浙 江 大 學

本 科 生 毕 业 论 文（设计）

题目　重大疾病相关的蛋白质组学的模式识别

姓名与学号　秦 臻 3130000210

指导教师　庞天晓、张南松

年级与专业　2013 级 统计学

所在学院　数学科学学院

浙江大学本科生毕业论文（设计）诚信承诺书

1.本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2.本人在毕业论文（设计）中引用他人的观点和参考资料均加以注释和说明。

3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

毕业论文（设计）作者签名：

_____年_____月_____日

本科生毕业论文（设计）任务书

一、题目：重大疾病相关的蛋白质组学的模式识别

二、指导教师对毕业论文（设计）的进度安排及任务要求：

（1）1 月——1 月 14 日：导师下达任务书，对进度、文献和开题提出要求；

（2）1 月 14 日——1 月 23 日：学生确认任务书，对确定的课题搜集相关文献资料，了解问题的背景、应用、研究历史与现状。从中确定论文最终题目。

（3）1 月 23 日——2 月 27 日：对确定的题目进一步展开学习，包括所必需的基础知识及近几年涉及此问题的文章。初步撰写并完成开题报告、文献综述，并提交导师审核。

（4）2 月 27 日——3 月 6 日：组织开题，每位学生准备 10 分钟左右的答辩；

（5）3 月 6 日——4 月 11 日：将定稿的开题报告、文献综述、外文翻译稿上传至教务系统。做中期检查报告。

（6）4 月 12 日——5 月 12 日：完成论文初稿，进行论文稿的修改并最终完成，向导师提交论文终稿。

（7）5 月 13 日——5 月 15 日：导师评阅，学生提交导师填写评语和签字的“毕业论文考核表”及符合规范格式要求的送审论文。

（8）5 月 16 日——5 月 21 日：毕业论文专家评阅.

（9）5 月 22 日——5 月 24 日：评阅结果有修改意见的，根据评阅意见对论文进行修改。

（10）5 月 24 日——5 月 30 日：组织毕业论文答辩。提交最终版毕业论文，并将论文上传至教务系统。

起讫日期 **2017** 年 **1** 月 **1** 日至 **2017** 年 **5** 月 **30** 日

指导教师（签名）_____职称_____

三、系或教研所审核意见:

负责人（签名）_____ 年 月 日

中文摘要

本文研究了人工神经网络在DNA芯片数据分析中的应用。DNA芯片技术是一种生物高新技术，在最近几十年发展迅猛，DNA芯片也可以称作做基因芯片（gene chip）或者基因微阵列（microarray），寡核酸芯片，或DNA微阵列，它通过微阵列技术将高密度DNA片段阵列以一定的排列方式使其附着在玻璃、尼龙等材料上面。通过样品与芯片的杂交，就可以获得样品的遗传信息，其发展和应用的前景广阔。

DNA芯片数据的分析方法是由DNA芯片数据的几个特点决定的。人工神经网络（ Artificial Neural Networks，简写为ANNs）也简称为神经网络（NNs）或称作连接模型（Connectionist Model），是对人脑或自然神经网络（Natural Neural Network）若干基本特性的抽象和模拟。人工神经网络以对大脑的生理研究成果为基础的，其目的在于模拟大脑的某些机理与机制，实现某个方面的功能。

R语言是统计领域广泛使用的一种用来进行数据探索、统计分析、作图的工具。本文在深入理解人工神经网络的基础上，利用R软件丰富强大的功能，编写基于人工神经网络的分析算法软件包，用于DNA芯片数据的分析。该软件包集成多种分析功能，充分发挥了人工神经网络的潜力，应用范围广，算法性能好，具有一定的应用价值。

关键词：DNA芯片；数据分析；人工神经网络；R软件

Abstract

This paper has studied the application of Artificial Neural Networks in data analysis of DNA chip. DNA chip technology is a biotechnology that has developing rapidly in recent years. DNA chip is also called gene chip, gene microarray, Oligonucleotide microarray or DNA microarray. It uses the micro array technology to make high-density DNA fragment arranged in a certain way and attach it to the glass, nylon and other materials. The genetic information of samples can be obtained by hybridization between samples and chips, as a result it has amplitude development and application prospects.

The analysis methods depend on the characteristics of DNA chip data. Artificial Neural Networks(ANNs) are also referred to as Neural Networks(NNs) or Connectionist Model. It is the abstraction and simulation of some basic characteristics of human brain and Natural Neural Networks. The Artificial Neural Networks are based on the research results of the brain. The purpose is to simulate some mechanism and mechanism of the brain to realize the function of a certain aspect

R language is a widely used tool in the field of statistics for data exploration, statistical analysis and mapping. On the basis of deep understanding of Artificial Neural Networks, the passage uses rich and powerful features of R software to write R packages based on Artificial Neural Networks on purpose of analyzing DNA microarray data. This R

package includes many functions and gives full play to the potential of Artificial Neural Networks. What is more, its wide range of applications and the algorithm performance prove its certain commercial value.

Key words: DNA chip; Data Analysis; Artificial Neural Networks; R Software

目　录

1 绪论

在过去的十几年里，计算机技术在生命科学和生物信息学的各个领域中发挥了前所未有的关键作用。新的高效的实验技术不断出现，新技术导致描述 DNA，RNA 和蛋白质的数据增长迅速。这让计算机在实验设计，数据处理和结果解释中扮演着重要角色，推动了生物信息学的发展。本文正是在导师的指导下，对生物信息学的一个具体分支进行了一些讨论和研究。

1.1 选题背景及意义

当今，生物领域的新发现越来越多，传统领域的研究逐渐转向依赖多个纬度和不同尺度的数据分析得出的结果。DNA序列和结构与功能数据，基因表达数据，临床数据等相互集成。

DNA 芯片技术是一种交叉技术，综合了微电子和分子生物学，借助计算机芯片的概念完成生物学实验。早在 1991 年，美国的 Affymetrix 公司生产了世界第一块合成的微阵列芯片，之后 1994 年斯坦福大学制备了第一块微阵列 cDNA 芯片。从此，DNA 芯片技术进入了大规模研究与应用的时代。Affymetrix 公司经过发展，率先使用了光导向平板印刷技术，这种方式可以在硅片上合成寡核苷酸点阵的高密度芯片，因此该公司在芯片分析领域取得领先。该公司与惠普公司合作开发出专用的能扫描 40 万点点阵的基因芯片扫描仪，同时又开发出同时可平行通过几块芯片的流路工作站和计算机软件分析系统。组合成一套较完整的芯片制造、杂交、检测扫描和数据处理系统。

其他发达国家的生物技术公司也纷纷加入 DNA 芯片的研究，例如 Q-Pix 克隆挑拣仪及 Q—Fill 制芯片设备。德国肿瘤研究所则用就位合成的肽核酸低密度的作表达谱及诊断用的探针芯片。如今，DNA 芯片已经在基因序列分析、基因诊断、基因表达研究、基因组研究、发现新基因及各种病原体的诊断等生物医学领域表现出巨大的应用价值。

DNA 芯片技术为生命研究带来了光明的研究前景，相对应的极大的数据量，在生物信息的存储，获取，分析和可视化方面，对于现有的分析算法和软件提出了更高的需求。而与此同时，计算机技术的发展提供了多种数据分析方法。

1.2 研究框架与研究方法

DNA 芯片数据分析的研究尚处于发展阶段，仍有很多方面存在不足。虽然人工神经网络算法的发展已经经历了很长一段时间，但是对于 DNA 芯片数据上的应用，人们还没有给予足够的重视。同时，算法本身也有很多可以优化的环节。

本文针对 DNA 芯片数据中的问题，对在基于人工神经网络的算法设计和实际

应用中存在的问题进行系统而深入的研究。内容主要包括四个方面：DNA 芯片数据的预处理；人工神经网络算法的设计；基于神经网络的数据分析；算法的性能比较和理论解释。本论文的研究框架如图 1.1 所示：

第一章 绪论。本章简单介绍了生物信息学和 DNA 芯片的技术发展历史与研究现状，提出了本文的研究框架与研究方法。

第二章 DNA 芯片数据。本章详细描述了 DNA 芯片数据的特点，从 DNA 芯片中的一部分入手，揭示芯片数据特点和数据分析之间的联系。之后本章介绍了获取 DNA 芯片数据的途径。

第三章 人工神经网络算法及实现。本章介绍了人工神经网络的发展进程，基本概念和深刻原理。根据人工神经网络的基本理论建立分析数据的多种算法，利用算法对 DNA 芯片数据进行分析。

第四章 基本结论及应用方向。通过前三章的介绍，本章对于论文的内容进行总结，指出论文创新之处，简述今后的应用方向。

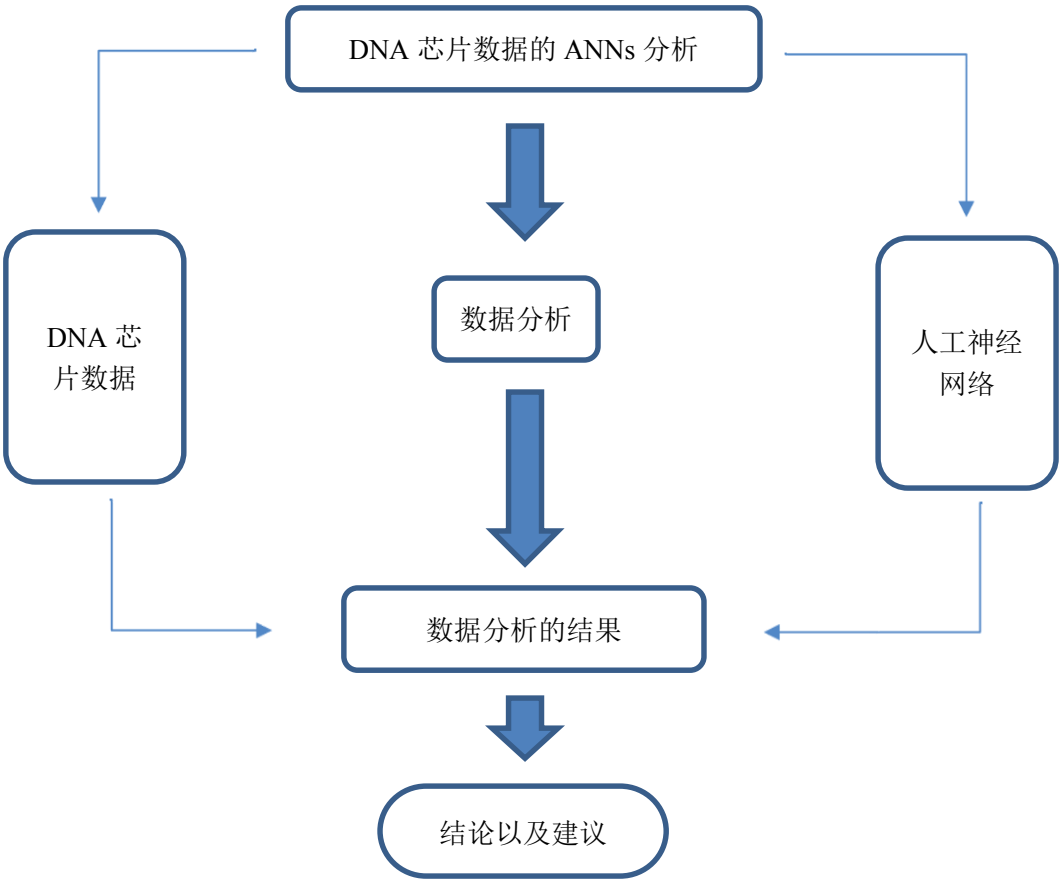

图 1.1 论文研究框架

为了深刻表现人工神经网络在 DNA 芯片分析中的应用，本文使用了多种研究方法。首先是文献研究法。文献研究法是从某一个研究课题和角度出发，通过调

查文献来获得资料，从而正确地全面地了解掌握所要研究问题的一种方法。然后是模拟法。模拟法是先依照原型的主要特征，创设一个相似的模型，然后通过模型来间接研究原型的一种研究方法。Gene Expression Omnibus（GEO）中有很多DNA 芯片数据，可以用这些数据来进行分析。接着是数量研究法。数量研究法也称“统计分析法”,指通过对研究对象的多种不同的可能的数量关系的分析研究，认识和揭示事物间的相互关系、变化规律和发展趋势，借以达到对事物的正确解释和预测的一种研究方法。最后是探索性研究法。探索性研究法是高层次的科学研究活动。它是用已知的信息，探索、创造新知识，产生出新颖而独特的成果或产品。探索性研究在本论文中也是非常重要的一部分，是论文的精华。

2 DNA 芯片数据

DNA 芯片可以获得多种数据，其中最重要的一种就是基因表达谱数据。基因表达谱是一种在分子生物学领域，借助 cDNA、表达序列标签（EST）或寡核苷酸芯片来测定细胞基因表达情况（包括特定基因是否表达、表达丰度、不同组织、不同发育阶段以及不同生理状态下的表达差异）的方法。基因表达图谱从逻辑上说是基因测序的下一个步骤：基因序列包含细胞可能存在的功能的信息，而基因表达谱则包含细胞实际上正在完成的工作的信息。

2.1 DNA 芯片数据的特点

微阵列基因芯片技术的成熟带来了基因表达谱数据分析方法的日新月异。从本质上讲，通过基因芯片技术实验所直接获得的是一个基因表达谱，基因芯片技术的实际应用就是通过对基因表达谱的生物信息学处理来实现的。本文采用的数据是使用测量相对荧光强度而获得的基因表达谱数据，为了进行随后的数据处理，首先要把基因表达谱数据从杂交图像中提取出来，基因表达谱数据是以矩阵的形式表示的，也即基因表达矩阵，矩阵的各行表示不同的基因，矩阵的各列表示不同的样本或实验条件。例如对于肿瘤芯片的数据，红光和绿光表示不同的 RNA，那么有以下矩阵：

$$x_{np} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

其中 n 表示基因个数，p 表示样本个数，矩阵元素是红光强度与绿光强度之比。通过基因芯片，就可以把复杂的人类 DNA 特征转化为矩阵形式，为接下来的分析工作提供便利。

DNA 芯片数据应用到癌症诊断和分型时，从两类到多类问题的扩展还需要寻

找更专用更有效的方法。再次，算法性能的好坏在很大程度上依赖于数据本身和初始条件的选择，需要寻找可靠性更好的算法。DNA 芯片技术为生物学和医学研究带来前所未有的机遇的同时，其所产生的海量和复杂的 DNA 芯片数据却对现有的数据处理和分析方法提出了巨大的挑战。

第一，DNA 芯片数据具有很高的维度，每一个维度代表一个独立的基因，通常有五千至一万五千维，而且这些基因维度之间又有非常复杂的关系。第二，实验的复杂和费用的昂贵导致 DNA 芯片数据具有较少的样本，并与巨大的基因数目构成不平衡的矛盾。这种矛盾造成大多数经典模式识别方法不能被直接应用，比如，Fisher 线性分析所要求的总类内样本协方差矩阵将成为奇异阵。第三，DNA 芯片数据天生具有高噪声和高变异等数据分析难点。第四，DNA 芯片数据中大量有用变量被隐藏。这可能需要使用概率统计的方法以挖掘和推导这些潜在的生物信息。

2.2 DNA 芯片数据的获取——重要的数据库

许多生物学家面临着一个问题：如何获取和保存实验数据。为了解决这个问题，许多国家和组织建立了基因表达谱的公共数据库，其中最著名的是NCBI-Gene Expression Omnibus(GEO)基因表达数据专用库。

GEO 是美国国家生物技术信息中心（NCBI，National Center for Biotechnology Information）建立的开放型基因表达数据库（如图 2.1）。这个数据库给生物信息学研究人员提供了可靠专业的数据平台。数据库网站同时提供搜索工具，下载工具和简单的分析工具，而且一部分数据经过专员整理，方便访客了解 DNA 芯片信息。

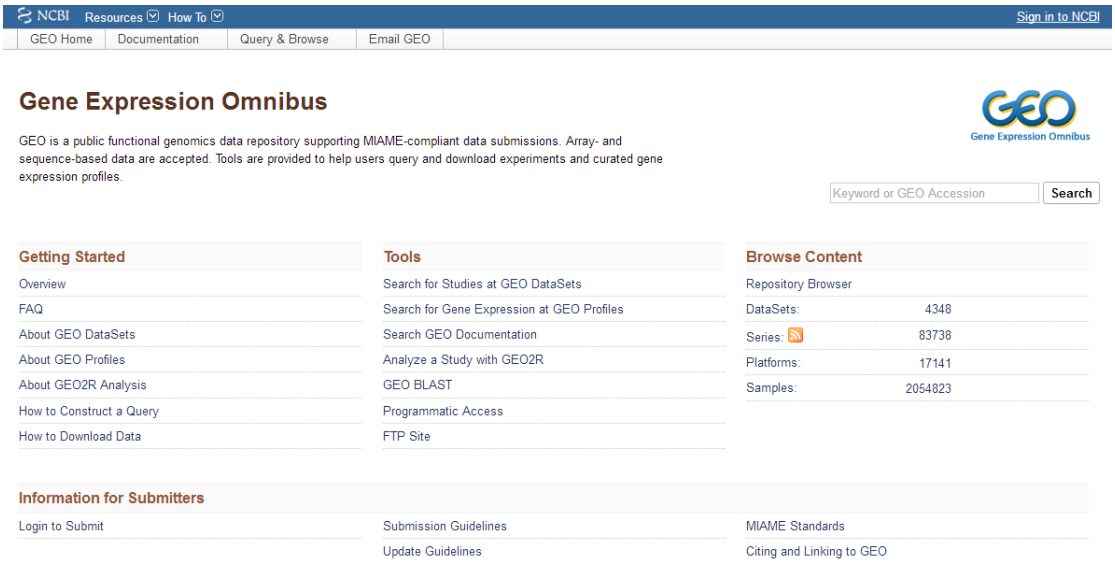

图 2.1 GEO 首页

GEO 是一个国际公共数据库，它可以存档和并且自由地发布研究人员提交的微阵列、下一代测序和其他形式的高通量功能性基因组学数据。GEO 的三个主要目标是：提供稳健而且通用的数据库；提供简单的提交程序和格式；提供用户友好的机制，让用户查询，定位，审查和下载感兴趣的研究和基因表达谱。

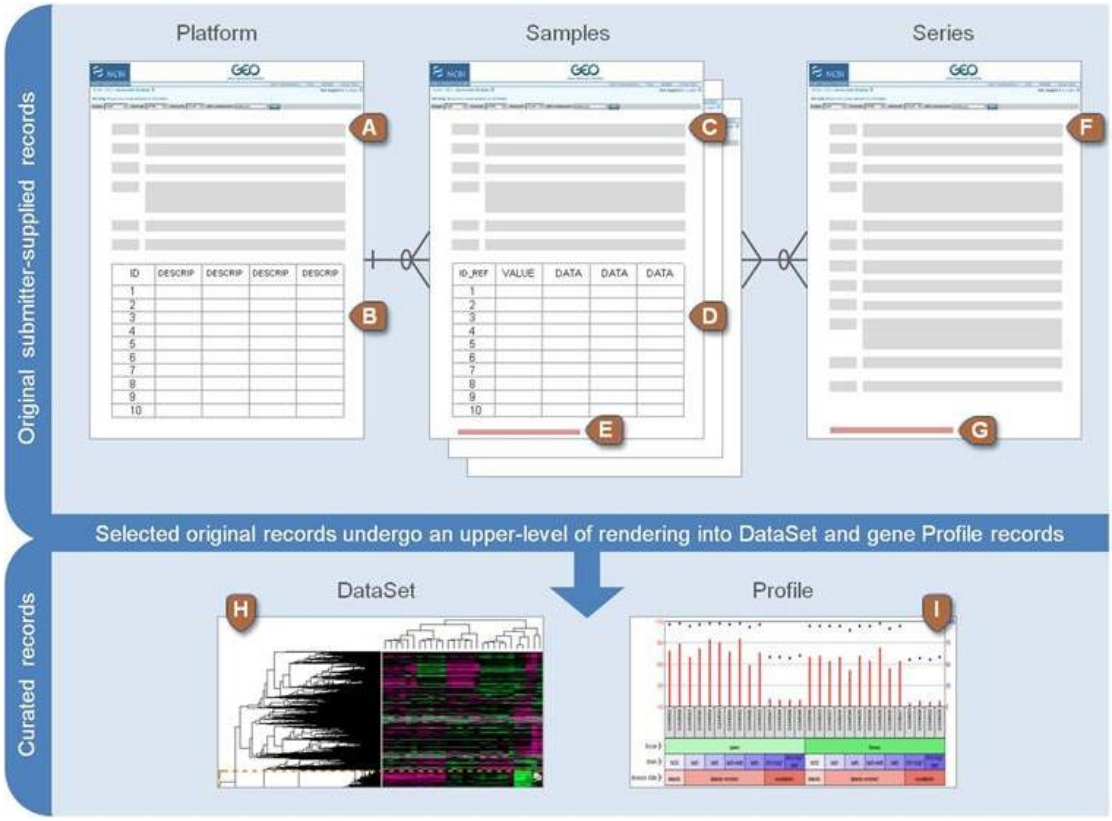

图 2.2 GEO 组织结构

GEO 数据的组织结构分为两个层次：原始数据记录和经过整理的数据记录。原始数据记录包括平台、样本和系列，包含 DNA 芯片数据的全部信息。而精心选择的一部分初始数据，经过详细整理成为数据集和基因谱数据。在本文进行数据分析的时候，既可以利用原始数据记录，也可以利用经过整理的数据记录。通常情况下，他们的处理步骤有细微的不同。GEO 数据集是一个研究级数据库，用户可以搜索他们感兴趣的研究。数据库和数据集包含了所有的原始的数据描述。GEO 谱是一个基因级数据库，用户可以搜索与他们的兴趣相关的基因表达谱。

3 人工神经网络算法及实现

人工神经网络是一种计算模型，它以大脑的神经网络结构为原型。在大脑的简化模型中，它包含了大量的基本计算单元，也就是神经元，以复杂的通信网络形式彼此互联，通过他们，大脑才能够执行高度复杂的计算。人工神经网

络是规则的计算结构，它模仿大脑的计算框架进行建模。

人工神经网络的基本构成单元就是神经元，它是一种有多个输入和一个输出的非线性单元，可以有反馈输入和阈值参数。连接模式是指神经元之间的连接关系，有单层、多层和循环连接模式。前两种连接模式构成的都是前向网络。第三种是包含反馈的连接模式。神经网络模型的基本特征是由其结构决定的，可归纳为：非线性、非局域性、非定性和非凸性。

3.1 算法的深刻原理

神经网络开始使用感知器作为神经元，而现代神经网络使用的通常是 S 型神经元（如图 3.1）。S 型神经元的输入端有多个变量，变量的定义域是[0, 1]。

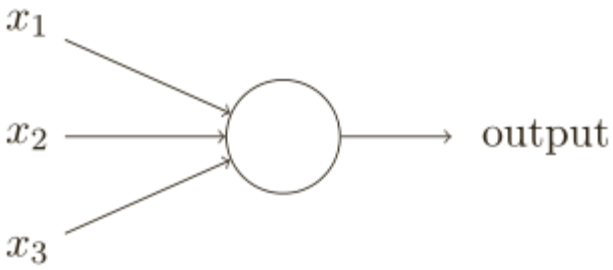

图 3.1 S 型神经元

这里引入权重表示输入端的重要性，记为$w_i, i = 1,2,3 \ldots$，神经元的输出由权重和神经元偏置 b 决定，输出值为$\sigma(w \times x + b)$，这个函数叫做 sigmoid 函数，这个函数定义域是 R，值域是(0, 1)，在定义域上严格单调递增（如图 3.2）。其中：

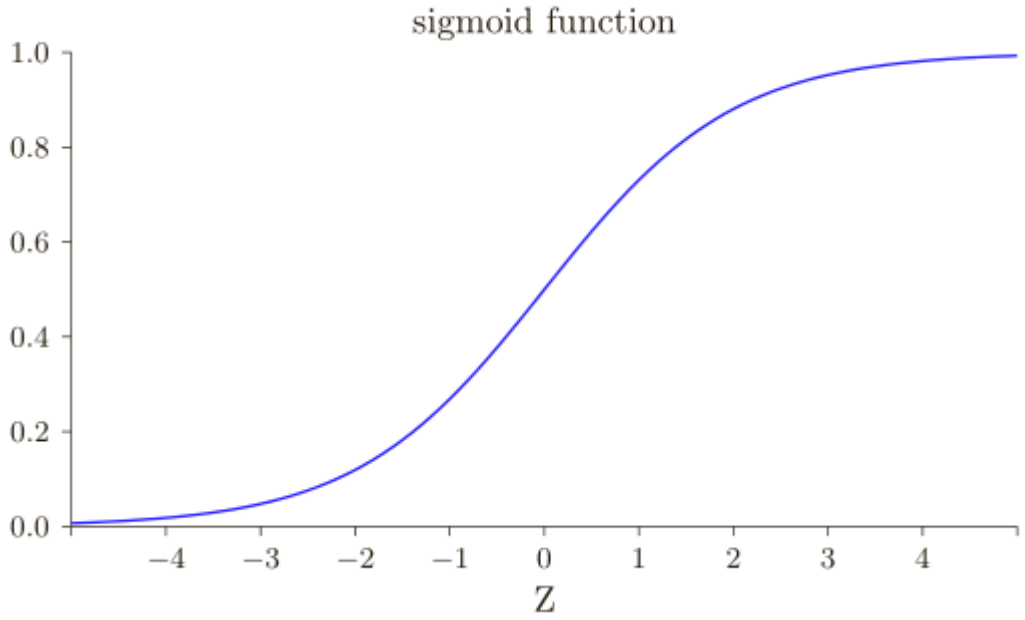

图 3.2 S 型函数

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

S 型函数非常平滑，所以权重和偏置的微小变化Δw_i和Δb会引起输出值的微小变化Δoutput。事实上有：

$$\Delta output \approx \sum_i \frac{\partial output}{\partial w_i}\Delta w_i + \frac{\partial output}{\partial b}\Delta b \tag{2}$$

这时，我们把这个函数σ称为激活函数。激活函数不只有一种形式，这里我们取最常见的激活函数 sigmoid 函数。激活函数反映了每一个感知器输出值与输入值之间的连接关系。

既然神经网络由 S 型神经元组成，那么他们的拓扑结构也是多样化的。最常用的和最经典的是多层感知器。多层感知器是一种前馈的神经网络。前馈型网络的输出只由当前输入、网络参数和结构决定，而与之相对应的是循环网络。循环网络的输出会被先前的输出所影响，所以有短期记忆的性质。这里神经网络的下一层的输入完全依赖于上一层的输出，所以这意味着信息总是向同一个方向传递的。

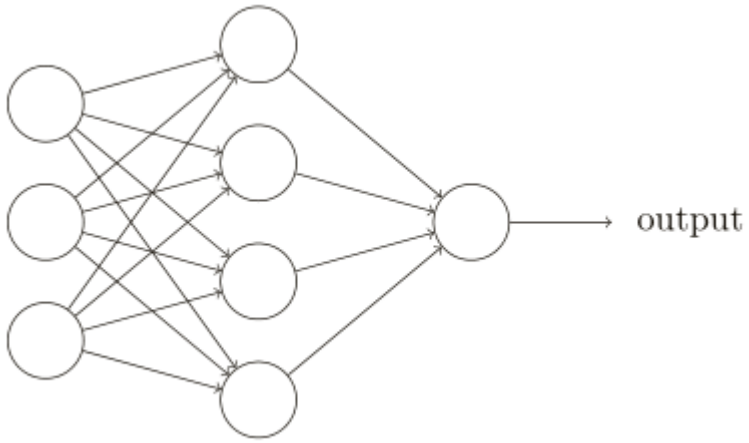

图 3.3 神经网络结构

在这个网络中（如图 3.3），左边的一层叫做输入层，右边的一层叫做输出层，中间的一层叫做隐藏层。输入层和输出层总是非常直观的，而且节点的数量非常自由。然而，隐藏层的设计十分具有工程性。有一些经验公式在实际应用中起到了很不错的作用，例如对于单隐藏层的神经网络，如果我们假设输入端节点是n，输出端节点是l，那么隐藏层的节点m可以是：

$$m = \sqrt{n+l} + \alpha, \alpha \leq 10 \tag{3}$$

$$m = \sqrt{nl} \tag{4}$$

事实上，最精准的方法是逐个实验，在初始网络的基础上对节点进行增加

和删除。然而，这种方法费时费力，每一次变化都要考验神经网络的性能和收敛性，丧失了神经网络方法本身的优势。有一种方法仍然是基于经验的，但是利用了多篇文献和最小二乘方法，有一定的合理性：

$$m = \sqrt{0.43nl + 0.12l^2 + 2.54n + 0.77l + 0.35} + 0.5 \quad (5)$$

我们仍然希望可以从理论的角度得到一个具有足够解释性的公式或方法，然而现有的理论并不能让人满意，所以神经网络的应用从某种角度讲仍然是非常有潜力的，而且依赖于设计人员的灵感和创造力。

3.1.1 普遍性定理

神经网络的拓扑结构和人脑的结构有相似之处，如果从认识论的角度出发，那么我们可以粗略的认为每一个隐藏层上，都包含着对于事物的一层认识，随着输入端到输出端，经历了多个隐藏层，即使没有足够的理论说明结论的来源，得到输出端的结果仍然包含了足够复杂的认知过程。

但是只有粗略的解释并不能说明为什么神经网络可以完成这些任务，所以普遍性定理告诉我们，神经网络可以计算任意一个连续函数。也就是说，随着隐藏层节点数量的增加，一个单层或者包含多个隐藏层的神经网络，作为一个映射可以逼近定义域上的任意连续函数。这个定理还有一个推广：可以利用利普希茨性质证明神经网络是通用拟合器，假设$f: [0,1]^n \to [0,1]$是ρ利普希茨函数，对于$\forall \varepsilon > 0$，可以构造神经网络$N: [0,1]^n \to [0,1]$，满足对于任意$\boldsymbol{x} \in [0,1]^n$有$|f(\boldsymbol{x}) - N(\boldsymbol{x})| \leq \varepsilon$。

定理的证明可以参考实变函数论中的富比尼定理的证明。首先证明神经网络可以构造一个符号函数，显而易见的是，由于感知器可以通过设计权重和偏置形成与门和与非门，两种门在同一层叠加可以得到任何定义域为闭区间，值域为布尔值的函数。然后在输出值中加上权重，可以得到符号函数。符号函数的线性和可以得到简单函数，所以通过增加隐藏层的节点个数，得到多个符号函数，就可以在输出端得到一个简单函数。而我们知道，任意连续函数都可以用一列简单函数逼近。由此我们可以知道，神经网络可以逼近连续函数。本文重点不在于此，所以详细过程不予给出。

这个定理保证了神经网络具有高度的适应性和广阔的应用空间，看似离散的结构，在权重和 S 型函数的加持下，只要有合适的高性能的学习算法，就可以解决各种实际问题。

3.1.2 反向传播算法

神经网络的拓扑结构保证了神经网络的潜能，发挥这些潜能还是要有训练网络的方法。我们的最终目标是找到合适的权重和偏置，使神经网络逼近需要的函数。这里，我们利用均方误差作为代价函数，假设a是训练数据的输出，g(x)代表神经网络，n 是样本容量，那么代价函数可以写作：

$$C(w,b) = \frac{1}{2n}\sum_{x}\left\|g(x) - a\right\|^2 \tag{6}$$

所以，构造神经网络的过程其实就转化为求 w 和 b，使代价函数最小。

反向传播算法的核心内容就是计算代价函数关于权重和偏置的表达式，表达式的背后是权重和偏置影响神经网络的方式。首先定义变量w_{jk}^{l}表示从第 *l-1* 层的第 k 个神经元传递到第 l 层的第 j 个神经元的权重，同样的，记b_j^l为第 l 层的第 j 个神经元的偏置，记a_j^l为第 l 层的第 j 个神经元的激活值。显然我们有以下公式：

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{7}$$

如果我们把这个公式写成向量形式，同样可以得到：

$$a^l = \sigma\left(w^l a^{l-1} + b^l\right) \tag{8}$$

还可以进行进一步的简化，设$z^l = w^l a^{l-1} + b^l$。我们称之为 l 层的带权输入。接下来为了方便，我们使用 Hadamard 乘积，也就是向量按元素乘积，符号记为 #，满足$(s\#t)_j = s_j t_j$。

引入中间变量，记为δ_j^l，作为第 L 层第 j 个神经元的计算误差。定义为：

$$\delta_j^l = \frac{\partial C}{\partial z_j^L} \tag{9}$$

那么反向传播算法的基本内容和结构由四个方程给出：

$$\delta^L = \nabla C \# \sigma'\left(z^L\right) \tag{10}$$

$$\delta^l = \left(\left(w^{l+1}\right)^T \delta^{l+1}\right) \# \sigma'\left(z^l\right) \tag{11}$$

$$\frac{\partial C}{\partial b_j} = \delta_j^l \tag{12}$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l \tag{13}$$

复合函数微分的链式法则可以用于证明上述方程，代价函数的梯度可以成功计算。具体过程参考论文文献。如果输入神经元激活值很低，或者输出神经元已经饱和了也就是拥有过高或者过低的激活值，学习过程会非常缓慢。这个结果其实也是意料之中的。然而关于神经网络学习的背后的思维模型就是以这些方程为基础的。而且，我们可以将这种推断方式进行拓广。四个基本方程也其实对任何的激活函数都是成立的，其实从证明中也可以看到，其实推断本身不依赖于任何具体的代价函数。所以，我们可以使用这些方程来设计有特定学习属性的激活函数。

3.2 算法的具体实现

3.2.1 代码结构

神经网络的反向传播算法让我们可以计算代价函数的梯度。根据上述思想，我们可以把代码结构写出来。整体算法记为 backpropagation。

1、输入x：为输入端设置对应的激活值a^1。
2、 前向传播：对于$l = 2,3,\dots,L$，计算$z^l = w^l a^{l-1} + b^l$和$a^l = \sigma(z^l)$。
3、 输出误差：计算向量$\delta^L = \nabla C \# \sigma'(z^L)$。
4、 反向传播：对于$l = L-1, L-2, \dots, 2$，计算$\delta^l = ((w^{l+1})^T \delta^{l+1}) \# \sigma'(z^l)$。
5、 输出：计算代价函数的梯度$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$，$\frac{\partial C}{\partial b_j^l} = \delta_j^l$。

这种反向移动其实是代价函数是网络输出的函数的结果。为了得到代价随前面层的权重和偏置变化的规律，我们需要重复利用微分的链式法则，反向地获得需要的表达式。

在实践中，通常将反向传播算法和诸如随机梯度下降这样的学习算法进行同时使用，我们会对许多训练样本计算对应的梯度。因为我们的目的是极小化代价

1、确定参数：迭代最大次数m，步长序列$\theta_1, \theta_2, \dots, \theta_m$，正则参数λ。
2、输入拓扑结构：建立分层图(V, E)，R 上的可微激活函数σ。
3、 初始化向量：随机选取$|E|$维向量ω_1。ω_1的分布要趋于 0。
4、 循环：对于$i = 1, \dots, m$，计算梯度$v_i = backpropagation(x, y, w, \sigma)$。赋值$w_{i+1} = w(i) - \theta_i(v_i + \lambda w_i)$。
5、 输出：取$\bar{w}$为空间中最优的w_i。

函数，那么随机梯度下降是一个迭代的优化策略，通过取沿着函数当前迭代点的负梯度方向的步长来提高界的精度。尽管我们不知道具体的分布，但是通过取一

个随机方向的步长，梯度下降法就可以取期望值来完成目的。例如给定一个训练样本，利用样本进行学习。

3.2.2 数据预处理

下面我们开始使用 Gene Expression Omnibus(GEO)的数据库进行数据分析。这里我们使用的数据集是GDS5627,研究对象是敏达沙替尼和抗达沙替尼细胞株。达沙替尼（Dasatinib），别名 DASA 锡 IB，是一种灰白至黄色固体的化学品。达沙替尼为抗肿瘤药，临床上主要治疗对甲磺酸伊马替尼耐药，或慢性髓细胞白血病成年患者。但 PC 细胞有一定可能具有内在或获得性的达沙替尼耐药性，可能使药物无效。研究与之相关的基因表达谱，就可以从基因层面上了解药物的作用基因，筛选试用药物人群等，具有广泛的应用价值。

数据集一共有 18 个样本，9 个敏达沙替尼细胞株，9 个抗达沙替尼细胞株（图 3.4）。检测平台是 Illumina 公司制作的微珠芯片，名称是 Illumina HumanHT-12 V4.0 expression beadchip。

Samples	Factors		
	cell line	cell type	
GSM1435684	Panc0403	dasatinib-sensitive pancreatic cancer cells	Panc0403 rep-1
GSM1435685	Panc0403	dasatinib-sensitive pancreatic cancer cells	Panc0403 rep-2
GSM1435686	Panc0403	dasatinib-sensitive pancreatic cancer cells	Panc0403 rep-3
GSM1435687	Panc0504	dasatinib-sensitive pancreatic cancer cells	Panc0504 rep-1
GSM1435688	Panc0504	dasatinib-sensitive pancreatic cancer cells	Panc0504 rep-2
GSM1435689	Panc0504	dasatinib-sensitive pancreatic cancer cells	Panc0504 rep-3
GSM1435690	Panc1005	dasatinib-sensitive pancreatic cancer cells	Panc1005 rep-1
GSM1435691	Panc1005	dasatinib-sensitive pancreatic cancer cells	Panc1005 rep-2
GSM1435692	Panc1005	dasatinib-sensitive pancreatic cancer cells	Panc1005 rep-3
GSM1435693	SU8686	dasatinib-resistant pancreatic cancer cells	SU8686 rep-1
GSM1435694	SU8686	dasatinib-resistant pancreatic cancer cells	SU8686 rep-2
GSM1435695	SU8686	dasatinib-resistant pancreatic cancer cells	SU8686 rep-3
GSM1435696	MiaPaCa2	dasatinib-resistant pancreatic cancer cells	MiaPaCa2 rep-1
GSM1435697	MiaPaCa2	dasatinib-resistant pancreatic cancer cells	MiaPaCa2 rep-2
GSM1435698	MiaPaCa2	dasatinib-resistant pancreatic cancer cells	MiaPaCa2 rep-3
GSM1435699	Panc1	dasatinib-resistant pancreatic cancer cells	Panc1 rep-1
GSM1435700	Panc1	dasatinib-resistant pancreatic cancer cells	Panc1 rep-2
GSM1435701	Panc1	dasatinib-resistant pancreatic cancer cells	Panc1 rep-3

图 3.4 数据集样本

DNA 芯片实验中的任意的差异性都可能会导致所得到的基因表达谱数据产生变化，由于从基因芯片的制各到从杂交图像中提取出基因表达数据有多个步骤，因此可能会有多种噪声来源：肿瘤组织内不同的细胞组成，样本组织中 mRNA 组分的不同，样本选择个体间的差异，样本制备手段和杂交手段的不同，以及不同

的基因芯片都可能会引起各种数据误差。其中最主要的差异体现在个体基因组间所造成的基因间的差别。针对该类数据必须进行一定的预处理再进行分析，才能获得真实反映生物信息的处理结果。

为了方便随后进行的数据分析工作，经过筛选的基因表达数据要进行调整，主要包括对数变换和归一化。本文对数据集进行了分位数归一化处理，以消除微阵列实验过程中混杂在变量中的噪声的影响，使变量之间具有可比性，发现基因间真正的联系。本文还对数据进行对数变换。变换的一个明显优点是可以从生物学意义上进行解释。对数变换减少了方差和平均值，使得表达的变化独立于其产生强度的位置，低强度值的倍数改变和高强度值的倍数改变具有可比性。进行对数变换同时与分布相关联。对数变换使数据分布具有对称性和近似正态分布。未经过对数变换的数据经常分布在非常宽的领域内，呈明显的偏态分布，对数变换之后就具有一定的对称性和正态性。在此基础上，可以使用多种方便的统计检验方法进行分析。

具体到操作层面，本文使用 Bioconductor 的 R 软件包 GEOquery 获取数据详细信息，然后利用 beadarray 进行分位数归一化和以 2 为底的对数变换。详细 R 代码见附录内容。Illumina 公司 beadarray 软件包提供数据质量评估和低数据量水平的分析。该软件包具有多个函数，规范了数据处理程序，方便而且有效。经过初步处理之后，数据可以导入到 R 软件的 GSE 类型的变量（如图 3.5）。

```
> show(gse)
$GSE59357_series_matrix.txt.gz
ExpressionSet (storageMode: lockedEnvironment)
assayData: 47323 features, 18 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1435684 GSM1435685 ... GSM1435701 (18 total)
  varLabels: title geo_accession ... data_row_count (32 total)
  varMetadata: labelDescription
featureData
  featureNames: ILMN_1343291 ILMN_1343295 ... NA.33865 (47323 total)
  fvarLabels: ID Species ... GB_ACC (30 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL10558
```

图 3.5 经过处理的数据

3.2.3 基于 BP 神经网络的分类器

在 2015 年，Chien W 等人在 molecule oncology 杂志上发表了一片论文，题目是 Activation of protein phosphatase 2A tumor suppressor as potential treatment of pancreatic cancer，这篇论文分析了 GDS5927 中的数

据，他们得出的结果显示，66 种激酶抑制剂抑制 14 人胰腺癌细胞株的增殖，这些细胞系表现出对激酶抑制剂和数据不同的灵敏度可以归纳为一个 heatmap。任何分类过程的第一步，都要首先分析各种属性的有效性并选出最有代表性的属性。因此在设计并利用分类器进行样本分类之前，要进行特征选择或提取。DNA 芯片

```
> eset
ExpressionSet (storageMode: lockedEnvironment)
assayData: 48107 features, 18 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1435684 GSM1435685 ... GSM1435701 (18 total)
  varLabels: sample cell.line cell.type description
  varMetadata: labelDescription
featureData
  featureNames: ILMN_1343048 ILMN_1343049 ... ILMN_3311190 (48107 total)
  fvarLabels: ID Gene title ... Platform_SEQUENCE (22 total)
  fvarMetadata: Column labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 25637283
Annotation:
```

数据的一大特点是样本量相对比较小，由于成本控制的考虑，但是影响因子很多，由于人体基因排列的多样性。因此直接运用提取数据是不现实的，效率很低，需要使用一些方法提取出我们需要的基因，为下一步分析做准备。

图 3.6 基因表达数据集

最简单的筛选方法当然是 t 检验。对于每一个基因，把样本分成两类进行检验，原假设是两类样本的基因表达水平的均值相同。把所有拒绝原假设的基因挑选出来作为差异表达基因。Bioconductor 中的 R 软件包 genefilter 包含了一个函数 fastT，可以快速计算所有基因的 t 检验，筛选出合适的基因，针对我们的 DNA 芯片数据，这种方法比起传统的 R 软件函数更为可靠稳定，效率也更高。其中一个参数 var.equal 是一个逻辑变量，指示是否将样本中的方差视为相等。如果“TRUE”，一个简单的 F 测试的均值在单向方差分析。如果“FALSE”， 使用 1951 年提出的韦尔奇近似的方法，俗称任意多个样品的情况下的推广的双样本韦尔奇检验。函数的返回结果是一列 p-value 向量，根据参数可以确定某一个基因是否被筛选进来。之后可以继续使用 SAM 方法，这种方法比 T 检验更加稳健，可以更好解决上述问题。

还有一种基因筛选方法是 genefilter 提供的 filterfun 方法。这个函数能

```
> f1<-kOverA(9,10)
> wh1<-genefilter(set1,filterfun(f1))
> sum(wh1)
[1] 4034
```

从整个数据集中挑选出区分度强的基因子类。这些基因中包含最多的特征信息，与敏感度分类的相关度也最高。换句话说，这些基因可以在不同的细胞株中具有不同的表达水平，而在同一种细胞株中，可能具有相似的表达水平。而且，同一

```
[1]  2264  3483  3595  3728  5356  6274  6715  7456  7759  8483  9065  9180
9234  9888 11666 12347 12471 12585 12619 12764 14174 15716 15799 15989 16968 17163
17244
[28] 17454 18047 18302 18778 18845 18875 19579 20135 20363 20614 21094 21946 23089
23281 23668 27356 27936 28702 29851 32172 32959 33073 33444 33512 33571 33850 36951
[55] 36981 37164 39020 39025 39076 39096 39229 40625 42870 42880 43013 43481 43769
45506 46412
```

种细胞株中的数据也不能有很大的起伏，方差要控制在一定范围。经过基因筛选之后的数据同样储存在 R 软件之中。经过两种不同方法筛选之后，还剩下 48000 多个基因中的 69 个基因。可能还有其他基因能够表征有关达沙替尼的特性，但是这些基因所包含的信息是最多的。

筛选完成之后，我们也不必直接开始人工神经网络的学习。首先观察得到的数据是否可以进行机器学习。箱线图可以看出数据的特点，对于数据的分布情况很好，近似正态分布，所以我们后面的分析就有了坚实的理论基础，而且也可以用一些常见普通的方法进行验证。

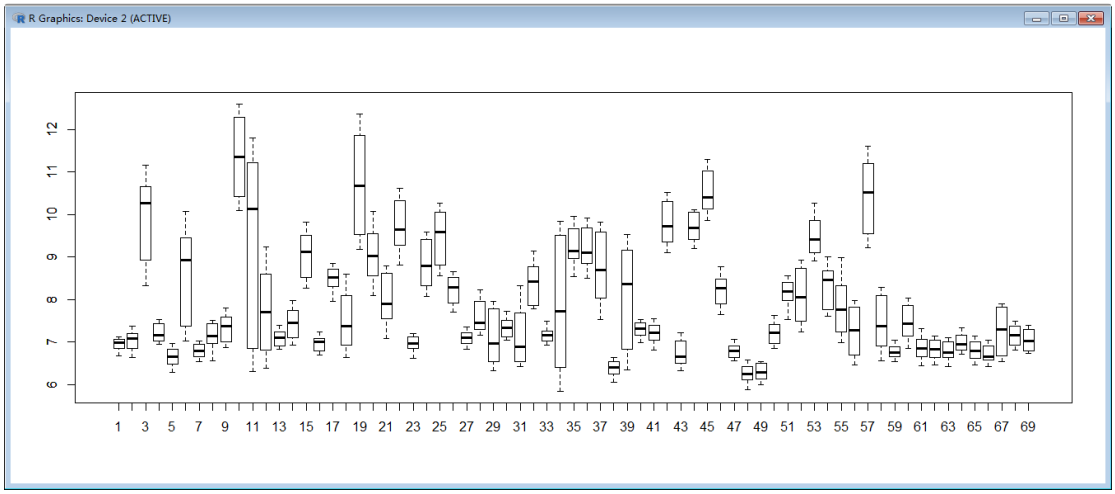

图 3.7 数据箱线图

之后我们就可以利用人工神经网络开始使 DNA 芯片数据的分类。为了建立神经网络分类器，我们首先提取敏达沙替尼细胞株样本和抗达沙替尼细胞株样本。将两个数据集分别分成两部分，一部分作为训练数据，另一部分作为检验数据。其中训练样本占到总样本量的三分之二。事实上，我们从敏达沙替尼样本中取出 GSM1435690 GSM1435691 GSM1435692 三个样本作为新数据，从抗达沙替尼样本中取出 GSM1435699 GSM1435700 GSM1435701 三个样本作为新数据，这样我们就可以利用其他样本建立分类器，用新数据检验分类器的正确性。下面我们使用

RSNNS 软件包（如图 3.7）中的函数 mlp 作为工具。RSNNS 是基于 SNNS 编写的一个 R 软件包。SNNS，也称斯图加特神经网络仿真器是一个函数库，包含神经网络的许多标准使用方法。这个包中丰富功能使它能在 R 中发挥很大的作用。这个函数创建多层感知器（MLP）并训练它。

Usage

```
mlp(x, ...)

## Default S3 method:
mlp(x, y, size = c(5), maxit = 100,
  initFunc = "Randomize_Weights", initFuncParams = c(-0.3, 0.3),
  learnFunc = "Std_Backpropagation", learnFuncParams = c(0.2, 0),
  updateFunc = "Topological_Order", updateFuncParams = c(0),
  hiddenActFunc = "Act_Logistic", shufflePatterns = TRUE, linOut = FALSE,
  outputActFunc = if (linOut) "Act_Identity" else "Act_Logistic",
  inputsTest = NULL, targetsTest = NULL, pruneFunc = NULL,
  pruneFuncParams = NULL, ...)
```

图 3.8 RSNNS 包函数 MLP 用法

MLP 的完全连接的前馈网络，也许是最常见的网络体系结构使用。训练通常是由误差反向传播或其他相关程序，有很多不同的学习功能。SNNS 让我们可以利用这些功能，目前 std_backpropagation 可以与很多函数一起使用，例如，backpropbatch，backpropchunk，backpropmomentum，backpropweightdecay，RPROP，Quickprop，SCG（SCG）。std_backpropagation，backpropbatch 等函数有两个重要的参数，学习率和最大输出差。学习率是一个介于 0.1 到 1 之间的值，代表了梯度下降的步长。最大输出差指的是目标函数与输出值的最大偏差，用于避免过度学习。

得到的结果是GSM1435684 GSM1435685 GSM1435686 GSM1435687 GSM1435688 GSM1435689 GSM1435690 GSM1435691 GSM1435692 这些样本是同一类，GSM1435693 GSM1435694 GSM1435695 GSM1435696 GSM1435697 GSM1435698 GSM1435699 GSM1435700 GSM1435701 这些样本是同一类。如果按照样本的特性

```
> mlp1
Class: mlp->rsnns
Number of inputs: 69
Number of outputs: 1
Maximal iterations: 100
Initialization function: Randomize_Weights
Initialization function parameters: -0.3 0.3
Learning function: Std_Backpropagation
Learning function parameters: 0.2 0
Update function:Topological_Order
Update function parameters: 0
Patterns are shuffled internally: TRUE
Compute error in every iteration: TRUE
Architecture Parameters:
$size
[1] 8

All members of model:
 [1] "nInputs"                "maxit"                  "initFunc"
"initFuncParams"        "learnFunc"              "learnFuncParams"
 [7] "updateFunc"             "updateFuncParams"       "shufflePatterns"
"computeIterativeError" "snnsObject"             "archParams"
[13] "IterativeFitError"     "fitted.values"          "nOutputs"
```

分类，则全部正确。本次采用的拓扑结构是一层隐藏层和多层隐藏层。其中一层隐藏层使用的是 8 个节点，因为由经验公式计算得出的一个值。经过多种不同拓扑结构和经验公式的组合，发现只需要 8 个节点就可以满足。其他的方法还可以有 11 个节点，10 个节点，9 个节点等。但是显而易见的是，只用八个节点可以提高分类器的效率。而一层拓扑结构已经满足我们的分类效果。两层隐藏层同样可以达到相同的效果，只是隐藏层节点增加的同时，训练时间大大加长。与之相比较，我们可以使用马氏距离对数据进行分类，结果得到的效果同样可以令人满意，当然使用距离函数十分快速。从我们分析数据的过程角度，筛选起到了重要的作用。其中建立的一个分类器特性如上所示，这个分类器的分类情况如表 3.1 所示。

表 3.1 使用上述神经网络分类器所做的结果

classification	First group	Second group
Dasatinib-sensitive cell line	3	0
Dasatinib-rejective cell line	0	3

从上述结果可以看出，人工神经网络方法起到了至关重要的作用。理论上讲，经过筛选的 69 个基因数量和论文中提到的 66 个差别不大。同时这些基因从另一个角度验证了论文中的观点，也就是这些基因表达强度对于达沙替尼的作用效果是有紧密联系的，虽然这些基因不一定包含了所有的论文中出现的影响因素，但是我们使用机器学习的方法恰恰不需要输入端的数据满足过多的条件。所以这个分类器的效果是非常理想的，包含了数据背后的生物学原理。虽然所有的方法都可以起到相似的效果，但是当样本量巨大的时候，使用距离分类的方法不能令人满意。尤其是相对与大数据量的 DNA 芯片数据来讲，运用人工神经网络分类器明显优于其他方法。另一个角度讲，人工神经网络可以通过增加训练样本的时候直接学习，这样可以避免后续处理过程和预测新数据时候的很多麻烦，比起简单的分类方法具有高效精准的独特优势，只可惜由于 DNA 芯片数据的稀有性，不可能拥有大量数据，在这种情况下我们可以做到尽量精确。

4 基本结论及应用方向

生物界中，有一条遗传信息在细胞内的生物大分子间转移的基本法则，就是中心法则。中心法则的内容是，遗传信息的标准流程大致可以这样描述：“DNA 制造 RNA，RNA 制造蛋白质，蛋白质反过来协助前两项流程，并协助 DNA 自我复制。”其中 RNA 是基因表达的重要环节，是 DNA 和蛋白质之间的桥梁。通过研究 mRNA，可以了解人类疾病与药物作用的内在机制，了解蛋白质影响人体的过程，了解人类遗传密码背后的意义。

本文通过研究两种不同的达沙替尼细胞株，筛选出差异表达基因，证实表达情况与细胞信使 RNA 的紧密联系。同时积极运用人工神经网络，理解了人工神经网络非线性和可扩展性的特点，建立模型。BP 神经网络作为一种相对成熟的机器学习技术，仍然有很强的灵活性，本文针对神经网络的拓扑结构和隐藏层节点个数，利用 DNA 芯片数据进行分析实验。得到的结果一方面证明人工神经网络具有强大的性能，比起一般的分类方法可以大大提高精确度。然而人工神经网络的性能和效率不能让人完全满意，这是一个可以进行进一步深入研究的领域。另一

方面，本文建立了基因水平的细胞株分类器，可以在已知遗传信息的基础上，分析和预测细胞株对于达沙替尼的敏感性和抗性，对于达沙替尼药物的运用具有一定的应用价值。

DNA 芯片技术的飞速发展不仅仅使得生物医学领域扩展到了一个新的领域，同时因为 DNA 芯片数据的多种特点，从不同的角度看，DNA 芯片技术越来越被看作是一个综合技术，与许多不同范畴的知识大范围交叉，例如计算机软件，应用数学，统计分析等。本文的创新点在于使用了多种不同的拓扑结构，隐藏层节点和学习算法，让人工神经网络在不同的情况下对于同一个数据集进行学习。最终得到满意的分析结果。

参考文献

[1] Andreas Zell, G unter Mamier, Michael Vogt, Niels Mache, Ralf Hubner, Sven Doring, Kai-Uwe Herrmann, Tobias Soyez, Michael Schmalzl, Tilman Sommer, Artemis Hatzigeorgiou, Dietmar Posselt Tobias Schreiner, Bernward Kett, Gianfranco Clemente Jens Wieland, Jurgen Gatter, Stuttgart Neural Network Simulator User Manual, Version 4.2[OL], UNIVERSITY OF STUTTGART.

[2] Beatriz A. Garro a, Katya Rodríguez a, Roberto A. Vázquez. Classification of DNA microarrays using artificial neural networks and ABC algorithm[J]. US: Applied Soft Computing. 2016.

[3] Christoph Bergmeir and José M. Benítez, Neural Networks in R using the Stuttgart Neural Network Simulator (SNNS)[OL], CRAN, 2016-12-16.

[4] D. Juan, O. Gran˜a, F. Pazos, P. Fariselli, R. Casadio, and A. Valencia, A Neural Network Approach to Evaluate Fold Recognition Results[J]. PROTEINS: Structure, Function, and Genetics, 2003, Vol.50: pp.600–608.

[5] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, Learning Representations by Back-propagating Errors[J], Nature, 1986 October, Vol.323: pp.533-534.

[6] Hieu Trung Huynh, Jung-Ja Kim and Yonggwan Won. Classification Study on DNA Microarray with Feedforward Neural Network Trained by Singular Value Decomposition[J]. US: International Journal of Bio-Science and Bio-Technology. 2006.

[7] Jagdish Chandra Patra, Lei Wang, Ee Luang Ang, and Narendra S. Chaudhari. Neural Network-Based Analysis of DNA Microarray Data[J]. US: Proceedings of International Joint Conference on Neural Networks. 2005.

[8] Kyung-Joong, Kim, Sung-Bae, Cho. Evolving Artificial Neural Networks for DNA Microarray Analysis[J]. US: IEEE. 2003.

[9] Leif E. Peterson, Mustafa Ozen, Halime Erdem, Andrew Amini, Lori Gomez, Colleen C. Nelson, and Michael Ittmann. Artificial Neural Network Analysis of DNA Microarray-based Prostate Cancer Recurrence[J]. US: IEEE. 2005.

[10] Mark Dunning, Analysis of Bead-level Data using beadarray[OL], Bioconductor, April 24, 2017.

[11] Mark Dunning, Analysis of Bead-summary Data using beadarray[OL], Bioconductor, April 24, 2017.

[12] Mike Smith, Image Analysis with beadarray[OL], Bioconductor, April 24, 2017.

[13] N. F. Britton, Xihong Lin, Hershel M. Safer, Maria Victoria Schneider, Mona Singh,

Anna Tramontano, Statistics and Data Analysis for Microarrays Using R and Bioconductor Second Edition[M], CRC Press, Taylor & Francis Group, 2012.

[14] Paolo Arena, Maide Bucolo, Luigi Furtuna, Luigi Occhipint. Cellular neural networks for real-time DNA microarray analysis[J]. US: MEMB. 2002.

[15] Shai Shalev-Shwartz, Shai Ben-David, Understanding Machine Learning from Theory to Algorithms[M], Cambridge University Press, 2014.

[16] 孙啸等，R 语言及 BIOCONDUCTOR 在基因组分析中的应用[M]，科学出版社，2006。

[17] 高山等，R 语言与 Bioconductor 生物信息学应用[M]，天津出版传媒集团，2014。

[18] 皮埃尔·巴尔迪，生物信息学——机器学习方法[M]，中信出版社，2003

[19] 李瑶，基因芯片数据分析与处理[M]，化学工业出版社，2006。

[20] 李银山，杨春燕，张伟，DNA 序列分类的神经网络方法[J]，计算机仿真，2003，第 20 卷第 2 期：pp.65-66。

[21] 桂江生，张青，基于神经网络的 DNA 序列分类的研究[J]，工业控制计算机，2015，第 28 卷第 10 期：pp.90-94。

[22] 敖丽敏，罗存金，基于神经网络集成的 DNA 序列分类方法研究[J]，计算机仿真，2012，第 29 卷第 6 期：pp.171-175。

附　录

附录 1

DNA 芯片数据集：

Series GSE59357	Query DataSets for GSE59357
Status	Public on Feb 24, 2015
Title	Gene expression profiles of dasatinib-resistant and dasatinib-sensitive pancreatic cancer cell lines
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Gene profiles from three dasatinib-resistant and three dasatinib-sensitive pancreatic cancer cell lines were compared by microarray analysis.
Overall design	RNA from three dasatinib-resistant (MiaPaCa2, Panc1, SU8686) and three dasatinib-sensitive (Panc0504, Panc0403, Panc1005) pancreatic cancer cell lines were extracted. Biological triplicates were employed for each cell line. Complementary DNA microarray analysis was performed using Illumina Human HT-12 v4 BeadChip (Illumina, San Diego, CA) at the National University of Singapore Core Facility following the manufacturer's instructions.
Contributor(s)	Chien W, Sun Q, Lee KL, Ding L, Wuensche P, Torres-Fernandez LA, Tan SZ,Tokatly I, Zaiden N, Poellinger L, Mori S, Yang H, Tyner JW, Koeffler HP
Citation(s)	Chien W, Sun QY, Lee KL, Ding LW et al. Activation of protein phosphatase 2A tumor suppressor as potential treatment of pancreatic cancer. *Mol Oncol*2015 Apr;9(4):889-905. PMID: 25637283
Submission date	Jul 11, 2014
Last update date	Feb 27, 2017
Contact name	Phillip H Koeffler
E-mail	phillip_koeffler@nuhs.edu.sg
Organization name	Cancer Science Institute
Lab	H. Phillip Koeffler's lab
Street address	14 Medical Drive
City	Singapore
ZIP/Postal code	117599
Country	Singapore
Platforms (1)	GPL10558 Illumina HumanHT-12 V4.0 expression beadchip

Samples (18)	GSM1435684	Panc0403 rep-1
More...	GSM1435685	Panc0403 rep-2
	GSM1435686	Panc0403 rep-3

Relations

BioProject	PRJNA255130

附录 2

基因筛选函数：

`filterfun`	*Creates a first FALSE exiting function from the list of filter functions it is given.*

Description

This function creates a function that takes a single argument. The filtering functions are bound in the environment of the returned function and are applied sequentially to the argument of the returned function. When the first filter function evaluates to FALSE the function returns FALSE otherwise it returns TRUE.

Usage

```
filterfun(...)
```

Arguments

...	Filtering functions.

Value

`filterfun` returns a function that takes a single argument. It binds the filter functions given to it in the environment of the returned function. These functions are applied sequentially (in the order they were given to `filterfun`). The function returns FALSE (and exits) when the first filter function returns FALSE otherwise it returns TRUE.

Author(s)

R. Gentleman

See Also

`genefilter`

Examples

```
set.seed(333)
x <- matrix(rnorm(100,2,1),nc=10)
cvfun <- cv(.5,2.5)
ffun <- filterfun(cvfun)
which <- genefilter(x, ffun)
```

附录 3

数据分析部分 R 代码：

```
#加载各种程序包
library(BiocGenerics)
library(parallel)
library(Biobase)
library(ggplot2)
library(beadarray)
library(GEOquery)
……
#数据预处理
Data(dData)
exprs(eData)[1:5,1:5]
eData.log2
eData.norm <- normaliseIllumina(eData.log2,
method="quantile", transform="none")
……
#神经网络分析
Gds<-getGEO("GDS5***", GSEMatrix=TRUE)
Eset<-GDS2eSet(gds)
F1<-kOverA(9, 7)
Ffun<-filterfun(f1)
Set1<-genefilter(exprs(eset), ffun)
F2<-kOverA(18, 6)
Ffun<-filterfun(f2)
Set2<-genefilter(exprs(eset), ffun)
Set3<-fastT(exprs(eset), 1:9, 10:18)
Set<-exprs(eset)[which(set3==TRUE),]
analyzeClassification(set, method = "WTA", l = 0, h = 0)
inputs<-t(set)
targets<-c(seq(0, 9), seq(1, 9))
artmap(set, nInputsTrain, nInputsTargets)
mlp<- mlp(inputs, targets, size = c(8), maxit = 100,
initFunc = "Randomize_Weights", initFuncParams = c(-0.3, 0.3),
learnFunc = "Std_Backpropagation", learnFuncParams = c(0.2, 0))
wm<-weightMatrix(mlp)
predictions<-predict(mlp, newdata)
Cm<- confusionMatrix(targets, predictions)
……
```

致　谢

论文即将完稿之际，心里涌起一丝激动和快慰。本研究及学位论文是在我的导师庞天晓教授的亲切关怀和悉心指导下完成的。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从课题的选择到项目的最终完成，庞老师都始终给予我细心的指导和不懈的支持。一个学期以来，他不仅在学业上给我以精心指导，同时还在思想上给我以无微不至的关怀，在此谨向庞老师致以诚挚的谢意和崇高的敬意。我还要感谢在一起愉快的度过本科生生活的数学科学学院的各位同门，正是由于你们的帮助和支持，我才能克服一个一个的困难和疑惑，直至本文的顺利完成。

毕 业 论 文（设计）考 核

一、指导教师对论文（设计）的评语：

人工神经网络是对人脑或自然神经网络若干基本特性的抽象和模拟。DNA 芯片技术通过微阵列技术将高密度 DNA 片段阵列以一定的排列方式使其附着在玻璃、尼龙等材料上面。在本毕业论文中，作者研究了人工神经网络在 DNA 芯片数据分析中的应用。

指导教师(签名)_____ 年 月 日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

秦臻同学论文研究人工神经网络在 DNA 芯片数据分析中的应用。全文主要内容由三部分组成：首先介绍研究背景和拟研究的问题；然后，介绍了 DNA 芯片数据的特点和获取；最后介绍了如何用人工神经网络进行 DNA 芯片的数据分析。 秦臻同学在答辩期间解释了选题的动机和主要目的，展示了主要结果， 并回答了答辩老师的提问。 答辩委员会认为该论文选题恰当，所获结果有一定的应用价值， 结构完整清楚，写作基本规范， 图表配置恰当，文献综述和引用符合要求。回答问题基本正确。一致认为秦臻同学的毕业论文达到本科学位论文要求。

成绩比例	文献综述占（10%）	开题报告占（20%）	外文翻译占（10%）	毕业论文（设计）质量及答辩占（60%）	总 评成 绩
分 值					

答辩小组负责人（签名）_____ 年 月 日