全国第七届研究生数学建模竞赛

题 目 **神经元的形态分类和识别**

摘 要：

本文主要利用支持向量机(SVM)理论、聚类分析理论、主成分分析与典型相关分析等有关知识，分析和解决了基于神经元空间几何形态特征的特征提取、样本分类、类间和类内样本形态特征的显著性分析，并根据分类的结果，给出了基于神经元形态特征的命名方案；最后建立了以神经元的几何拓扑空间结构为隐变量的隐马尔科夫模型(HMM)，从而预测神经元形态随时间的生长变化规律，给出了预测模型的实时性和有效性，得到了比较好的仿真结果。

问题一首先提取神经元的 43 种几何形态特征，然后采用主成分分析法，在这些形态特征中提取 6 种对神经元形态特征有较强解释力的几何特征，最后以此特征建立支持向量机分类器模型。

问题二首先主观标记附录 B 中神经元的类型；然后利用问题一中所建立的 SVM 分类器模型对其进行分类，并进行主客观的交叉验证，达到 90%的准确率。验证了所建立模型在解决非线性及高维模式识别中表现出极大优势。最后根据对分类结果中误分的神经元的几何特征的研究，建议引入新的神经元名称。

问题三考虑到待分类神经元的类别数未知，用层次聚类方法进行神经元的未知类型数目的聚类，得到谱系聚类图；然后根据 R^2 统计量的方法确定待分类神经元样本分为 8 类最合适；最后在已经确定类别数目的情况下，针对层次聚类算法的时间复杂度较高、容易陷入局部最优解等问题，采用 k 均值聚类算法确定每一类神经元具体样本，并给出了一种新的神经元命名方案。

问题四从分类的逆向思维出发，基于样本间的分离度准则，建立了一种不同物种神经元的区分方法，同时给出了不同物种同一神经元差异较大的形态特征，并利用统计中的典型相关分析给出了改进的模型。

问题五针对神经元的实际形态是随时间的流逝不断发生变化的问题，建立了以神经元的几何拓扑空间结构为隐变量的 HMM，其显状态为神经元的几何特征；利用已有的数据预测出神经元显式的状态，并预测神经元形态的生长变化规律。

关键词 神经元 几何特征 支持向量机 主成分分析 k 均值 隐马尔科夫模型

参赛队号 10701001

队员姓名 何立火 朱明敏 侯伟龙

参赛密码 _____

（由组委会填写）

中山大学承办

一. 问题重述

大脑是生物体内结构和功能最复杂的组织，其中包含上千亿个神经细胞(神经元)。人类脑计划(Human Brain Project, HBP)的目的是要对全世界的神经信息学数据库建立共同的标准，多学科整合分析大量数据，加速人类对脑的认识。

作为大脑构造的基本单位，神经元的结构和功能包含很多因素,其中神经元的几何形态特征和电学物理特性是两个重要方面。其中电学特性包含神经元不同的电位发放模式；几何形态特征主要包括神经元的空间构象，具体包含接受信息的树突，处理信息的胞体和传出信息的轴突三部分结构。由于树突，轴突的生长变化，神经元的几何形态千变万化。电学特性和空间形态等多个因素一起，综合表达神经元的信息传递功能。

对神经元特性的认识，最基本问题是神经元的分类。目前，关于神经元的简单分类法主要有：(1)根据突起的多少可将神经元分为多极神经元；双极神经元和单极神经元。(2)根据神经元的功能又可分为主神经元，感觉神经元，运动神经元和中间神经元等。主神经元的主要功能是输出神经回路的信息。例如大脑皮层的锥体神经元，小脑皮层中的普肯野神经元等。感觉神经元，它们接受刺激并将之转变为神经冲动。中间神经元，是介于感觉神经元与运动神经元之间起联络作用的。运动神经元，它们将中枢发出的冲动传导到肌肉等活动器官。不同组织位置，中间神经元的类别和形态，变化很大。动物越进化，中间神经元越多，构成的中枢神经系统的网络越复杂。

如何识别区分不同类别的神经元，这个问题目前科学上仍没有解决。生物解剖区别神经元主要通过几何形态和电位发放两个因素。神经元的几何形态主要通过染色技术得到，电位发放通过微电极穿刺胞内记录得到。利用神经元的电位发放模式区分神经元的类别比较复杂，主要涉及神经元的 Hodgkin-Huxley 模型和 Rall 电缆模型的离散形式(神经元的房室模型)。本问题只考虑神经元的几何形态，研究如何利用神经元的空间几何特征，通过数学建模给出神经元的一个空间形态分类方法，将神经元根据几何形态比较准确地分类识别。

神经元的空间几何形态的研究是人类脑计划中一个重要项目，NeuroMorpho.Org 包含大量神经元的几何形态数据等,现在仍然在不断增加，在那里你们可以得到大量的神经元空间形态数据，例如附录 A 和附录 C。对于神经元几何形态的特征研究这个热点问题，不同专家侧重用不同的指标去刻画神经元的形态特征，例如图 1 给出的神经元的粗略空间刻画以及附录 A 和附录 C 用标准的 A.SWC 格式给出的刻画。你们需要完成的任务是：

1. 利用附录 A 中和附录 C 样本神经元的空间几何数据，寻找出附录 C 中 5 类神经元的几何特征(中间神经元可以又细分 3 类)，给出一个神经元空间形态分类的方法。
2. 附录 B 另外有 20 个神经元形态数据，能否判定它们属于什么类型的神经元。在给出的数据中，是否有必要引入或定义新的神经元名称。
3. 神经元的形态复杂多样，神经元的识别分类问题至今仍没有解决，你们是否可以提出一个神经元分类方法，将所有神经元按几何特征分类。你们能否给生物学家为神经元的命名提出建议(附录 A 和附录 C 的神经元是比较重要的类别，实际应该有很多其他类别)。
4. 按照你们的神经元形态分类方法，能否确定在不同动物神经系统中同一类神经元的形态特征有区别吗？例如，附件 A 中有猪的普肯野神经元和鼠的普肯

野神经元，它们的特征有区别吗？

5. 神经元的实际形态是随着时间的流逝，树突和轴突不断地生长而发生变化的，你们能预测神经元形态的生长变化吗？这些形态变化对你们确定的几何形态特征有什么影响。

二. 基本假设

（1） 神经元细胞的生理学特征不做为分类的特征；
（2） 神经元胞体表面积按照球型近似计算；
（3） 神经元的房室的表面按照当前点为底面的圆柱的面积，并且忽略圆柱上下底面积；在计算整个神经元的所有房室的表面积时，忽略节点处的表面积；
（4） 假设每个节点的半径的测量值是无误差的；
（5） 假设每一个神经元的测量数据中，没有噪声点，或者说对这些点的测量都是准确值。

三. 符号定义

$X=\{x_1,x_2,\ldots,x_n\}$：待分类的神经元样本总体；

$F_p=\{f_1,f_2,\ldots,f_p\}$：神经元的 p 个形态特征集合；

x_{ij}：第 i 个神经元的 j 个形态特征；

β_i：第 i 个主成分的方差贡献率；

$\beta(k)$：前 k 个主成分的累积方差贡献率；

C_i：第 i 类神经元；

S：样本训练集；

$f_{ij}(x)$：决策函数；

$d(x_i,x_j)$：第 i 个神经元与第 j 个神经元之间的欧式距离；

$D^2(C_p,C_q)$：第 p 类神经元与第 q 类神经元类间距离平方；

$Com(C)$：类 C 内神经元样本紧密度；

$Sep(C_i,C_j)$：第 i 类神经元与第 j 类神经元类间分离度；

π：初始状态概率矩阵；

A：状态转移概率矩阵；

四． 问题的分析与求解

4.1 问题一

4.1.1 问题的分析

本题要求根据附件数据，提取出附录 C 中 5 类神经元的几何特征，然后以这 5 类神经元为标准，给出一种基于神经元空间几何特征的分类方法。根据题目所给的附录 A 和 C 中的 51 条数据分析可知，每一类神经元都有众多鲜明的形态学特征。例如神经元的胞体表面积，干的数目，分叉数目，分支数目，宽度，高度，深度，直径，长度，表面积，体积，树干锥度，分支幂律，分支角度或者其他形态参数等等。并且，利用附录中标准 SWC 文件所给的神经元形态数据可以容易地计算出每个神经元的具体的特征参数。

对于这些形态学分布有明显差异的神经元，在进行了大量的观测试验后，发现有若干种形态学特征对这些差异有较强的解释力，可以对具有不同形态学特征的神经元进行较好的区分。然而在随后的分类中，通过主观选择的特征向量对不同的形态学差异的解释并不尽如意人。因此，需要从众多特征中通过数学方法解析出影响神经元分类的主要因素。考虑到不同类神经元的几何特征差异很大，并且他们之间是相互独立的，基于此，本题采用主成分分析法进行特征提取，将所有特征参数作为变量，通过统计分析确定一组“主要特征元素”，这些主要“主要特征元素”能够很好的说明神经元的分类属性，从而达到特征提取的目的。

针对附录 C 中带有标签的 5 类神经元，即待分类目标的类别数已知，只需要根据这些已训练好的类别，找到能够描述它们之间差异的标准，这是一个多类分类问题，目前求解这类问题的方法非常多，考虑到待分类目标的类别数已知，基于此，将支持向量机的二分类方法推广到多类分类问题中，通过构造任意两类神经元的决策函数，将待分类问题转化为一个线性规划的求解问题，该算法计算量小，而且易于操作。

4.1.2 模型的建立

通过附录中提供的描述神经元形态学信息的标准 SWC 文件，共计算得到 43 种形态学特征，包括胞体表面积(Soma_Surface)、干的数目(N_stems)、分叉数目(N_bifs)、分支数目(N_branch)、末梢数目(N_tips)、宽度(Width)、高度(Height)、深度(Depth)、房室类型(Type)、直径(Diameter)、直径幂律(Diameter_pow)、长度(Length)、表面积(Surface)、截面积(SectionArea)、体积(Volume)、胞体到末梢的直线距离(EucDistance)、胞体到末梢的路径距离(PathDistance)、胞体到末梢的分支层数(Branch_Order)、所有房室可达的末梢的数目(Terminal_degree)、所有房室可达的末梢的最大数目(TerminalSegment)、相邻房室之间的锥度(Taper_1)、分支两端的直径之比(Taper_2)、分支路径长度(Branch_pathlength)、分支上的欧式距离与路径距离之比(Contraction)、分支上的房室个数(Fragmentation)、分叉口上较大直径与较小直径之比(Daughter_Ratio)、父支与子支的直径之比(Parent_Daughter_Ratio)、两分支之间的末梢数目的不对称性(Partition_asymmetry)、分支幂律(Rall_Power)、分支幂律之比(Pk)、幂为 1.5 的父子直径幂律之比(Pk_classic)、幂为 2 的父子直径幂律之比(Pk_2)、两分叉的近距离张角(Bif_ampl_local)、两分叉的远距离张角(Bif_ampl_remote)、前分叉与后分叉的近平面交角(Bif_tilt_local)、前分叉与后分叉的远平面交角(Bif_tilt_remote)、

前分叉与后分叉的近端截面平面交角(Bif_torque_local)、前分叉与后分叉的远端截面平面交角(Bif_torque_remote)、与末梢最近分叉的直径(Last_parent_diam)、与末梢最近第一分支的直径(Diam_threshold)、父分叉与子分叉的加权平均(HillmanThreshold)、分形维数(Hausdorff)、神经元的螺旋性(Helix)。这些特征分别就不同的侧重点描述了神经元的空间形态分布特征。

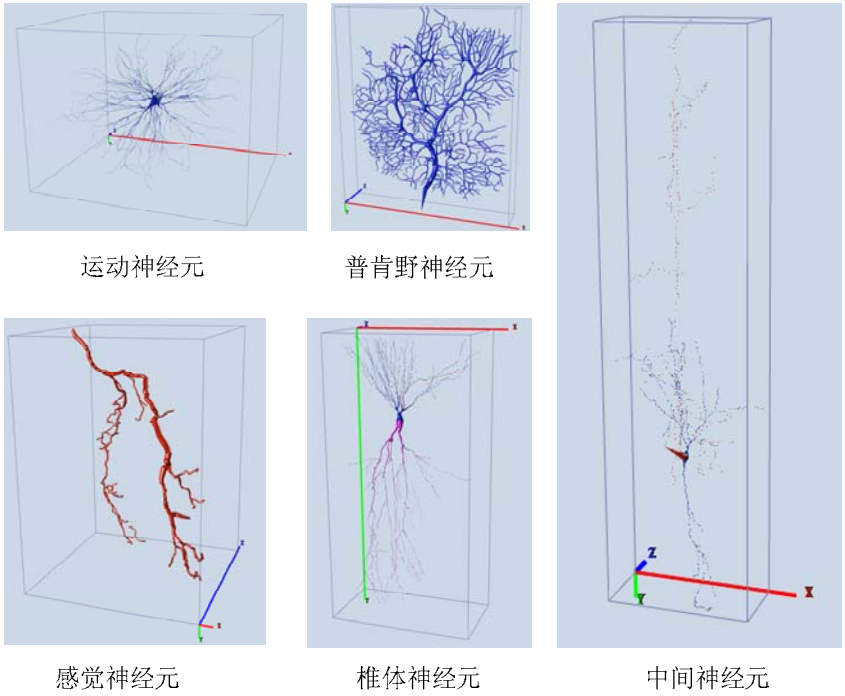


图 1 5 类不同神经元的三维视图

在附录中提供的 5 类重要神经元分类分别为运动神经元、普肯野神经元、椎体神经元、中间神经元和感觉神经元。如图 1 所示，这些不同类型的神经元在形态学特征上有着显著的差异。其中，运动神经元表现为关于胞体对称的发散型结构，分支多、密集；普肯野神经元在近似一个平面内延伸，类似于二维树状结构；椎体神经元表现为对称的锥体状发散；中间神经元表现为指向特定方向的发散结构，分支较少；而感觉神经元则表现为无胞体的树枝状结构。

我们主观地选定神经元的胞体表面积(Soma_Surface)，干的数目(N_stems)，分支数目(N_branch)，末梢数目(N_tips)，深度(Depth)，表面积(Surface)共 6 种特征作为用来区分神经元形态的依据，组成对神经元形态描述的特征向量 F：

$$F=\{f_1, f_2, f_3, f_4, f_5, f_6\}$$

其中 $f_i,(i=1,...,6)$ 分别对应 6 种形态学特征。

然而在随后的分类实验中，通过主观选择的特征向量对不同的形态学差异的解释并不尽如意人。而且，利用主观方法观察、选择形态学特征用于区别不同类别的神经元费时费力，因此，为了避免主观选择所带来的种种弊端，建立一种适当的数学模型自动地解析出影响神经元分类的主要因素便显得尤为重要。考虑到不同类神经元的几何特征差异很大，并且他们之间是相互独立的，基于此，本题采用主成分分析法进行特征提取。

假设 $X=\{x_1, x_2, \ldots, x_n\}$ 为待分类的神经元样本总体， $x_i=\{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ 表示第 i 个神经元， x_i 的 p 个分量 $x_{ij}(j=1,2,\ldots,p)$ 表示第 i 个神经元的 p 个形态特征，

由以上分析可知，神经元的 $p(p=43)$ 个形态特征构成了神经元的形态特征空间，为了对附录 A，C 中有限的数据进行准确的分类，我们首先对构成特征空间的数据 $(x_{ij})_{n\times p}$ 进行降维，即特征提取。由于神经元的几何特征都是相互独立的，因此，我们考虑用主成分分析法对神经元的形态特征进行特征提取。

首先对样本数据进行标准化，计算各指标的均值为：

$$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij},\quad j=1,2,...p \tag{4.1-1}$$

原始数据的协方差为：

$$S_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j) \tag{4.1-2}$$

标准化后的数据为：

$$\delta_{ij}=\frac{x_{ij}}{\overline{x}_j} \tag{4.1-3}$$

标准化后数据的协方差矩阵为：$V=(V_{ij})_{p\times p}$，则$V_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(\delta_{ki}-\overline{\delta}_i)(\delta_{kj}-\overline{\delta}_j)$，根据(4.1-3)式可知，

$$\begin{aligned}V_{ij}&=\frac{1}{n-1}\sum_{k=1}^{n}(\delta_{ki}-\overline{\delta}_i)(\delta_{kj}-\overline{\delta}_j) \\ &=\frac{1}{n-1}\sum_{k=1}^{n}(\frac{x_{ki}-\overline{x}_i}{\overline{x}_i})(\frac{x_{kj}-\overline{x}_j}{\overline{x}_j}) \\ &=\frac{S_{ij}}{\overline{x}_i\overline{x}_j}\end{aligned} \tag{4.1-4}$$

取 $i=j$，得：

$$V_{ii}=\frac{S_{ii}}{\overline{x}_i\overline{x}_i}=(\frac{\sqrt{S_{ii}}}{\overline{x}_i})^2 \tag{4.1-5}$$

由(4.1-5)式可知，均值化后的数据的协方差矩阵$V=(V_{ij})_{p\times p}$的主对角线元素是各指标变异系数平房，因此我们取V_{ii}作为各指标变异系数的差异。

具体算法流程如下：

步骤 1. 对原始数据进行均值化处理；

步骤 2. 计算均值化后 p 个指标的协方差矩阵；

步骤 3. 计算均值化后的 p 个指标的协方差矩阵 V 的特征根 $\lambda_1\geq\lambda_2\geq...\geq\lambda_p\geq 0$ 及相应的标准化正交特征向量 $(\alpha_1,\alpha_2,...,\alpha_p)$ 其中 $\alpha_i=(\alpha_{i1},\alpha_{ip},...,\alpha_{ip})',\ (i=1,2,...,p)$。$\lambda_i$ 反映了第 i 个指标在描述被评价神

经元上所起到的作用大小；

步骤 4. 计算主成分 y_i 的方差贡献率 $\beta_i = \lambda_i \Big/ \sum_{j=1}^{p} \lambda_j$ 及累积方差贡献率 $\beta(k) = \sum_{i=1}^{k} \left(\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j \right)$，其中 β_i 表示第 i 个主成分提取的原始 p 个指标的信息量，$\beta(k)$ 越大，前 k 个指标包含的原始信息越多；

步骤 5. 选取主要指标并进行综合评价：按累积贡献率不低于 99.9%的阈值确定 k 个指标，并计算综合评价指标 $T = a_1 y_1 + a_2 y_2 + ... + a_k y_k$. 其中 $y_i = \alpha_{i1}\delta_1 + \alpha_{i2}\delta_2 + ... + \alpha_{ip}\delta_p, \quad (i = 1, ..., k)$ 为第 i 个主成分，$\delta_j = (\delta_{1j}, \delta_{2j}, ..., \delta_{pj})'$ 为标准化后的数据。

4.1.3 模型的求解与结果分析

通过计算，得到神经元前 10 个特征的贡献率表，如表 1 所示：

表 1 神经元特征贡献率

序号	贡献率	累积贡献率
1	79.7409%	79.7409%
2	16.6950%	96.4359%
3	2.1877%	98.6236%
4	0.8566%	99.4802%
5	0.4136%	99.8938%
6	0.0482%	99.9420%
7	0.0333%	99.9753%
8	0.0174%	99.9927%
9	0.0038%	99.9964%
10	0.0022%	99.9987%

由表 1 可知，前 6 个主成分共同解释了总变异的 99.9753%，因此用着 6 个主成分代替原有的 43 个特征变量做系统分类识别。

表 2 附录 C 中 5 类神经元的主要特征

神经元类名称	特征
运动神经元	关于胞体对称的发散型结构，分支多、密集
普肯野神经元	在近似一个平面内延伸，类似于二维树状结构
椎体神经元	对称的锥体状发散
中间神经元	指向特定方向的发散结构，分支较少
感觉神经元	无胞体的树枝状结构

4.2 问题二

4.2.1 问题的分析

本题首先要求对未分类神经元依据其形态学特征进行分类，其次根据所得结果，讨论是否有必要引入或定义新的神经元名称。鉴于在问题一中，我们已经提取了若干神经元的形态学特征并分析了其对分类的影响。因此只需选择合适的分类方法依据问题一中所得到的神经元形态学特征进行分类。鉴于此分类问题的样本集较小，我们选择了 Corinna Cortes 和 Vapnik8 等于 1995 年提出的支持向量机作为神经元形态的分类器，它在解决小样本、非线性及高维模式识别中表现出许多特有优势。利用支持向量机，我们根据有限的神经元形态样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以期获得最好的推广分类能力。最后，根据分类结果的准确性分析误分的神经元形态学特征，探索其在形态学上所具有的特殊性质，并探讨是否有必要定义新的神经元名称对其进行描述。

4.2.2 模型的建立

假设给定的训练集 $S=\{(x_1,y_1),(x_2,y_2),...,(x_\ell,y_\ell)\}$，$S\in\{X\times Y\}^\ell$（其中 $x_i\in X\in\mathbb{R}^n, y_i\in Y=\{C_1,C_2,...,C_k\}, i=1,...,\ell$）。我们需要得到一个决策函数 $f(x)$ 来推断任一模式的 x 相对应的 y 值，采取“一对一”结构的算法，即对任意的一个类别对 $(C_i,C_j)\in Y\times Y$，希望建立一个决策函数 $f_{ij}(x)$，使其能够将 C_i,C_j 区分开来。

不失一般性，假设 $x_i, i=1,...,p$ 属于 C_i 类，并且它们的标号为+1；训练点 $x_i, i=p+1,...,\rho$ 属于 C_j 类，并且将他们标号为-1，即

$$\begin{aligned} S&=\{(x_1,y_1),...(x_p,y_p),(x_{p+1},y_{p+1}),...,(x_\ell,y_\ell)\}\\ &=\{(A_1,1),...,(A_p,1),(B_1,-1),...,(B_{\ell-p},-1)\}\end{aligned} \tag{4.2-1}$$

其中 $A_i=(a_{i1},...,a_{in})^T\in\mathbb{R}^n, i=1,...,p$，$B_i=(b_{i1},...,b_{in})^T\in\mathbb{R}^n, i=1,...\ell-p$.

设划分超平面为

$$(w\cdot x)+b=0 \tag{4.2-2}$$

其中 $w\in\mathbb{R}^n$，$b\in\mathbb{R}$,则所得分类问题的决策函数为：

$$f_{ij}(x)=sng((w\cdot x)+b) \tag{4.2-3}$$

记 $A=(A_1,...,A_p)^T\in R^{p\times n}$，$B=(B_1,...,B_{\ell-p})^T\in R^{(\ell-p)\times n}$，则寻找能正确划分所有训练集中的样本点的超平面，等价于寻找超平面(4.2-2)，其中的 w 和 b 满足：

$$-Aw-eb+e\le 0, Bw+eb+e\le 0 \tag{4.2-4}$$

其中 e 是分量全为 1 的向量。

由于实际问题可能是线性不可分的，即就是训练集中有被错划的样本点，因此，需要对以上模型进行修正，对任意的 n 维向量 $x=([x]_1,...,[x]_n)^T$，定义相应的 n 维向量 x_+

$$x_+=(\max(0,[x]_1),...,\max(0,[x]_n))^T \tag{4.2-5}$$

则正类样本点和负类样本点平均被错划的程度可以分别用如下两个量来度量

$$\frac{1}{p}\left\|(-Aw-eb+e)_+\right\|_1,\frac{1}{\ell-p}\left\|(Bw+eb+e)_+\right\|_1 \tag{4.2-6}$$

其中 $\|\cdot\|_1$ 为 1 范数。

为了使被错划的程度降到最低，从而有

$$\min_{w,b} g(w,b)=\frac{1}{p}\left\|(-Aw-eb+e)_+\right\|_1+\frac{1}{\ell-p}\left\|(Bw+eb+e)_+\right\|_1 \tag{4.2-7}$$

由(4.2-7)式得到如下的线性规划问题

$$\min_{w,b,y,z}\frac{e^Ty}{p}+\frac{e^Tz}{\ell-p} \tag{4.2-8}$$

$$s.t.\quad -Aw-eb+e\le y \tag{4.2-9}$$

$$Bw+eb+e\le z \tag{4.2-10}$$

$$y\ge 0,z\ge 0 \tag{4.2-11}$$

设 (w^*,b^*,y^*,z^*) 是(4.2-8)-(4.2-11)的最优解，则决策函数如下：

$$f_{ij}(x)=\operatorname{sgn}((w^*\cdot x)+b^*) \tag{4.2-12}$$

在 k 类多分类问题中，考虑其中任意的两类 (C_i,C_j)，都可以构造一个决策函数 $f_{ij}(\cdot)$ 把相应的两类 C_i,C_j 分开，这样共得到 $k(k-1)/2$ 个决策函数。对于一个新的训练点 x_k 可以得到 $k(k-1)/2$ 个输出，我们采用以下的投票方式判断它所属的类别：当 $f_{ij}(x_k)=+1$ 时，表明 x_k 属于 C_i 类，就在 C_i 类上投上一票，其他类为 0 票；当 $f_{ij}(x_k)=-1$ 时，表明 x_k 属于 C_j 类，就在 C_j 类上投上一票，其他类为 0 票。当我们检验完所有的 $k(k-1)/2$ 个输出后，会得到每一类的所有票。最后，x_k 属于获得票最多的那一类。具体算法描述如下：

步骤 1. 设已知训练集： $S=\{(x_1,y_1),(x_2,y_2),...,(x_\ell,y_\ell)\}$，$S\in\{X\times Y\}^\ell$，其中 $x_1,x_2,\ldots,x_\ell$ 表示待分类的样本数据，$\{y_1,y_2,\ldots,y_\ell\}\in\{-1,1\}^\ell$ 样本类别；

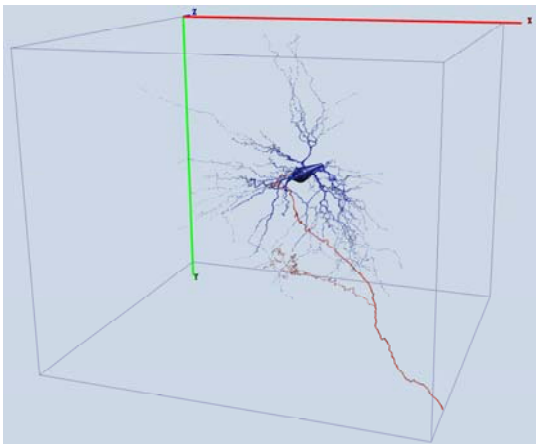
步骤 2. 在 K 类问题中任取一个类别对 (C_i,C_j)，按照式(4.2-8)-(4.2-11)求解关于每个类别对的线性规划问题，得到 w^*_{ij} 和 b^*_{ij}；

步骤 3. 关于每个类别对构造决策函数 $f_{ij}(x)$，共得到 $k(k-1)/2$ 个决策函数；

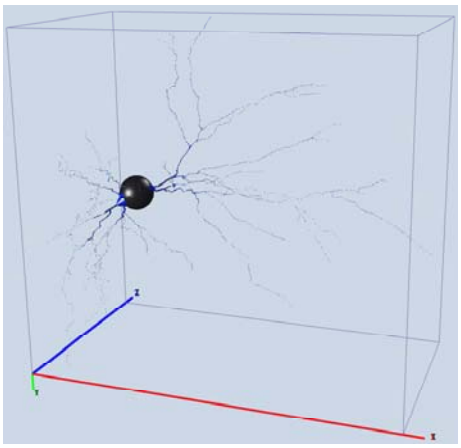
步骤 4. 对于一个新的训练点 x_k，可以得到 $k(k-1)/2$ 个函数值，然后按照投票规则，x_k 属于获得票最多的那一类。

4.2.3 模型的求解与结果分析

利用附录 A 与 C 中的 51 个神经元数据，利用第一问中所抽取的特征作为训练集，然后对附录 B 中的 20 个数据进行分类，其分类结果如表 3 所示。



(19)



(20)

图 2 新类型神经元三维图示

表 3 中，同时给出了我们对附录 B 中神经元依据主观判断给出的分类结果(类别为：1：运动神经元；2 中间神经元；3 普肯野神经元；4 椎体神经元；5 感觉神经元)。可以看出，支持向量机的分类结果与主观判断十分吻合，准确率达到 90%。这也从另一方面证明了所选取特征对形态学区分的有效性。

图 2 显示了误判的两个神经元的空间结构，由图可以看出，其形态与训练集中 5 大类神经元并不完全相似，反而展示出一些特殊的形态学特征了。这也许是支持向量机发生误判的原因之一。因此，针对这种形态上与原有 5 类均不相似的神经元，我们有必要引入并且定义新的神经元类型。

表 3 附件 B 中神经元的分类结果

序号	SVM 分类结果	主观分类结果
1	4	4
2	4	4
3	4	4
4	4	4
5	3	3
6	3	3
7	1	1
8	1	1
9	1	1
10	1	1
11	1	1
12	1	1
13	5	5
14	5	5
15	2	2
16	2	2
17	5	5
18	5	5
19	4	1
20	4	1

4.3 问题三

4.3.1 问题的分析

本问是一个按照神经元几何特征的分类问题，并且是不知道神经元需要分成多少类，这是本问的重点也是难点所在。神经元的形态复杂多样，提取不同的特征，可能分类的结果就不太一样，即使提取的特征一样，所选取的方法不同也可能出现不一致的分类结果。

针对这样一个问题，以及对问题一和问题二的认真分析之后，提取能够反映同类神经元几何特征的指标；以此为基础，用层次聚类方法进行神经元的未知类型数目的聚类，得到谱系聚类图；最后根据 R^2 统计量的方法确定分类个数，得到所有神经元按照几何特征的无监督分类结果。

由于神经元的类别数目是未知的，所以运用层次的无监督的分类方法。这样能较为有效的解决神经元的分类问题，甚至是可以得到一种新的分类方法或者分类模式，这是因为它的分类完全是根据其内在的几何特征之间的相似性得到的分类结果。得到的分类结果，以及参考附录 A 和附录 C 的分类类别，可以找到新的神经元类型，并且可以建立新的分类体系。

4.3.2 模型的建立

由于本题待聚类对象的类别数目未知，因此，首先考虑用层次聚类法来确定分类对象的类别数目，然后利用 k 均值聚类算法给出每个类别中所包含的具体神经元个体。根据附录 A，B，C 中的数据分析，假设有 n 个数据的观测值，用

$X=\{x_1,x_2,\cdots,x_n\}$表示所有的观测数据的集合，即待聚类分析的对象的全体。每一个x_j对象包含p个参数值$x_{j1},x_{j2},\ldots,x_{jp}$，分别为神经元的胞体表面积、干的数目、分叉数目、分支数目、宽度和高度等。层次聚类的基本思想是：首先定义样本间的距离和类与类之间的距离；初始将n个神经元数据看成n类(每一类包含一个神经元样本)，这时类间的距离与神经元样本间的距离是等价的；然后将距离最近的两类合并成新类，并计算新类与其他类的类间距离，再按最小距离准则并类；这样每次缩小一类，直到所有的数据都并成一类为止。基于以上思想，我们首先给出样本间的距离公式：

$$d(x_i,x_j)=(\sum_{k=1}^{p}(x_{ik}-x_{jk})^2)^{1/2} \tag{4.3-1}$$

利用离差平方和法定义类间的距离公式：

$$D^2(C_p,C_q)=\frac{|C_p|\cdot|C_q|}{|C_p|+|C_q|}d(\overline{x_p},\overline{x_q}) \tag{4.3-2}$$

其中C_p,C_q分别表示第p类和第q类神经元，$|C_p|$表示第p类神经元包含的样本个数，$\overline{x_p},\overline{x_q}$分别为第$p$类和第$q$类神经元的重心.

下面给出算法的具体步骤：

步骤 1. 数据变换：为了便于比较和计算，我们首先利用如下公式进行数据的标准化处理：$\delta_{ij}=x_{ij}-\overline{x}_j$，$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$j=1,2,...p$；

步骤 2. 计算n个样本两两间的距离，得样本间的距离矩阵$D^{(0)}$；

步骤 3. 初始(第一步：$i=1$)n个样品各自构成一类，类的个数$k=n$，即第t类$C_t=\{x_t\}(t=1,\ldots,n)$，此时类间的距离就是样本间的距离(即$D^{(1)}=D^{(0)}$)，然后对步骤$i=2,\ldots,n$执行并类过程的步骤 4 和步骤 5；

步骤 4. 对步骤 i 得到的距离矩阵$D^{(i-1)}$，合并类间距离最小的两类为新一类，此时，类的总个数减少一类，即$k=n-i+1$。

步骤 5. 计算新类与其他类的距离，得新的距离矩阵$D^{(i)}$。若合并后类的总个数 k 仍大于 1，重复步骤 4 和 5，直到类的总个数为 1 时转到步骤 6；

步骤 6. 画谱系聚类图；

步骤 7. 决定分类的个数及各类成员。

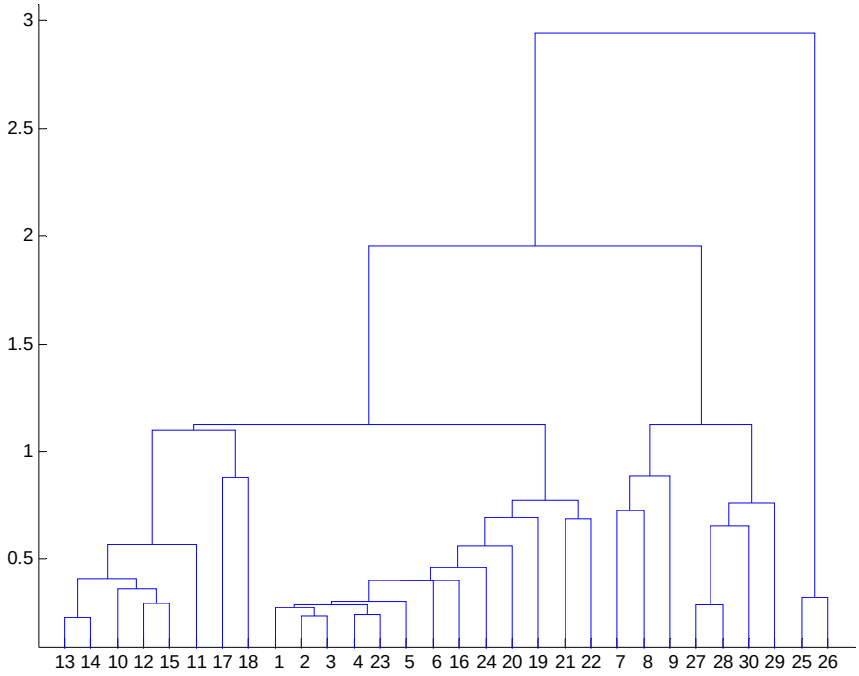


图 3 谱系聚类图

根据层次聚类法得到分类的谱系聚类图形如图 3，下面利用 R^2 统计量确定层次聚类算法分类的个数。假定已将 n 个样本分为 k 类，记为 $C_1, C_2, ..., C_k$，n_t 表示 C_t 类的样本个数 $(n_1 + ... + n_k = n)$，$\bar{X}^{(t)}$ 表示 C_t 的重心。$X_{(i)}^{(t)}$ 表示 C_t 中第 i 个样本 $(i = 1, ..., n_t)$，$\bar{X}$ 表示所有样本的重心，则 C_t 类中 n_t 个样本的离差平方和为：

$$W_t = \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X})'(X_{(i)}^{(t)} - \bar{X}^{(t)}) \tag{4.3-3}$$

其中 $X_{(i)}^{(t)}$，$\bar{X}^{(t)}$ 和 $\bar{X}$ 均为 p 维向量，W_t 为一数值；所有样品的总离差平方和为：

$$T = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X}^{(t)})'(X_{(i)}^{(t)} - \bar{X}^{(t)}) \tag{4.3-4}$$

T 又可分解为：

$$T = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X}^{(t)} + \bar{X}^{(t)} - \bar{X})'(...) \tag{4.3-5}$$

$$= \sum_{t=1}^{k} W_t + \sum_{t=1}^{k} n_t (\bar{X}^{(t)} - \bar{X})'(\bar{X}^{(t)} - \bar{X})$$

$$= P_k + B_k$$

令 $R_k^2 = B_k / T = 1 - P_k / T$，则 R_k^2 值越大，也就是 B_k / T 越大，表示 k 个类的类间偏差平方和的总和 B_k 在总离差平方和 T 中占的比例越大，这说明 k 个类越能

够区分开。因此R_k^2统计量可用于评价合并为k个类时的聚类效果。R_k^2越大，聚类效果越好。又因为R_k^2的值总是在 0 和 1 之间，当n个样品各自为不同的类时$(T=B_n)$，$R_k^2=1$；当n个样品最后合并成同一类时$(T=P_n)$，$R_k^2=0$，而且R_k^2的值总是随着分类个数k的减少而变小。因此我们可以通过分析R_k^2值的变化来确定n个样品应分为几类更合适。

表 4 不同特征形态下的R^2

R_1^2	R_2^2	R_3^2	R_4^2	R_5^2	R_6^2	R_7^2	R_8^2	R_9^2
0	0.308496	0.408776	0.428386	0.516017	0.643851	0.655322	0.950740	0.951418

根据上表可知，在分为 8 类之前的并类过程中R_k^2的值减少是逐渐的，改变不大；并且分为 8 类时的$R_8^2=0.950740$，而下一次合并后分为 7 类时$R_7^2=0.655322$的值下降较多，因此通过分析R_k^2统计量的变化可得出，分为 8 个类时比较合适的。

然而层次聚类算法的时间复杂度较高，当样本量很大时，不容易得到精确的聚类结果；而且在构建分类的谱系聚类图时，已被合并的向量不再参与以后的分类，容易陷入局部最优解；同时，如果一旦某一步做出了错误的合并决定，由于层次聚类法每步所做的处理不能被撤销，类与类之间也不能交换对象，因而这些错误会在以后的凝聚过程中叠加，会导致低质量的聚类结果。

基于以上的考虑，下面我们采用 k 均值聚类算法在已经确定类别数目的情况下，对神经元样本进行再次分类。分析论域X中n个样本所对应的模式相似性，按照样本间的亲疏关系把X划分成多个不相交的子集$C_1,C_2,...,C_k$，并求每类的聚类中心，设为Z_i，选择第i类C_i中的向量x_k与相应聚类中心Z_i之间的度量为欧几里德距离，建立 k 均值聚类模型：

$$J=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\Big(\sum_{k,X_k\in C_i}\|x_k-Z_i\|^2\Big) \tag{4.3-6}$$

这里$J_i=\sum_{k,X_k\in C_i}\|x_k-Z_i\|^2$是类$C_i$内目标函数，$J_i$的值依赖于$C_i$的几何形状和$Z_i$的位置。显然，$J$的值越小，表明聚类效果越好。下面给出 K 均值聚类算法的具体描述：

步骤 1. 随机选取k个向量作为每类的中心；

步骤 2. 设U是一个$k\times n$的二维隶属矩阵，如果第j个向量x_j属于类i，则U中的元素u_{ij}为 1；否则，该元素取 0，即

$$u_{ij}=\begin{cases}1,\forall k\neq i,\|x_j-Z_i\|^2\le\|x_j-Z_k\|^2,\\0,other.\end{cases}$$

步骤 3. 根据u_{ij}计算目标函数式(1)的值，如果它低于一个给定的最小阈值或者连

续两次值之差小于一个参数阈值则停止；

步骤 4. 根据 u_{ij} 更新各个聚类中心：$Z_i = \frac{1}{|C_i|}\sum_{k, X_k \in C_i} X_k$，这里 $|C_i| = \sum_{j=1}^{n} u_{ij}$ 表示类 C_i 内元素个数，然后重新回到步骤 2。

4.3.3 模型的求解与结果分析

根据前面所给的算法，通过计算机仿真，在对附录中 A，B，C 中的 71 条神经元数据进行标准化处理后，利用 K 均值算法共得到 8 类，并取其前最重要的二维特征画出示意图(图 4)。

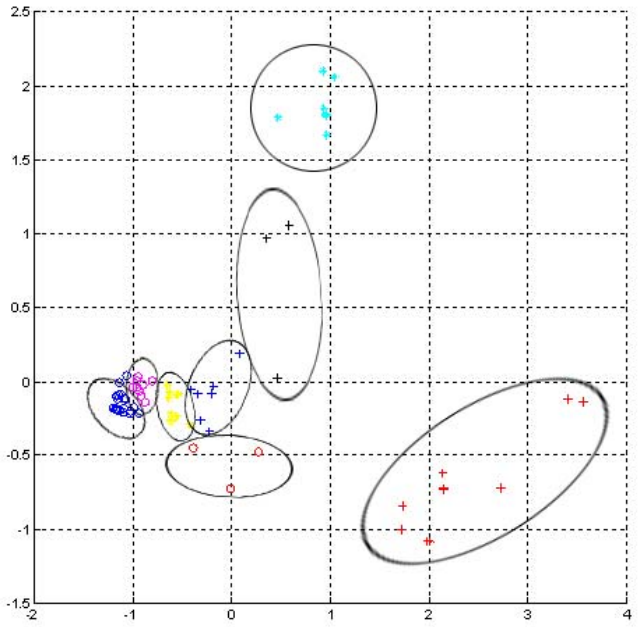


图 4 k 均值聚类结果

结果分析：

由图 4 可知，分类结果与实际情况相符；整体上来说，应用该模型得到的分类结果比较准确，各类的聚集点都比较集中，特别是运动神经元(青色)，双极神经元(黑色)，浦肯野神经元(红色+)，椎体神经元(蓝色)，感觉神经元(红色圈)这五种神经元的几何形态特征差异比较大，因此得到的结果比较准确，从仿真结果来看，所建立的模型具有很好的性能。

此外，由于对数据的正确分类与否直接影响到后续问题结果的正确性，而想要正确分类，如何选取聚类中心的初始位置又起着至关重要的作用，距离阈值决定了聚类的数目，阈值选得太小，聚类太多；阈值选得太大，聚类太少，如果能用事先知道的模式类分布的先验知识指导阈值的选择，可以得到满意的结果，否则只能选择一些大小不同的阈值进行多次聚类、仿真试探。因此从实验结果来看，也证明了通过多次聚类、仿真试探选取的阈值是合理的。表 5 给出了新的神经元命名规则。

表 5 神经元新命名规则和新命名

类别	新命名	样本神经元	主要特征
1	网状型	Guinea-pig Purkinje cell (1,2,3), Purkinje cell (1,2,3), purkinje neuron-A, 附录 B (5,6,7,8,9)	分支多，整体呈平面网状
2	沙漏型	Bipolar interneuron 3, pyracidal cell (2,3), pyramidal neuron-A, 附录 B (1,3,4)	分支呈锥状沙漏型伸展
3	折线运动型	附录 B (11,12,19)	分支较多而细，分叉角度大
4	均匀丝状型	Bipolar interneuron (1,2,4), Bipolar interneuron-A, pyracidal cell (1,4,5,6,7), 附录 B (15,20)	分支粗细均匀且较长
5	直线运动型	Motoneuron (1,2,3,4,5,6), motor neuron-A, 附录 B (10)	分支较多而细，分叉角度小
6	均匀粗线型	Multipolar cell (1,3,4,6,8,9), multipolar interneuron-A, Tripolar interneuron 2, 附录 B (16)	分支粗细均匀，其干较少
7	稀疏平滑型	Sensory neuron (1,2,3,4,5,6,7), sensory neuron-A, Tripolar interneuron (1,3,4), Tripolar interneuron-A, Multipolar cell (2,5,7) P44-DEV191.CNG, Bipolar interneuron 5, 附录 B (13,14)	分支较少，主干比较平滑
8	稀疏粗糙型	附录 B (17,18,2)	分支较少，主干上有球状隆起

4.3.4 模型的改进

聚类的目的是给出数据的最优划分，所谓最优就是，每个类中的数据相似性最大，而不同类中的数据最大程度的不相似。令 $C=\{C_1,C_2,\ldots,C_k\}$ 是数据集的一组聚类，聚类质量主要体现为聚类内的紧密度和聚类间的分离度。

紧密度是同一个类别之间相似的程度，紧密度越大，聚类内的相似度越大，聚类的质量也越高，表示为：

$$Com(C)=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{|C_i|^2}\sum_{m=1}^{|C_i|}\sum_{n=1}^{|C_i|}d(x_m,x_n) \tag{4.3-7}$$

其中 $x_m,x_n\in C_i$.

分离度反映属于不同类别之间相似的程度，分离度越大，聚类间的相似性越小，聚类的质量也越高，表示为：

$$Sep(C)=1-\frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{1}{|C_i|\cdot|C_j|}\sum_{m=1}^{|C_i|}\sum_{n=1}^{|C_j|}d(x_m,x_n) \tag{4.3-8}$$

其中 $x_m\in C_i,x_n\in C_j$.

针对 k 均值算法的主要不足之处，本章提出了相应的改进算法——自适应聚类算法。该算法能够自适应地确定聚类的个数，避免了在聚类数目选取上存在的主观性和盲目性，在一定程度上解决了聚类中的局部最优化问题。该方法的基本思路是：如果聚类数 k 定得比较合适，对应的聚类结果就能较好地分辨出不同类的神经元，那么所有类别内神经元相似度的平均值与不同类别间神经元相似度的

平均值的比值将趋向最大化，即聚类算法的同构度与异构度之比趋近于最大化，因此定义如下判别函数：

$$\Delta(k)=k\sum_{c_i\in c,x_m,x_n\in c_i}\frac{d(x_m,x_n)}{|c_i|\cdot(|c_i|-1)}\Big/\sum_{x_m\in c_i,x_n\in c_j}d(x_m,x_n) \tag{4.3-9}$$

其中 $d(x_m,x_n)=(\sum_{k=1}^{p}(x_{mk}-y_{nk})^2)^{1/2}$ 用来计算神经元之间的相似度，$|c_i|$ 表示第 i 个类别中神经元的数目，k 为类别个数，k 的取值依次从 2 迭代至待聚类神经元类别数，其中当判别函数取最大值时，所对应的 k 即为最佳的聚类数目，称其为判断函数的最大值属性。由判断函数的最大值属性来自适应地确定最终的聚类个数 k，进而自动待分类神经元潜在的类别数目。算法的具体描述如下：

步骤 1. 随机选择 k 个神经元作为初始类；

步骤 2. 通过计算剩余待聚类神经元特征向量与 k 个类均值向量之间的相似度，将这些神经元分派到与之最相似的 k 个类中的一个；

步骤 3. 计算聚类结果对应的自定义判别函数 $\Delta(k)$，并将当前自定义判别函数值与判别函数库中的数值进行比较，保存结果较大的判别函数值以及对应的 k 值到判别函数库中；

步骤 4. 更新 k 值，重新计算 k 个类的均值向量，重复步骤 2 和 3，直到获得最大判别函数值为止，保存最大判别函数值以及对应 k 值入库；

步骤 5. 此时判别函数库中保存的 k 值即为聚类类别的个数，聚类过程结束。

4.4 问题四

4.4.1 问题的分析

针对本问，需要研究同一类神经元中不同物种神经元之间的形态差异。根据问题三的方法，同一类的神经元必然在大体上是相似的，才能分成一类；本问要对这一类大体相似的神经元，进一步细化，以区分出同一类中的不同物种的神经元，这类内的差异性的挖掘是本问的核心问题。

通过上面分析，建立了一种分离度准则的不同物种神经元的区分方法，同时可以提取出差异较大的神经元的形态特征，运用这些提取出来的较大差异的特征，可以区分或区别不同物种的同种神经元。

本算法从分类的逆向思维出发，不是从相似度出发，而是以样本间的分离度出发，这是一种扩展，能较为有效的找到样本间的差异。

4.4.2 模型的建立

假设某一类神经元 C 中含有 r 个动物的神经元 $C_1,C_2,\ldots,C_r$，分别取每一种神经元的 m 条数据，利用问题 3 中提出的聚类方法计算类 C 中子类 $C_1,C_2,\ldots,C_r$ 之间的分离度：

$$Sep(C)=\sum_{C_i\neq C_j}Sep(C_i,C_j) \tag{4.4-1}$$

$$Sep(C_i,C_j)=\frac{1}{|C_i|\cdot|C_j|}\sum_{t=1}^{|C_i|}\sum_{s=1}^{|C_j|}d(x_t,y_s) \tag{4.4-2}$$

$$d(x_t, y_s) = (\sum_{k=1}^{p}(x_{tk} - y_{sk})^2)^{1/2} \quad (4.4\text{-}3)$$

其中，C_i, C_j 表示在这 r 个动物的神经元 $C_1, C_2, \ldots, C_r$ 中任取两种，x_t, y_s 分别表示 C_i, C_j 中包含的神经元，$t, s = 1, 2, \ldots, m$，x_{tk} 表示 C_i 中第 t 个神经元的第 k 个形态特征，$k = 1, 2, \ldots, p$.

由于分离度反映这 r 种神经元之间相似的程度，分离度越大，r 种神经元之间的相似性越小，则参与分离度 $Sep(C)$ 计算的形态特征区分这 r 种神经元的能力越强，同时也说明这 r 种神经元在这些形态特征上差异比较显著。因此，我们采用穷举法对这些形态特征进行组合，并计算在各种组合下类 C 的分离度，从而对区分神经元的主要形态特征进行显著性分析。具体算法描述如下：

步骤 1. 令 $F_p = \{f_1, f_2, \ldots, f_p\}$ 表示神经元的 p 个形态特征集合，分别计算第 $i(i = 1, 2, \ldots, p)$ 种形态特征 f_i 下式(4.4-1)的值，并求 $f_{j_1} = \max_{f_1, f_2, \ldots, f_p} Sep(C)$，记 $F_{p-1} = F_p - \{f_{j_1}\}$；

步骤 2. 分别取 F_{p-1} 中的元素分别与 f_{j_1} 组合，计算 $p-1$ 个组合下下式(4.4-1)的值，得到 $\{f_{j_1}, f_{j_2}\} = \max_{\{f_{j_1}, f_{j_k}\}, k=1,2,\ldots,p-1} Sep(C)$，记 $F_{p-2} = F_p - \{f_{j_1}, f_{j_2}\}$；

步骤 3. 令 $i = 2, \ldots, p-1$，分别取 F_{p-i} 中的元素与 $\{f_{j_1}, \ldots, f_{j_i}\}$，计算 $p-i$ 个组合下下式(4.4-1)的值，得到 $\{f_{j_1}, \ldots, f_{j_i}, f_{j_{i+1}}\} = \max_{\{f_{j_1}, \ldots, f_{j_i}, f_{j_k}\}, k=1,2,\ldots,p-i} Sep(C)$，记 $F_{p-(i+1)} = F_p - \{f_{j_1}, \ldots, f_{j_i}, f_{j_{i+1}}\}$；

步骤 4. 重复步骤 3，直到 $F = \varnothing$；

步骤 5. 记录按区分强度由大到小排列的的 p 个形态特征 $f_{j_1}, f_{j_2}, \ldots, f_{j_p}$.

4.4.3 模型的求解与结果分析

从附件 A 中分别选取 m 个猪的普肯野神经元和鼠的普肯野神经元，分别表示为 $X = \{x_1, x_2, \ldots, x_m\}$ 和 $Y = \{y_1, y_2, \ldots, y_m\}$，假设先将这两种动物的普肯野神经元神经元近似的看作两类神经元，分别记为 C_1, C_2，根据(4.4-1)式计算在各种形态特征下他们之间的分离度：$Sep(c_1, c_2) = 1/(|c_1| \cdot |c_2|) \sum_{i=1}^{|c_1|} \sum_{j=1}^{|c_2|} d(x_i, x_j)$，得到表 6。

分析表 6 可知，猪的普肯野神经元和鼠的普肯野神经元主要在胞体表面积、高度、深度、末梢数目、分支数目、长度，这些特征上差异比较大，分别对应表 6 中的第 1、7、8、5、4、12 种形态特征。

表 6 各种形态特征下类 C 内的分离度

Selected Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Max_Index
NULL	0.547331	0.333333	0.500112	0.503134	0.506173	0.484168	0.512873	0.521757	0.508391	0.492601	0.461363	0.495762	0.487208	0.475789	1
1	0	0.685402	0.864095	0.866557	0.869169	0.837038	0.870997	0.834439	0.863536	0.829715	0.822512	0.819505	0.840028	0.830895	7
1,7	0	0.989611	1.056632	1.058404	1.060313	1.038536	0	1.119069	1.046194	1.090219	1.083417	1.101238	1.09832	1.090896	8
1,7,8	0	1.229699	1.301914	1.303561	1.305333	1.276021	0	0	1.280409	1.262067	1.260145	1.267803	1.275452	1.267815	5
1,7,8,5	0	1.410152	1.442815	1.444222	0	1.428022	0	0	1.433749	1.443525	1.43962	1.451921	1.453566	1.446666	13
1,7,8,5,4	0	1.550102	1.586673	1.588223	0	1.570892	0	0	1.57867	1.571134	1.563696	1.578302	0	1.569498	4
1,7,8,5,4,12	0	1.680865	1.702155	0	0	1.69101	0	0	1.699114	1.703619	1.695913	1.711724	0	1.701432	12
1,7,8,5,4,12,3	0	1.801171	1.824642	0	0	1.81233	0	0	1.820314	1.813546	1.806968	0	0	1.812397	3
1,7,8,5,4,12,3,10	0	1.911059	0	0	0	1.916082	0	0	1.924273	1.925013	1.918239	0	0	1.923441	10
1,7,8,5,4,12,3,10,9	0	2.008611	0	0	0	2.014453	0	0	2.022158	0	2.010722	0	0	2.016003	9
1,7,8,5,4,12,3,10,9,11	0	2.10161	0	0	0	2.105217	0	0	0	0	2.106988	0	0	2.112203	11
1,7,8,5,4,12,3,10,9,11,14	0	2.183984	0	0	0	2.188786	0	0	0	0	0	0	0	2.190185	14
1,7,8,5,4,12,3,10,9,11,14,6	0	2.264593	0	0	0	2.270812	0	0	0	0	0	0	0	0	6
1,7,8,5,4,12,3,10,9,11,14,6,2	0	2.343502	0	0	0	0	0	0	0	0	0	0	0	0	2

4.4.4 模型的改进

从以上的分析中，不难看出，对形态特征的显著性分析主要依赖于分离度的计算，当分离度不是很明显或者变化很小时，算法的误差比较大，且分析结果不是很精确。对于实际问题，一般只需要考虑两种变量之间的差异，对于多个变量，则可以考虑两两对比的办法，因此可以选择典型相关性分析对两个变量的特征进行区分。

设 $X=(X_1,...X_p)'$ 及 $Y=(Y_1,...Y_p)'$ 为随机向量，用 X 和 Y 的线性组合 $\alpha'X$ 和 $\beta'Y$ 之间的相关性来研究两组随机变量 X 和 Y 之间的相关性，希望找到 α 和 β，使 $\rho(\alpha'X,\beta'Y)$ 最大。由相关系数的定义：

$$\rho(\alpha'X,\beta'Y)=\frac{Cov(\alpha'X,\beta'Y)}{\sqrt{Var(\alpha'X)}\sqrt{Var(\beta'Y)}}. \tag{4.4-4}$$

易得出对任意常数 e，f，c 和 d，均有：

$$\rho[e(\alpha'X)+f,c(\beta'Y)+d]=\rho(\alpha'X,\beta'Y)] \tag{4.4-5}$$

这说明使得相关系数最大的 $\alpha'X$ 和 $\beta'Y$ 并不唯一，故求综合变量时常限定 $Var(\alpha'X)=1,Var(\beta'Y)=1$。令 $V=\alpha'X,W=\beta'Y$，则 V,W 的相关系数

$$\rho(V,W)=\frac{\alpha'\Sigma_{12}\beta}{\sqrt{\alpha'\Sigma_{11}\alpha}\sqrt{\beta'\Sigma_{12}\beta}} \tag{4.4-6}$$

求第一对典型相关变量就等价于求 $\alpha=(\alpha_1,\alpha_2,...,\alpha_p)'$ 和 $\beta=(\beta_1,\beta_2,...,\beta_p)'$，使得在条件 $Var(\alpha'X)=1$ 和 $Var(\beta'Y)=1$ 下，$\rho(\alpha'X,\beta'Y)=\alpha'\Sigma_{11}\beta$ 达到最大。这是条件极值问题，用拉格朗日乘子法，令：

$$\varphi(\alpha,\beta)=\alpha'\Sigma_{12}\beta-\frac{\lambda_1}{2}(\alpha'\Sigma_{11}\alpha-1)-\frac{\lambda_2}{2}(\beta'\Sigma_{22}\beta-1) \tag{4.4-7}$$

其中 λ_1,λ_2 为拉格朗日乘子。为求 φ 的极大值。对上式分别关于 α,β 秋偏导，并令其为零，得

$$\begin{cases}\dfrac{\partial\varphi}{\partial\alpha}=\Sigma_{12}\beta-\lambda_1\Sigma_{11}\alpha=0\\ \dfrac{\partial\varphi}{\partial\beta}=\Sigma_{21}\alpha-\lambda_2\Sigma_{22}\beta=0\end{cases} \tag{4.4-8}$$

再分别用 α',β' 左乘以上方程，得 $\lambda_1=\lambda_2=\alpha'\Sigma_{12}\beta=\rho(V,W)\overset{\text{def}}{=}\lambda$，则方程组(4.4-8)等价于：

$$\begin{cases} -\lambda\Sigma_{11}\alpha + \Sigma_{12}\beta = 0 \\ \Sigma_{21}\alpha - \lambda\Sigma_{22}\beta = 0 \end{cases} \tag{4.4-9}$$

方程组(4.4-9)有非零解的充要条件是：

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0 \tag{4.4-10}$$

该方程左端是 λ 的 $p+q$ 次多项式。求解 λ 的高次方程(4.4-10)，把求得的最大的 λ 代回方程组(4.4-9)，再求得 α 和 β，从而得出第一对典型相关变量；具体计算时，因 λ 的高次方程(4.4-10)不易解；将其代入方程组(4.4-9)后还需要求解 $(p+q)$ 阶方程组。为了计算上的简便，作以下变换，用 $\Sigma_{12}\Sigma_{22}^{-1}$ 左乘方程组(4.4-9)的第二式，并把 $\Sigma_{12}\beta = \frac{1}{\lambda}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha$ 代入方程组(4.4-9)的第一式得 $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\alpha - \lambda^2\Sigma_{11}\alpha = 0$，再用 Σ_{11}^{-1} 左乘上式得：$(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I_p)\alpha = 0$ 然后由 p 阶特征方程求解 λ 和 α。

类似地，可通过求解 q 阶特征方程 $(\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I_q)\beta = 0$ 来得到 λ 和 β。故求解方程(4.4-10)等价于求解方程组(4.4-11)：

$$\begin{cases} \left|\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \lambda^2 I_p\right| = 0 \\ \left|\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \lambda^2 I_q\right| = 0 \end{cases} \tag{4.4-11}$$

由于 $\Sigma_{11} > 0, \Sigma_{22} > 0$，故 $\Sigma_{11}^{-1} > 0, \Sigma_{22}^{-1} > 0$，所以有

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{11}^{-1/2}(\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21}) \overset{\text{def}}{=} AB \tag{4.4-12}$$

其中 $A = \Sigma_{11}^{-1/2}, B = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21}$，但 AB 与 $BA = (\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21})\Sigma_{11}^{-1/2}$ 有相同的非零特征值，如果记 $T = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$，则 $BA = TT'$，故 M_1 与 TT' 有相同的非零特征值。类似地，$M_2 = \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2}\Sigma_{12}$ 与 TT' 有相同的非零特征值。

由以上分析可知，M_1 与 M_2 有相同的非零特征值，且非零特征值的个数至多为 p 个(因 $p \le q$)。设 $\left|TT' - \lambda^2 I_p\right| = 0$ 的 p 个特征值依次为 $\lambda_1^2 \ge \lambda_2^2 \ge \ldots \ge \lambda_p^2 > 0$，则 $T'T$ 的 q 个特征值中，除以上 p 个外，其余 $q-p$ 个均为 0。故方程(4.4-10)的 $q+p$ 各跟依次为 $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p > 0 = \ldots = 0 > -\lambda_p \ge \ldots \ge \lambda_2 \ge \lambda_1$ (λ_i 是 λ_i^2 的正平方根，

$i=1,...,p$)。取其中最大的λ_1代入方程(4.4-9)，即可求得$\alpha=a_1,\beta=b_1$(设a_1,b_1满足$a_1'\Sigma_{11}a_1=1,b_1'\Sigma_{22}b_1=1$)。令$V_1=a_1'X,W_1=b_1'Y$，则$V_1,W_1$为第一对典型相关变量；而$\rho(V_1,W_1)=a_1'\Sigma_{12}b_1=\lambda_1$为第一个典型相关系数。

4.5 问题五

4.5.1 问题的分析

针对神经元的实际形态是随着时间的流逝，树突和轴突不断地生长而发生变化的问题，首先应该对同一个神经元的各个时间的几何拓扑结构以及几何特征进行分析，找到其中的一些变化规律，为后面模型的建立提供基础；然后利用隐马尔科夫模型(HMM)建立以神经元的几何拓扑空间结构为隐变量的数学描述模型，其显式状态为神经元的几何特征；接着利用已有的数据预测出神经元显式的状态；最后利用得到的显式状态，计算出神经元的拓扑结构；这样的一个模型应该可以描述其生长的过程，并且其隐状态与显式状态之间的关系就是形态变化与几何特征之间的关系描述，也可以找出它们之间的关系。

4.5.2 模型的建立

隐马尔可夫模型(HMM)可以用五个元素来描述，包括 2 个状态集合和 3 个概率矩阵：

1. 隐含状态 S：这些状态之间满足马尔可夫性质，是马尔可夫模型中实际所隐含的状态。这些状态通常无法通过直接观测而得到，本问中为神经元的空间拓扑结构，用S_1,S_2,S_3等表示。
2. 可观测状态 O：在模型中与隐含状态相关联，可通过直接观测而得到。本问中为可以得到的神经元的形态特征的统计量，用O_1,O_2,O_3等表示，可观测状态的数目不一定要和隐含状态的数目一致。
3. 初始状态概率矩阵 π：表示隐含状态在初始时刻$t=1$的概率矩阵。当$t=1$时，$P(S_1)=p_1,P(S_2)=p_2,P(S_3)=p_3$，则初始状态概率矩阵 $\pi=[p_1\ p_2\ p_3]$。
4. 隐含状态转移概率矩阵 A：描述了 HMM 模型中各个状态之间的转移概率。其中$A_{ij}=P(S_j\mid S_i),1\le i,j\le N$表示在 t 时刻、状态为S_i的条件下，在$t+1$时刻状态是S_j的概率。
5. 观测状态转移概率矩阵 B。令 N 代表隐含状态数目，M 代表可观测状态数目，则：

$$B_{ij}=P(O_i\mid S_j),1\le i\le M,1\le j\le N \tag{4.5-1}$$

表示在 t 时刻、隐含状态是S_j条件下，观察状态为O_i的概率。

总结：一般的，可以用$\lambda=(A,B,\pi)$三元组来简洁的表示一个隐马尔可夫模型。隐马尔可夫模型实际上是标准马尔可夫模型的扩展，添加了可观测状态集合和这些状态与隐含状态之间的概率关系。

HMM 模型可以解决三类问题，分别是：(1)评价问题：给定模型参数$\lambda=(A,B,\pi)$及观察序列$O=[o_1,o_2,o_3,...,o_r]$。求此模型产生此观察序列的概率

$P_r[O/\lambda]$；(2)解码问题：给定模型λ及观察序列O；问此观察序列是模型λ中取怎样的状态次序$[q_1 \to q_2 \to ... \to q_r]$得到的。解决此问题的关键是采用什么作为取得结论的判据。通常是取产生此观察序列概率最大的一组状态序列$Q=[q_1,q_2,q_3,...,q_r]$作为判决。(3)识别(或称训练)问题：给定 HMM 的结构(指状态数 N，观察类数 M)，由给定的一组供训练用的观察组$O_1,O_2,O_3,...,O_r$估计该模型的最优参数$\lambda=(A,B,\pi)$。

本问由于未知参数，所以先要对 HMM 模型进行训练，解决的第三类问题，然后在运用得到的并带有参数的 HMM 模型计算隐变量的值，由此得到预测的神经元的拓扑结构的变化过程。

显式状态变量的预测，运用 Logistic 回归模型，对神经元的每一个几何特征进行沿时间序列进行预测，此回归模型的表达式如下：

$$y=f(x)=\frac{\alpha_1-\alpha_2}{1+e^{-\left(\frac{x-\alpha_3}{|\alpha_4|}\right)}}+\alpha_2 \tag{4.5-2}$$

其中 x 是时间序列刻度，y 是神经元的某一个几何特征；$\alpha_1,..,\alpha_4$ 是对某一个特征的一组预测参数。

4.5.3 模型的求解与结果分析

本节首先要下载相关数据，包括同一神经元在不同阶段的实测数据。网站[14]上只提供青年时期(young)和老年时期(aged)的神经元，并且不是一一对应的关系，所以针对神经元几何特征的描述只能是选取不同神经元的同一特征的统计平均值作为某一个几何特征，用来反映和建立与神经元拓扑结构之间的关系。

Logistic 回归模型的参数训练，它实质上是一个累计增长或者生长曲线，略呈现拉长的“S”形。求 Logistic 曲线方程的一阶导数，可以得到 Logistic 增长或生长过程的数度函数。

$$\frac{dy}{dx}=\frac{(\alpha_2-\alpha_1)/|\alpha_4|\cdot e^{-\left(\frac{x-\alpha_3}{|\alpha_4|}\right)}}{\left[1+e^{-\left(\frac{x-\alpha_3}{|\alpha_4|}\right)}\right]^2} \tag{4.5-3}$$

这样的模型用来描述神经元的生长与发育问题。把神经元的几何特征代入所需要拟合的曲线方程中。

确定系数，即求出模型。记

$$J(\alpha_1,...,\alpha_4)=\sum_{i=1}^{n}\delta_i^2=\sum_{i=1}^{n}[f(x_i)-y_i]^2 \tag{4.5-4}$$

为求$\alpha_1,...,\alpha_4$使$\frac{\partial J}{\partial \alpha_k}=0$，使用最小二乘准则计算模型参数，只需要利用极值的必要条件$\frac{\partial J}{\partial \alpha_k}=0\,(k=1,\ldots,4)$，得到关于$\alpha_1,...,\alpha_4$的方程组

$$\begin{cases} \sum_{i=1}^{n} r_1(x_i)(f(x_i)-y_i)=0 \\ \qquad\qquad \cdots\cdots\cdots \\ \sum_{i=1}^{n} r_4(x_i)(f(x_i)-y_i)=0 \end{cases} \tag{4.5-5}$$

其中 $r_1(x),...,r_4(x)$ 分别是 $f(x)$ 关于参数 $\alpha_1,...,\alpha_4$ 的导函数。使的满足 $\frac{\partial J}{\partial \alpha_k}=0$ 的上述方程的解就是模型的最优参数。

HMM 的训练过程，这是 HMM 的学习问题——Baum-Welch 估计算法。该算法是一种迭代算法，初始时刻由用户给出各参数的经验估计值，通过不断迭代，使各参数逐渐趋向更为合理的较优值。算法描述如下：

步骤 1. 初始化：$\tilde{\pi}_i=\gamma_1(i)$ 当 t=1 时处于 S_i 期望值，$\lambda=(A_0,B_0,\pi)$

步骤 2. 迭代计算：令 $\zeta_t(i,j)$ 表示 t 时状态为 i 以及 t+1 时状态为 S_j 的概率。

$$\zeta_t(i,j)=\frac{P(q_t=i,q_{t+1}=j,O\mid\lambda)}{P(O\mid\lambda)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O\mid\lambda)}$$

$$=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}$$

其中 $\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$，t 时刻处于状态 S_i 的概率；$\sum_{t=1}^{T-1}\gamma_t(i)=$ 整个过程中从状态 S_i 转出次数的预期；$\sum_{t=1}^{T-1}\xi_t(i,j)=$ 从 S_i 跳转到 S_j 次数的预期；重估公式：$\tilde{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i,j)}$，和 $\tilde{b}_{ij}(k)=\frac{\sum_{t=1}^{T}\gamma_t(i,j)}{\sum_{t=1}^{T}\gamma_t(i,j)}$；

步骤 3. 终止条件：$\left|\log P(O|\lambda)-\log P(O|\lambda_0)\right|<\varepsilon$，其中 ε 是预先设定的阈值。

这样就得到了几何形态特征与形态变化的隐马尔科夫关系。预测的时候，先预测几何形态特征的变化，然后由隐马尔科夫模型预测出神经元形态的生长变化，以及其拓扑结构的变化。

结果分析：

通过模型对神经元的生长变化趋势的预测，在分析了这些大量的不同年龄段的神经元数据的形态特征，可以发现伴随着年龄的增长，神经元的胞体表面积、

表面积、长度、体积、以及所占空间范围等都有不同程度的增大，而干的数目、分支数目以及末梢数目等则呈下降趋势。由此可见神经元的生长伴随体积的增长以及细节的损失。

五． 模型的评价

5.1 模型的优点

（1） 采用较为成熟的数学理论建立模型，可信度比较高；

（2） 模型原理简单明了，容易理解与灵活运用；

（3） 建模的方法和思想对其他类型也适合，易于推广到其他领域；

（4） 模型方便、直观、易于在计算机上实现与推广；

（5） 模型的计算采用专业数学软件，可信度较高，便于推广。

5.2 模型的缺点

（1） 模型虽然综合考虑到了很多因素，但为了建立模型，理性化了许多影响因素，具有一定的局限性，得到的最优方案可能与实际有一定的出入；

（2） 建模过程中，简化了一些因素，因而与实际问题有偏差；

（3） 由于时间和缺少数据的原因，对某些问题提出了改进模型，但没有具体实现。

参考文献

[1] 姜启源，谢金星，叶俊，数学模型，北京：高等教育出版社，2005

[2] 萧树铁，姜启源，何青，数学实验，北京：高等教育出版社，2004

[3] 徐萃薇，孙绳武，计算方法引论，北京：高等教育出版社，2005

[4] 高惠旋，应用多元统计分析，北京：北京大学出版社，2005

[5] 邓乃杨，田英杰，数据挖掘中的新方法——支持向量机，北京：科学出版社，2004

[6] 孙啸，陆祖宏，谢建明，生物信息学基础，北京：清华大学出版社，2004

[7] 高新波，模糊聚类分析及其应用，西安：西安电子科技大学出版社，2004

[8] 王松桂，陈敏，陈立萍，线性统计模型，北京：高等教育出版社，1999

[9] Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition, Proc. of the IEEE, 1989, 77(2): 257-266.

[10]汪云九，神经信息学，北京：高等教育出版社，2006

[11]http://neuromorpho.org/neuroMorpho/index.jsp

[12]http://senselab.med.yale.edu/NeuronDB/ndbRegions.asp?sr=1

[13]http://krasnow.gmu.edu/L-Neuron/L-Neuron/database/index.html#Scorcioni

[14]http://www.compneuro.org/CDROM/nmorph/index/topindex_tn.html