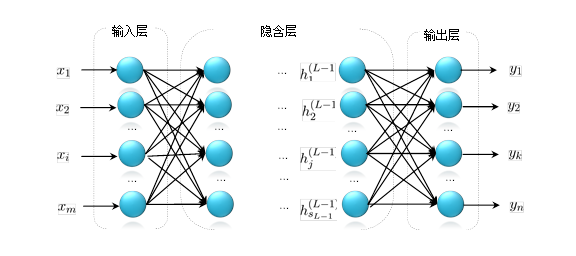


图 2.4 典型的人工神经网络结构

输入向量为：

输出向量为：

第隐含层各神经元的输出为：

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：

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：

### 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，是目前最流行的[深度学习](https://ww2.mathworks.cn/discovery/deep-learning.html)算法之一。深度学习是机器学习的一种，该算法的模型直接从图像、视频、文本或声音中学习执行分类任务。卷积神经网络尤其适用于发现图像中的模式，从而识别物体、人脸和场景。卷积神经网络直接从图像数据中学习，使用模式对图像进行分类，无需手动特征提取。卷积神经网络可能有数十个甚至数百个层，每个层学习检测图像的不同特征。卷积层滤波器会应用到不同分辨率的各个训练图像，且每个卷积图像的输出会用作下一层的输入。滤波器最初可以是非常简单的特征，例如亮度和边缘，然后增加复杂度，直至可以唯一地确定目标特征。

卷积神经网络的特别之处在于神经元之间的连接构造方式。在前馈神经网络中，神经单元被组织成层，给定层的神经单元仅从下面层的神经单元获得输入。卷积神经网络中的神经单元具有空间排列，在每一层，神经单元被组织成称为特征图的二维网格，每个特征图都是卷积的结果，这意味着将相同的卷积滤波器（权重集）应用于下一层的每个位置。因此，二维网格上特定位置的神经单元只能从下面一层的相似位置的单元接收输入。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2.5所示：

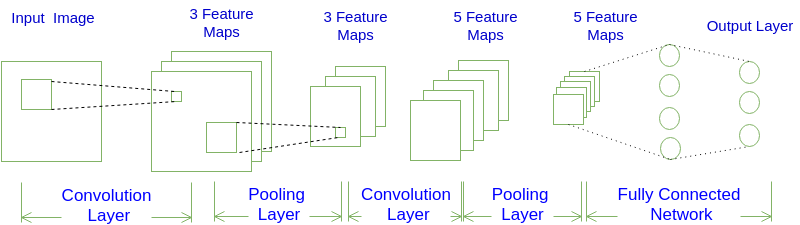


图 2.5 典型的卷积神经网络结构

卷积神经网络通常包含以下几种层：

1. 卷积层（Convolutional layer）

卷积层是从输入图像提取特征的第一层。卷积层通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。

1. 线性整流层（Rectified Linear Units layer, ReLU layer）

在神经网络中，激活函数负责将来自节点的加权总和输入转换为节点或该输入的输出的激活。整流线性激活函数是分段线性函数，如果为正，将直接输出输入，否则将输出零。它已成为许多类型的神经网络的默认激活功能，因为使用该模型的模型更易于训练，并且通常可以获得更好的性能。

1. 池化层（Pooling layer）

当图像太大时，“合并层”部分会减少参数的数量。空间池化也称为子采样或下采样，可降低每个地图的维数，但保留重要信息。空间池可以具有不同的类型：最大池化、平均池化和汇总池化。

1. 全连接层（Fully-Connected layer）

把所有局部特征结合变成全局特征，用来计算最后每一类的得分。

### 密集连接网络

密集连接网络（DenseNet）是一种结构特殊的卷积神经网络，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DenseNet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，Densenet网络可以减轻梯度弥散问题，加强特征传播，鼓励特征重用，大大减少参数数量。Densenet网络结构示意图如图2.6所示：

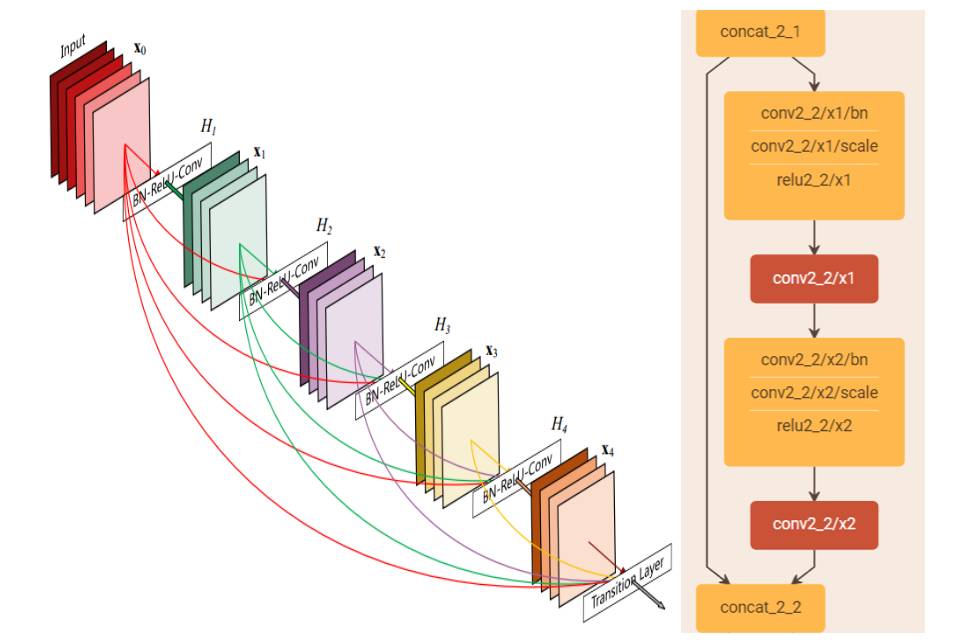


图 2.6 密集连接（Densenet）网络结构

### 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。具有不同模型配置的神经网络集合可减少过拟合，但需要训练和维护多个模型的额外计算费用。通过在训练期间随机删除节点，可以使用单个模型来模拟具有大量不同的网络体系结构。这称为[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) ，它提供了一种计算上非常低成本且非常有效的正则化方法，以减少过拟合并改善各种深度神经网络中的泛化误差。该方法如图 2.7 所示。

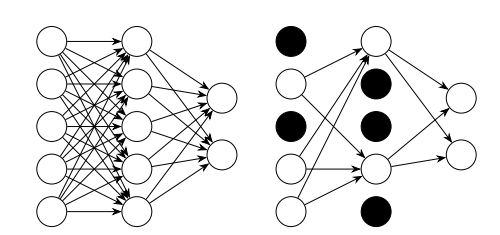


图 2.7 标准 Dropout 的示例

1. 损失函数（Loss Function）

在优化算法的上下文中，用于评估候选解决方案的函数（即一组权重）称为目标函数。我们可能寻求最大化或最小化目标函数，这意味着我们正在寻找分别具有最高或最低得分的候选解决方案。通常，使用神经网络，我们试图将误差最小化。这样，目标函数通常被称为成本函数或损失函数。

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。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。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

True Positive（真正， TP）被模型预测为正的正样本；True Negative（真负 ， TN）被模型预测为负的负样本； False Positive （假正， FP）被模型预测为正的负样本； False Negative（假负， FN）被模型预测为负的正样本。

# 数据集的构造

## 数据集的处理

### 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

本课题实验以长度为 L(L= 15，16,17……35）的氨基酸序列进行数值转化变成可计算的数据。这里参考普遍用于自然语言处理问题中的编码方式，既 one-of-K 编码。在one-of-K 编码中，对于氨基酸的编码 K 可以等于 20 代表 20 种常见的氨基酸，每一个氨基酸都用一个维度为 20 的向量表示，只有对应这个氨基酸的位置是 1 其他位置都是 0。然而对于那些蛋白质片段长度不足 K的，需要用额外的特殊符号 “-”表示，因此实际的 K=21。对于其他非常见氨基酸或者不确定的氨基酸均给定统一合适数值，最终一个蛋白质序列被转化成shape为（L，K）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。

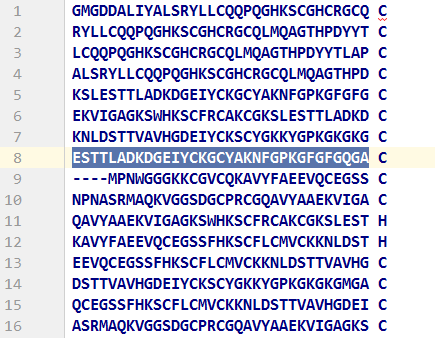


图 3.1 长度为 33 的蛋白质片段

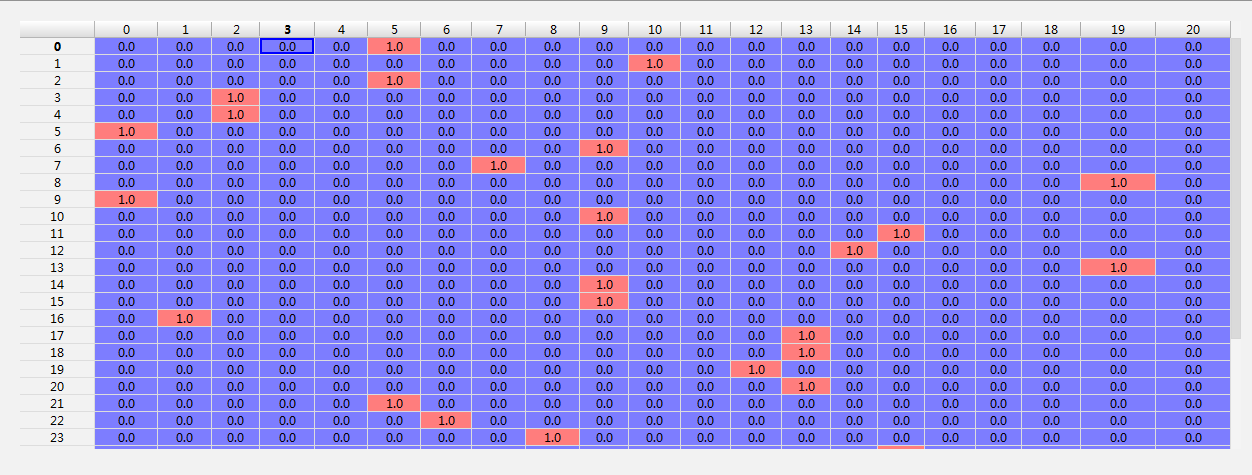


图 3.2 长度为 33 的蛋白质片段one-of-K编码部分数据

### 蛋白质位置特异性得分矩阵

位置特异性打分矩阵（PSSM）表达了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Lx20维特征。

### 蛋白质二级结构

蛋白质的二级结构是指由于骨架原子之间的相互作用而在多肽内形成的局部折叠结构。最常见的二级结构类型是α螺旋和β折叠片。两个结构都通过氢键保持形状，氢键形成在一个氨基酸的羰基O和另一个氨基酸的氨基H之间。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。这样，一个蛋白质二级结构序列由3个抽象字符构成蛋白质主要二级结构，把八类二级结构粗略地归类为α-螺旋（Helix）、β-折叠（Sheet）和无规则卷曲（Curly），则长度为L的蛋白质序列可提取Lx3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）分为两类：欠采样（UnderSampling）和过采样（OverSampling）。欠采样是从多数类样本中按照某种规则选择少量样本，再合并原有少数类样本作为新的训练数据集。过采样与欠采样相反，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。采样方法分类如图3.3所示：

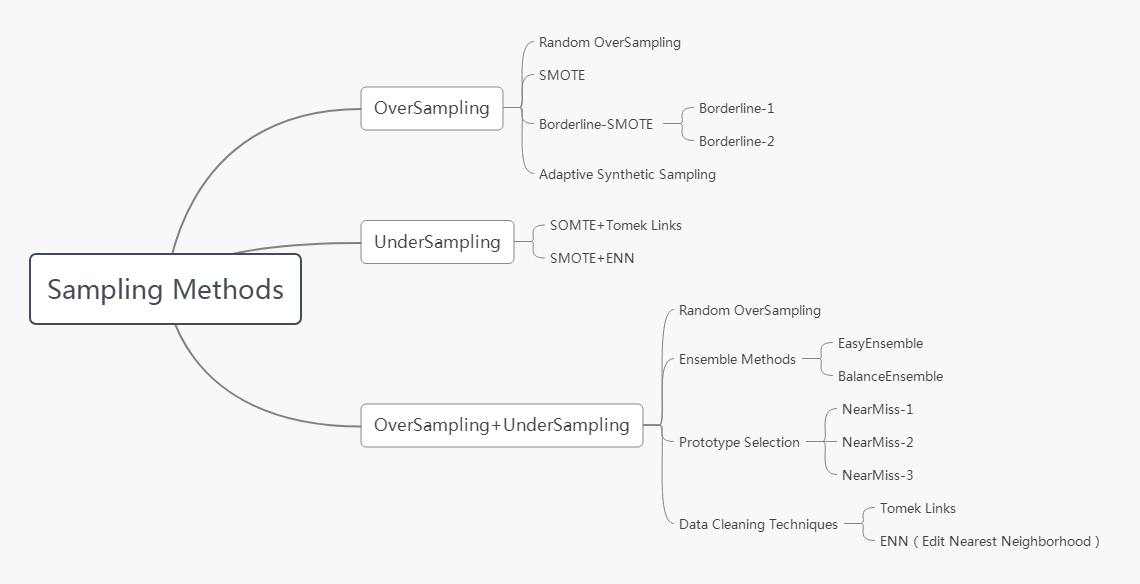


图 3.3 采样方法分类

### 欠采样

欠采样是从多数类样本中按照某种规则选择少量样本，再合并原有少数类样本作为新的训练数据集。本次课题实验中采用的是随机欠采样（RandomUnderSampling）来减少多数类的样本数量。如果随机删除的数量过大，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4.1.1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4.1、图4.2所示：

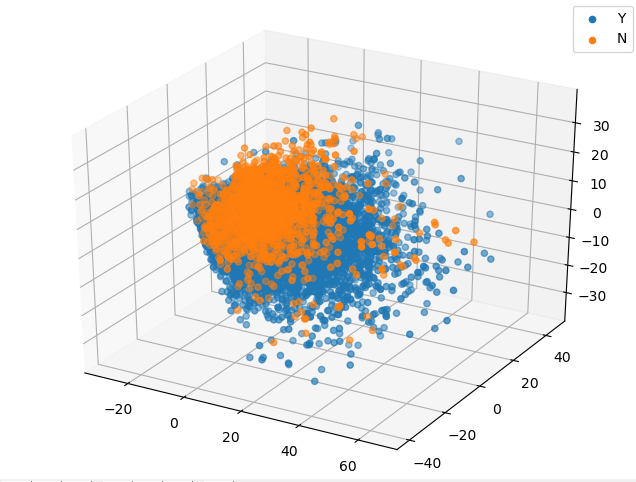


图 4.1 预处理前正负样本数据分布

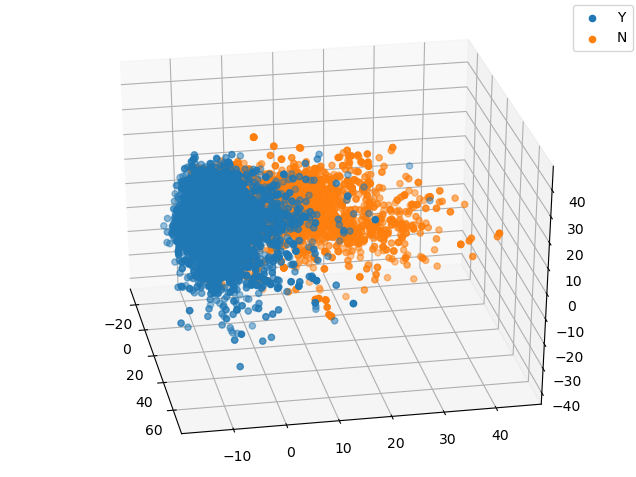
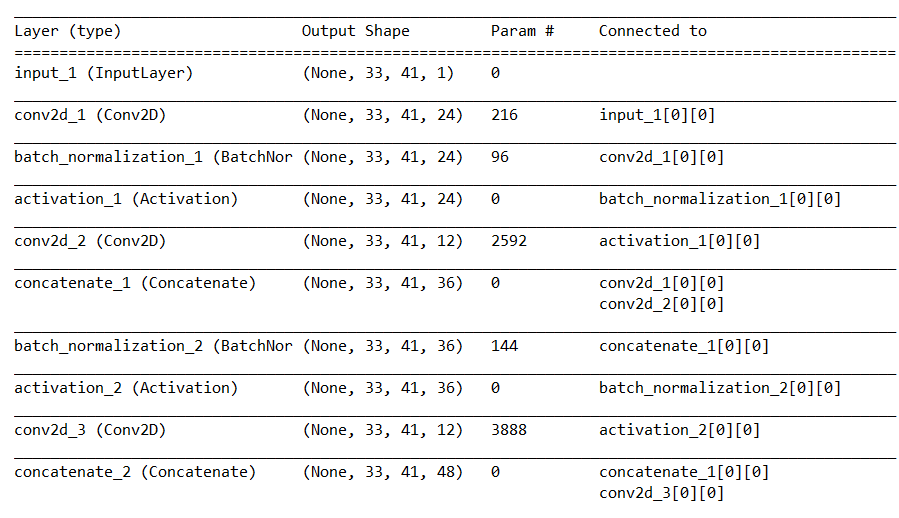
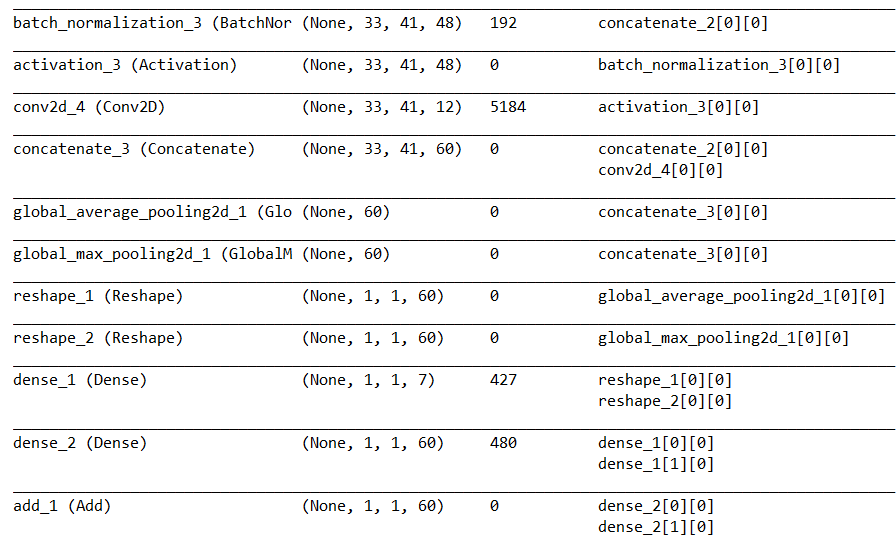


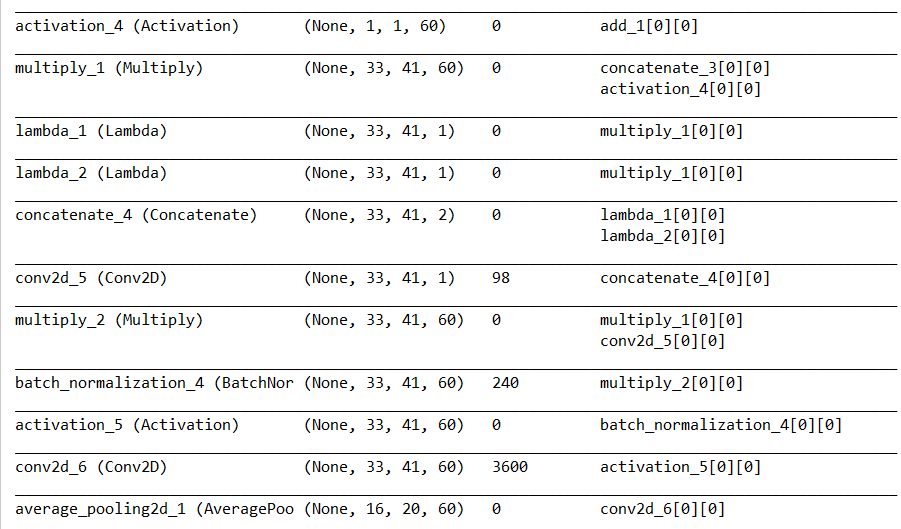
图 4.2 预处理后正负样本数据分布

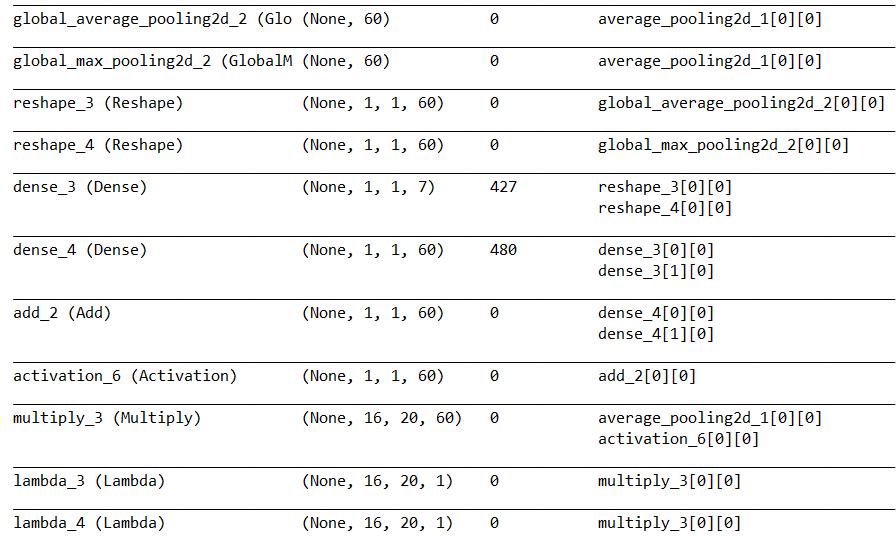
## 网络模型结构

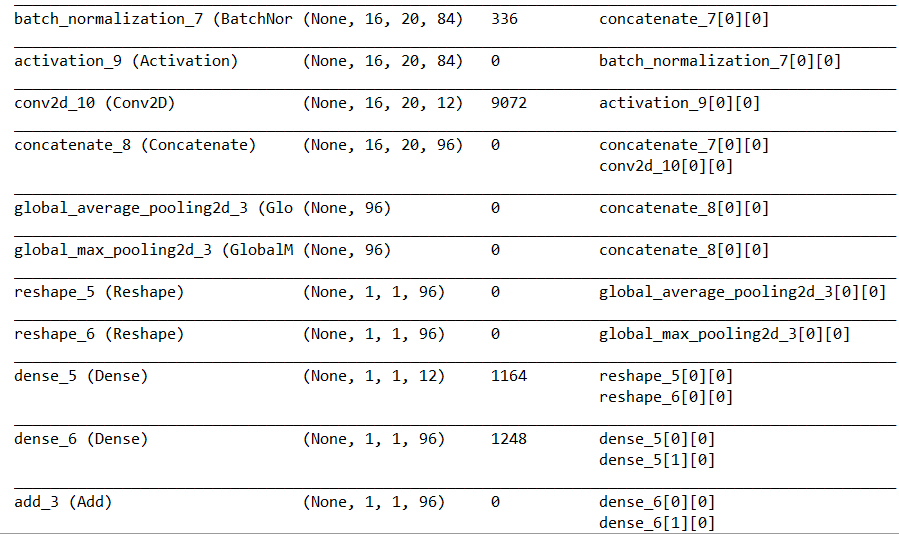
本次课题实验选用的是密集连接网络（Densenet），密集连接网络脱离了加深网络层数(ResNet)和加宽网络结构(Inception)来提升网络性能的定式思维,从特征的角度考虑,通过特征重用和旁路(Bypass)设置,既大幅度减少了网络的参数量,又在一定程度上缓解了gradient vanishing问题的产生。网络模型结构如图4.3所示：

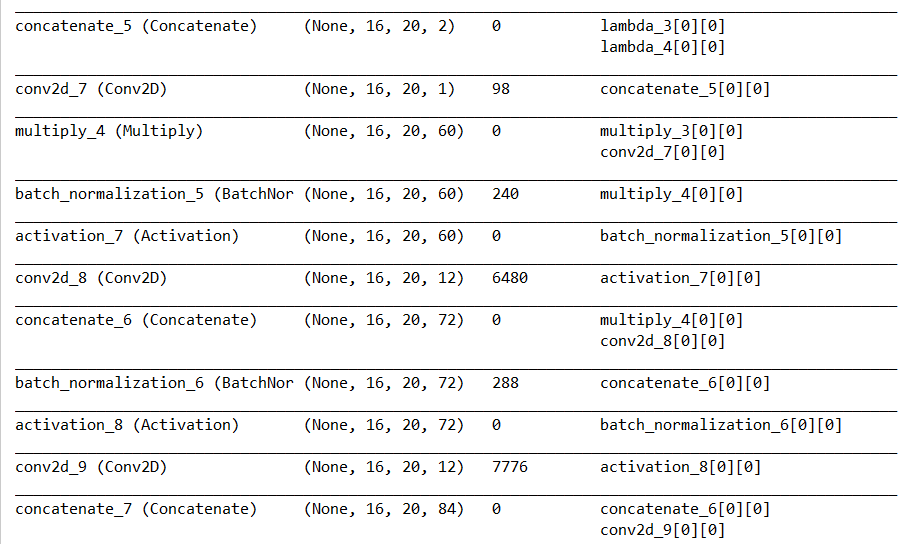


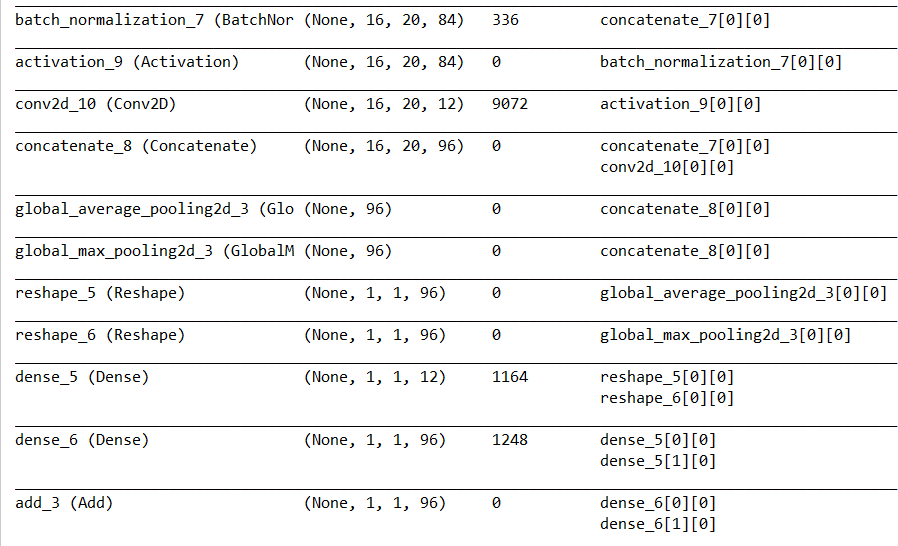


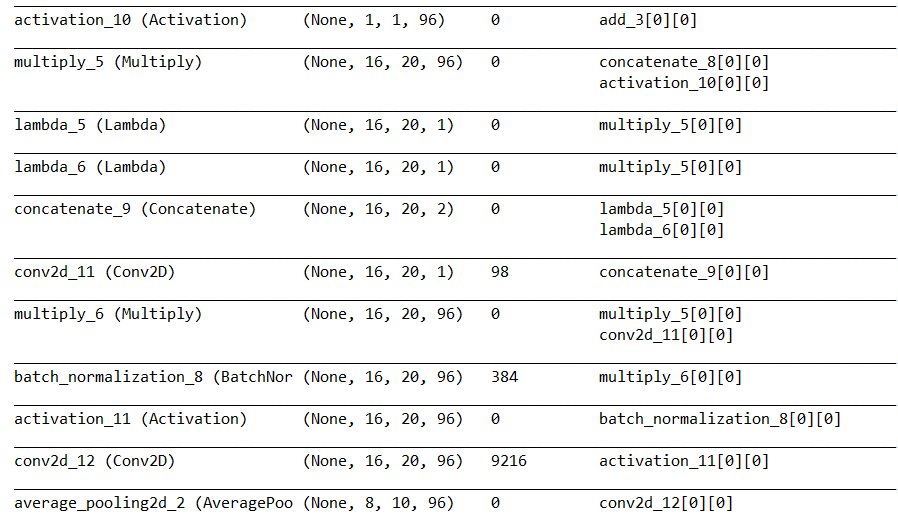


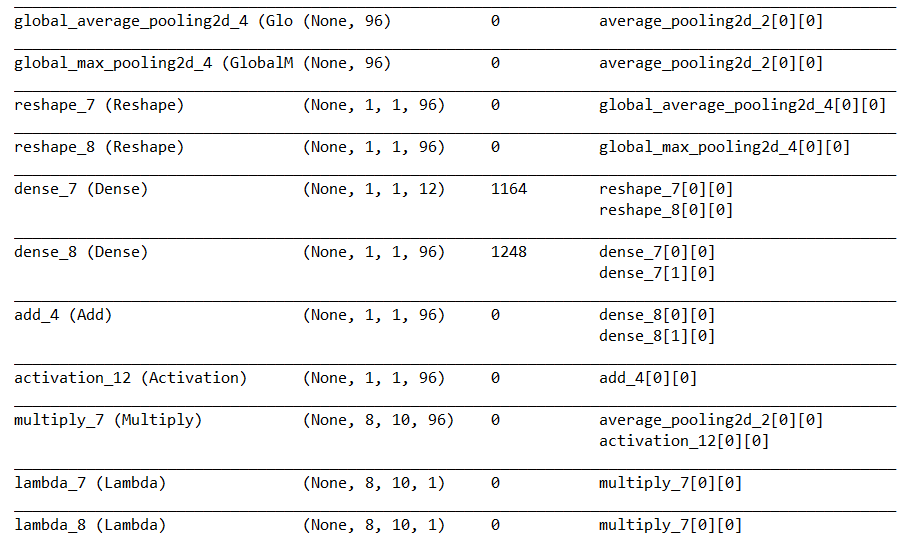


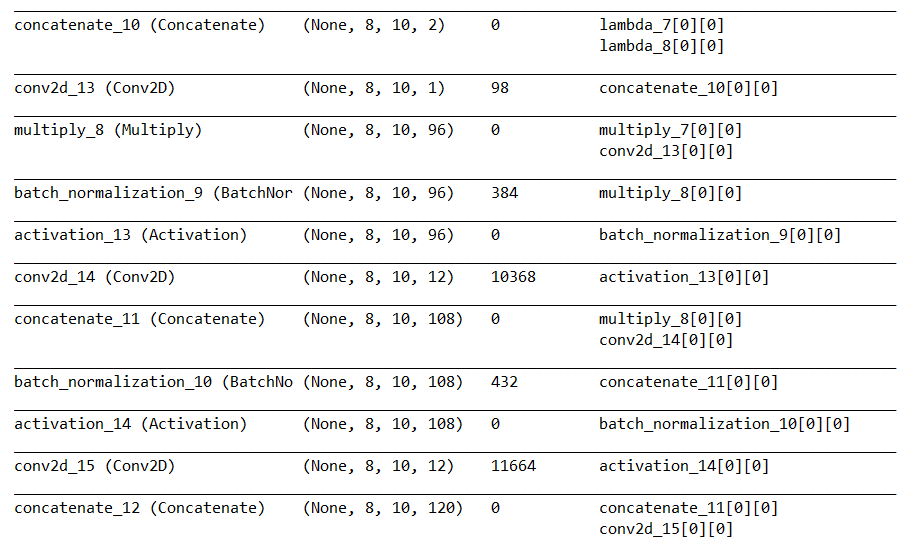


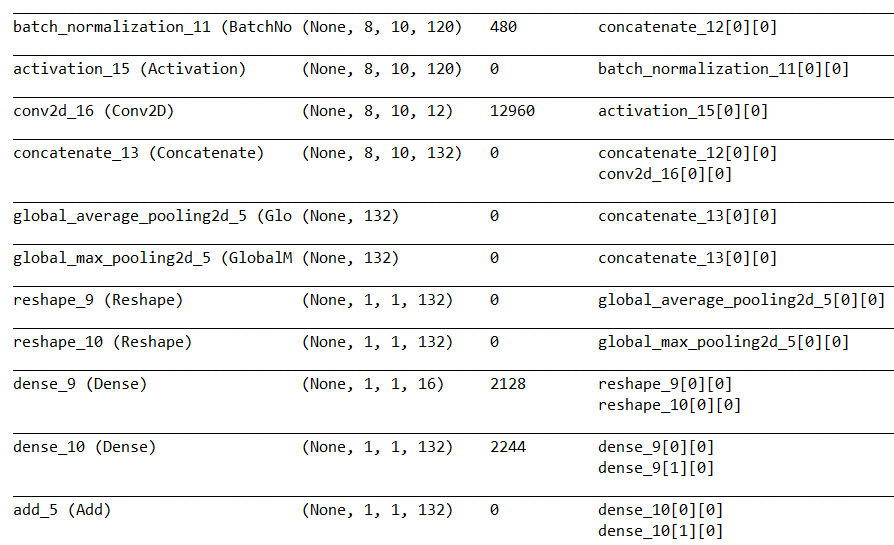


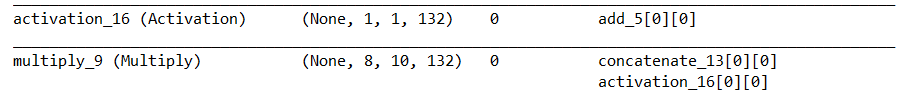












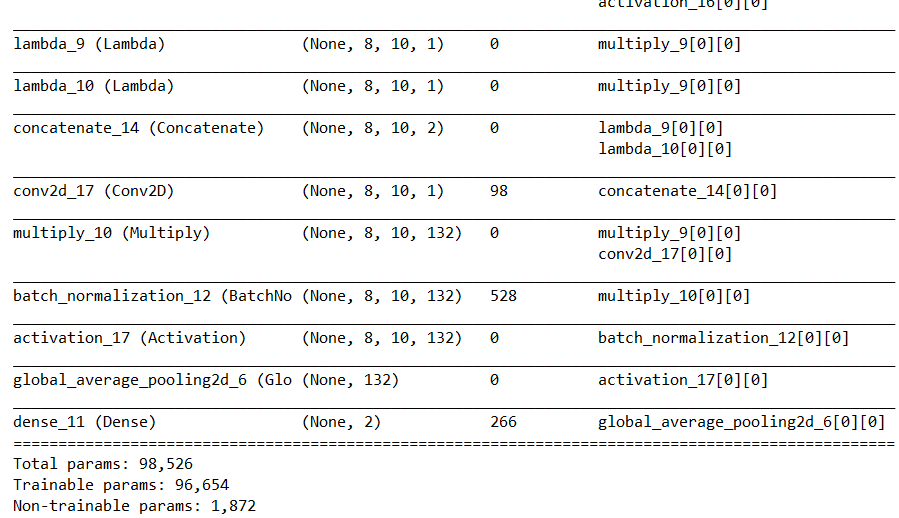


图 4.3 Densenet网络模型结构

## 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：

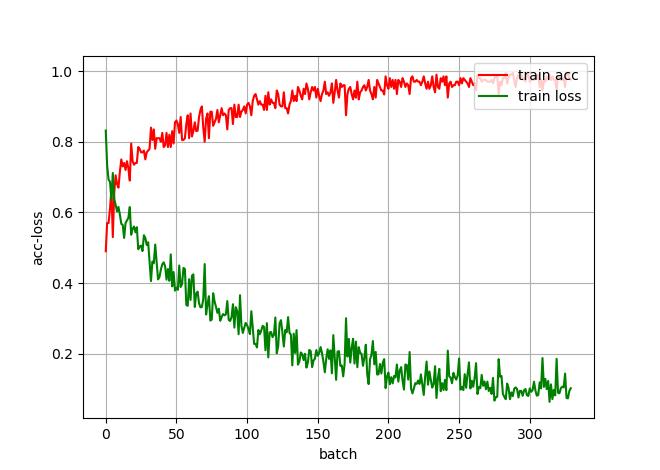


图 4.4 训练集的损失和准确度

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多数类样本进行欠采样、利用自适应综合过采样算法对少数类进行过采样，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Densenet网络模型喂入数据训练网络。利用验证集对训练好的网络进行测试，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

本章的数据预处理与第4章的数据预处理一致，所以不过多介绍。详见第4章4.1小节。

## SVM模型参数

SVM模型是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机。SVM模型参数众多，本章节主要关注核函数（Kernal）和惩罚系数（C）。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，在训练集上的准确度和验证集上的准确度分别如表5.1所示：

表 5.1 训练集和验证集的准确率

|  |  |
| --- | --- |
| 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

本章首先依旧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SVM模型喂入数据训练网络。利用验证集对训练好的网络进行测试，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

本章的数据预处理与第4章的数据预处理一致，所以不过多介绍。详见第4章4.1小节。

## 模型融合

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维特征做为输入数据喂入到SVM网络模型进行分类训练。

## 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。在训练集上的准确度和验证集上的准确度如图6.1所示：

表 6.1 训练集和验证集的准确率

|  |  |
| --- | --- |
| 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相互作用位点预测具有重要的意义。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的相互作用；针对蛋白质相互作用位点预测的典型常见问题——数据不平衡问题，本课题尝试利用采样方法对数据进行预处理，使用随机欠采样算法对正样本即多数类样本进行随机删除，减少多数类样本的数量。使用自适应综合过采样算法对负样本即少数类样本进行插入，增加少数类样本的数量，从而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%。通过以上一系列的实验结果对比，笔者不严谨的得到了一些结论：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；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.