编号

2022-2023学年第一学期



《深度学习应用开发基础》

课程期末大作业



题 目 人类染色体图像核型分类

姓 名 杨晨露

学 号 32001298

专业班级 计算机2001

二○二三 年 二 月

### 选题背景

染色体核型分析是细胞遗传学研究的主要技术之一，在现代生物和医疗领域是了解生物细胞染色体的构成、研究物种净化、诊断遗传疾病不能缺少的重要手段。一般在染色体核型分析的过程中，我们首先要在染色体中期图像中分割出单挑染色体，再将染色体一一进行分析、比较、排序和分类。以前由于技术支持的工具精确度低，起到的作用不大，而现在将计算机技术与染色体核型分析相结合，大大提高了分析效率，也方便了生物学和遗传学邻域的相关研究。

本系统主要通过计算机对人类染色体图像进行分类。

### 开发环境

本系统的开发工具使用visual Studio Code，在该免费平台下的python环境建立后端，第三方库函数有PyQt5、matplotlib、cv2、skimage 、tensorflow等。

### 数据处理

# 1.染色体图像预处理

## 1.1图像分割

每个染色体图像都有不同的对比度，将导致计数精度降低，为此，需要对染色体图像进行增强。首先进行图像分割，将染色体图像前景区域与背景分离开来。

最大类间法（OTSU），是一种图像二值化算法。其算法假设图像中存在一个阈值，判断图像中每个像素与的大小关系，可以将所有像素分为前景（物体）和背景两类，当选取到最佳阈值时，背景部分与前景部分的差别最大，而OTSU算法利用最大类间方差来衡量这一差别[2]。鉴于这一定义，OTSU法将图像分为明暗区域和

其中是阈值，是图像最大灰度等级。

OTSU的阈值处理方法是扫描所有可能的阈值，并计算阈值两侧的像素水平的最小值，目标是找到前景和背景之和最小熵的阈值，对于选定的阈值t，可计算出聚类和的方差。最佳阈值是通过最小化加权组方差之和来计算的，其中加权是各组的概率。

给出：为观察到的灰度值的直方图概率

其中、分别是图像的行和列的索引，、分别是图像的行数和列数。

、和分别为强度值在到之间的聚类的权重、平均值和方差。

、和分别为强度值在到之间的聚类的权重、平均值和方差。

是聚类间方差的加权和。

最佳阈值是聚类间方差加权和最小的值。

聚类间方差的定义：

其中

以训练集中的下图为例



图1 train中1号染色体中的“16w0078.043.K.R.PNG”原图

使用OTSU法后



图2 最大类间法处理后

经过OTSU处理后的染色体图像呈现异常凹陷、存在小的干扰块，需要进行图像形态学开运算处理，即对图像先进行腐蚀处理，再进行膨胀处理。利用开运算操作去噪，而保留目标区域的几何结构，起到消除孤立点的作用。

首先定义一个结构元素，结构元素是图像目标区域的更小的组成部分，多个结构元素构成目标图像。设IM(input image)表示输入图像，SE(structuring element)表示结构元素，开运算可以表示为

其中、分别表示腐蚀操作、膨胀操作[4]。

假设为平面坐标上的目标区域，结构元素为结构元素，使用结构元素对的腐蚀操作可以表示为

膨胀操作可以表示为

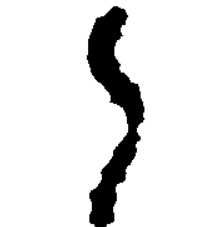


图3 开运算处理后

## 1.2边缘平滑

经过阈值分割的染色体图像边缘区域呈现较突出的锯齿形态和毛刺，为了降低后续对轮廓提取及特征提取的影响，需要使用边缘平滑方法处理二值化的染色体图像，使用高斯滤波处理来进行边缘平滑是一种很好的选择。

高斯滤波是一种线性滤波器，能够较好地平滑与抑制图像噪声，是对图像像素进行加权平均的一个过程，距离中心点越近的部分，其权重越大，反之越小，而权值的分布，则符合正态分布，对于一个二维图像来说，则符合二维高斯分布。

高斯滤波器的一个重要性质是它是唯一满足不确定性关系的算子

其中、分别是其在空间域和频率域的方差。这一特性使得高斯算子可以在空间域和频率域的定位的冲突目标之间进行最佳权衡。

一维高斯函数为

其中为高斯函数的标准差，也是图像的笛卡尔坐标。

二维高斯滤波器也是唯一的旋转对称滤波器，是可分离的笛卡尔坐标。可分离性对空间域卷积平滑运算的计算效率至关重要。二维高斯函数为[5]

通过对图像应用不同尺度的高斯滤波器，可以获得具有不同平滑程度的图像。

在实际操作中，需要对图像的边缘进行补偿，以便卷积核可以对图像边缘部分的像素进行计算，我们使用0补偿（补偿的像素值为0）方式来进行处理。结果如图5所示：



图5 边缘平滑处理后

## 1.3轮廓提取

染色体的外部轮廓直观地描述了染色体的大小、形状等几何信息，是染色体分析过程中的重要信息。另外，染色体轮廓在提取中轴线特征以及着丝点特征时具有重要运用，是特征提取的重要一步。提取图像轮廓的常用方法有边缘检测法、形态学边界提取法、边界追踪法等方法。

在通常的图像处理过程中，边缘检测法是最常用的一种方法，几乎所有的边缘检测思想都是以像素灰度差分为基础的，其基本原理是通过计算邻域像素的一阶导数、二阶导数以及灰度梯度来进行，将计算过程进行简化就形成检测算子，常用的梯度检测算子有Roberts、Sobel、Canny等[1]。

使用Canny边缘检测算法，在染色体图像边缘平滑步骤中，我们已经对图像进行高斯滤波去除噪声，则此时我们需要设置两个阈值来确定边界。设minv、manv分别为最小阈值和最大阈值，当图像的灰度梯度高于manv时，被认定为真的边界；反之图像的灰度小于minv时，则被舍弃；如果灰度梯度介于两者之间，丢弃与否取决于此点是否与与某个被认定为真的边界点相连，如果是，就被认定为是真的边界点，反正舍弃。

如图4所示，A点高于阈值maxv，所以直接被认定为真的边界点，B点和C点虽然都介于minv和maxv之间，但B点与A点相连，所以也被看作边界点，C点不与边界点相连，所以被舍弃[6]。

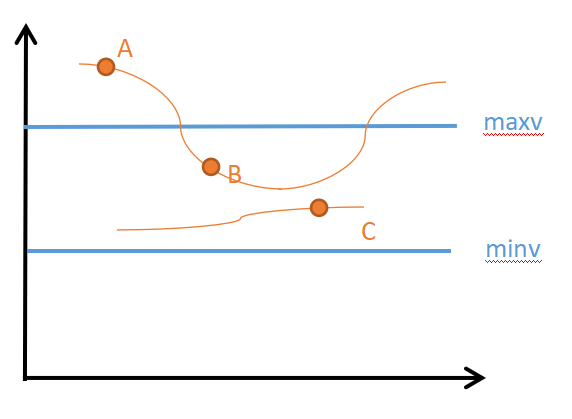


图5 Canny算法示例

使用Canny边缘检测，有以下步骤：

计算梯度振幅和方向，通常梯度方向采取4个角度：-0，45，90，135度

其中、作为一对卷积数组

再进行非最大限度的抑制，消除非边缘像素，只留下一些细线，最后选择滞后阈值，保留或排除像素，以选择边缘。处理结果如图6所示：

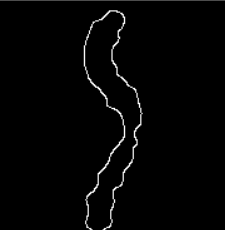


图6 轮廓提取处理后的图像

# 2.染色体图像特征提取

染色体图像中非同源染色体的形态、大小以及带纹特征各有不同，这些差异性是染色体识别分类的重要依据。染色体图像的特征是整个染色体核型分类的重要且关键的一步。特征的选择以及数据的质量直接影响到识别分类结果的准确率，根据染色体图像的差异性，选取染色体图像的长度、面积、着丝点指数、带纹分布特征作为主要的分类依据。

## 2.1染色体中轴线提取

染色体的中轴线是指贯穿整条染色体一条单像素的曲线，它描述了染色体的长度和弯曲形态等几何信息，是染色体长度数据提取的关键。在染色体着丝点定位以及带纹分布量化的过程中也会基于中轴线作量化计算，染色体中轴线的提取是整个染色体特征数据提取的重要环节。

### 2.1.1Zhang-Suen细化算法提取骨架

染色体中轴线的提取重要的一步是染色体图像骨架提取，再对染色体骨架进行延长，拟合等优化处理得到完整染色体图像中轴线。染色体图像骨架的提取有多种方法可以选择，本文采用Zhang-Suen细化算法提取骨架。

Zhang-Suen细化算法首先定义像素点的8邻域（前景为1，背景为0），假设点的邻居为点，，，，，，，。如图7所示：

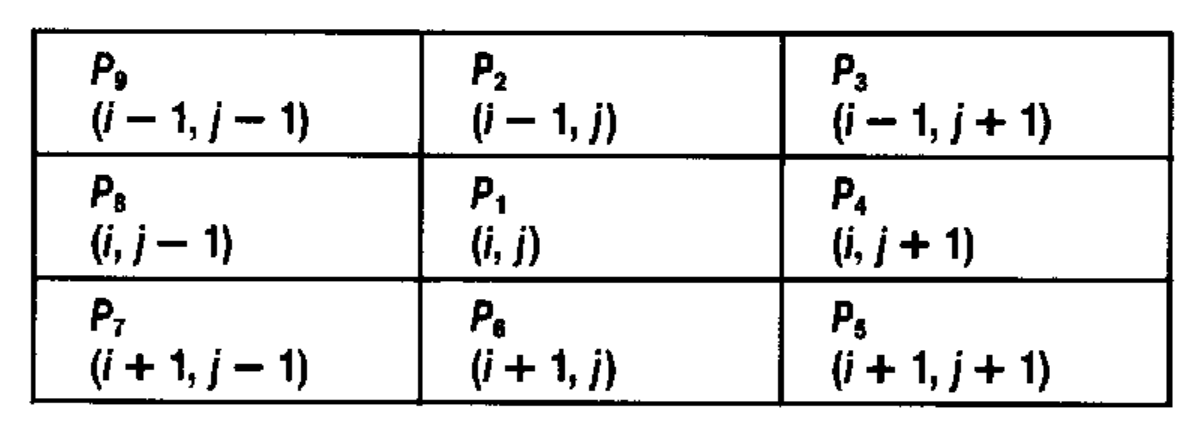


图7 3×3的二进制矩阵

在并行图像处理中，在第n次迭代中给予一个点新值，这个新值取决于它自己的值和八个邻居在第(n-1)次迭代时的值，所以所有图片点可以被同时处理。

Zhang-Suen细化算法通常是一个迭代算法，整个迭代过程分为两步：

第一步：循环所有前景像素点，对符合如下条件的像素点标记为删除：

(a)

(b)

(c)

(d)

其中表示P2、P3…P9像素中出现0~1的累计次数，其中0表示背景，1表示前景,表示跟P1相邻的8个像素点中，为前景像素点的个数

第二步，与第一步类似，只有(c)和(d)条件改变：

()

()

循环上述两个步骤，直到两步中都没有像素被标记为删除为止，输出结果即为二值化图像细化后的骨架[7]。处理结果如图8所示：

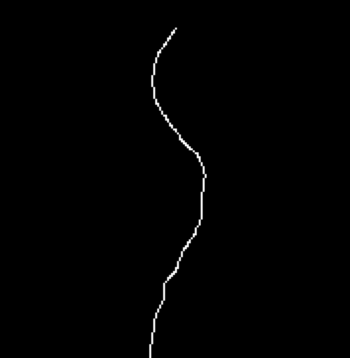


图7 提取骨架后的图像

### 2.1.2中轴线拟合及延长

目前提取的染色体中轴线并不完整，在染色体两端会产生部分缺少，需要将中轴线两端延长至边缘轮廓才能得到一条完整的染色体中轴线。初步提取的骨架端点通常位于分叉点，距离染色体两端边缘轮廓距离相对较近，因此分布取初步提取的骨架两端若干个点进行曲线拟合并延长至轮廓即可。

使用最小二乘法进行曲线拟合是一种常用的曲线拟合方法。主要思想是计算未知参数并使得理论值与观测值的误差平方和达到最小。假设需要拟合的多项式为：

离散的点到曲线的平方和为：

对分别求的偏导数，得到m个等式：

整理这m个等式形成的线性方程组：

解方程组得到系数、…确定拟合曲线[8]。

图8所示为最终效果：

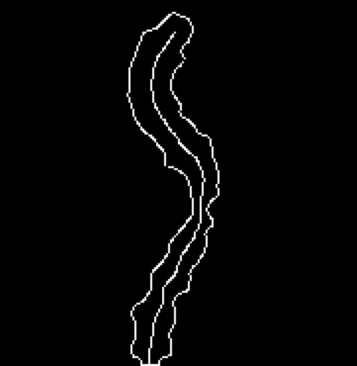
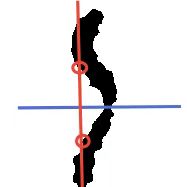


图8 拟合后的图片

2.2着丝点定位

在人体染色体图像上，位于着丝点处的染色体图像轮廓分别向内收缩形成凹陷，凹陷区域的横截宽度最短，因此可以计算染色体横截宽度最小值来确定最小宽度的位置，进而确定着丝点位置。染色体的中轴线的形态反映了染色体的弯曲状态，通过寻找染色体中轴线上的垂线交于染色体轮廓两交点形成的截线的方法可以较准确地得到染色体每一处的横截宽度，比较中轴线上每一点处的横截宽度大小即可确定着丝点的具体位置。如图，蓝色为染色体的中轴线，红色为中轴线的垂线，交点之间即为横截宽度。



具体步骤为：

1. 拟合中轴线：从染色体中轴线的某一端开始，假设染色体中轴线点集为,取连续的5个点使用最小二乘法进行二次曲线拟合（端点及端点附近取不到 5 个点时则只取 3 个或 4 个连续点），得到曲线表达式;
2. 中轴线上点的斜率：利用计算点的斜率，并求出该点法线斜率为,垂线方程公式为：
3. 中轴线的垂线与轮廓的交点：染色体轮廓的点集为,中轴线的垂线与轮廓交于两点，在图像坐标系中，这两点到垂线的距离最小，轮廓每个点到垂线距离计算方法为
4. 计算宽度：在确定了两点之后就可以计算两点之间的距离，即染色体的宽度
5. 定位着丝点：计算染色体中轴线上所有点处的横截宽度值，当宽度最小时的点即为着丝点的位置。



2.3带纹提取

在着丝点定位中已经计算出了染色体的中轴线的垂线，推导出垂线的解析式，计算出染色体中轴线的垂线所经过的像素点，并且这里的计算范围可以限定在教育轮廓的两端点纵坐标之间的范围内，降低计算量。

为了增强亮暗带之间的界限，需要对染色体图像进行灰度拉伸处理。灰度拉伸也叫做对比度拉伸，灰度拉伸是分段线性变化如公式所示：

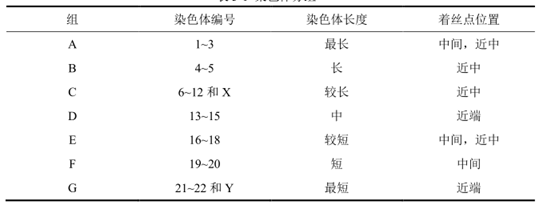
### 模型设计

# 1.染色体分类

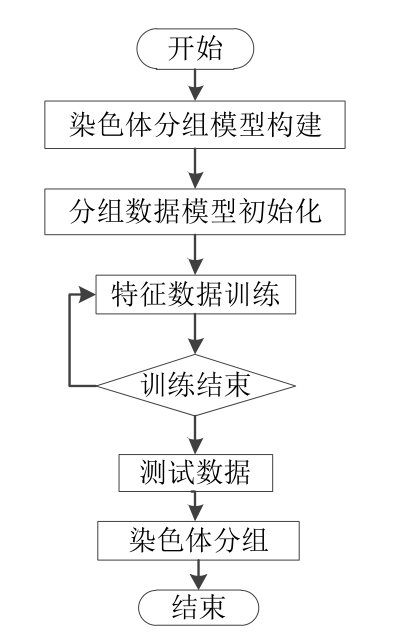
1.1基于BP神经网络分组

人类体细胞分裂中期含有 23 对染色体，其中有 22 对染色体为常染色体，男女共同拥有，按照长度划分为 1~22 号染色体，第 23 对染色体为性染色体，男女的性染色体不同，女性为两条 X 性染色体，男性为一条 X 性染色体和一条 Y 性染色体。

国际上根据染色体的形态特征和带型位置将染色体分为 A~G 共七个组



构建BP神经网络模型流程如图：



BP 神经网络的输入输出数据的特点决定了网络的结构，构建网络模型时，主要考虑输入节点数、输出节点数、隐含层节点数、隐含层数、激活函数、误差函数以及学校速率等。对于染色体图像的初步分组：

（1）本文输入参数包括染色体相对长度、相对像素面积、着丝点指数三个特征，输入层节点数为 3。所要分为 A~G 共 7 个组，输出节点为 7。

（2）隐含层数的选择没有确定的规则，一般来说网络层数越多，模型的预测精度越高，但训练时间也会更加复杂，网络的复杂性越高。考虑到样本的数量以及分类问题的复杂性，本文网络隐含层数为 1 层。

（3）隐含层结点数目的确定也并没有具体的要求，一般来说隐含层结点数的设置可以参考如下公式：

这里 n 为输入层节点数，l 为隐含层节点数，m 为输出层节点数，a 为 0~10 之间的常数。隐含层的节点数参考公式范围进行实验，选取合适的节点数。

（4）学习速率的设置也是一个重要因素，学习速率过小则网络计算量加大，收敛过慢，学习速率过大可能使网络不稳定。

（5）误差函数选择误差平方和函数。

（6）BP 神经网络隐含层一般选择 S 激活函数，输出层实验线性激活函数。

1.2基于k-means的组内分类

k均值聚类算法是一种[迭代](https://baike.baidu.com/item/%E8%BF%AD%E4%BB%A3/8415523?fromModule=lemma_inlink" \t "_blank)求解的聚类分析算法，其步骤是，预将数据分为K组，则随机选取K个对象作为初始的[聚类中心](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E4%B8%AD%E5%BF%83/55999889?fromModule=lemma_inlink" \t "_blank)，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695?fromModule=lemma_inlink" \t "_blank)。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，[误差](https://baike.baidu.com/item/%E8%AF%AF%E5%B7%AE/738024?fromModule=lemma_inlink" \t "_blank)[平方和](https://baike.baidu.com/item/%E5%B9%B3%E6%96%B9%E5%92%8C/783894?fromModule=lemma_inlink" \t "_blank)局部最小。

先随机选取K个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是以下任何一个：

1. 没有（或最小数目）对象被重新分配给不同的聚类。
2. 没有（或最小数目）聚类中心再发生变化。
3. 误差平方和局部最小。

使用 BP 神经网络分组的 A~G 共 7 个组中，不同的组中染色体类数不同，因此针对 A~G 组分别聚类。

在进行聚类时，染色体图像的特征参数可由相对长度、着丝点指数、相对面积、短臂带纹数目、长臂带纹数目共 5 个维度组成 [45] ，本例使用欧氏距离作为距离计算方式。聚类过程包括训练和测试两个过程，基于 BP 神经网络的分组结果，根据 A~G每组染色体类数进行 k-means 聚类训练，使用测试样本进行聚类测试。

对于需要分组聚类染色体图像，提取所需要的特征数据，使用训练完成的 BP神经网络进行分组，再对每组使用相应的训练完成的 k-means 分类器进行组内分类，完成所有染色体的分类。

### 项目组成员分工和贡献（分组作业适用）

小组成员信息：

组长：杨晨露、32001298、计算机2001、1908523502@qq.com、18758211467

组员1：何佳静、32001017、计算机2001、109147420@qq.com、15968175001

组员2：常睿嘉、32001027、计算机2001、532607269@qq.com、19818519663

贡献度（不管多少名成员，所有成员的贡献度之和为1）：

杨晨露贡献度：0.32，何佳静贡献度：0.33，常睿嘉贡献度0.35

### 课程感想

课本代码解说较详细，十分有利于学生学习。课程算法较难，训练时间较长，代码能力较弱时，需要花费大量时间学习，有时候收效甚微。

### 参考文献

1. Abid F, Hamami L. A survey of neural network based automated systems for human chromosome classification[J]. Artificial Intelligence Review, 2018, 49: 41-56.
2. Xu X, Xu S, Jin L, et al. Characteristic analysis of Otsu threshold and its applications[J]. Pattern recognition letters, 2011, 32(7): 956-961.
3. Yousefi J. Image binarization using Otsu thresholding algorithm[J]. Ontario, Canada: University of Guelph, 2011, 10.
4. Said K A M, Jambek A B, Sulaiman N. A study of image processing using morphological opening and closing processes[J]. International Journal of Control Theory and Applications, 2016, 9(31): 15-21.
5. Basu M. Gaussian-based edge-detection methods-a survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2002, 32(3): 252-260.
6. Xu Z, Baojie X, Guoxin W. Canny edge detection based on Open CV[C]//2017 13th IEEE international conference on electronic measurement & instruments (ICEMI). IEEE, 2017: 53-56.
7. Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of the ACM, 1984, 27(3): 236-239.
8. Gavin H P. The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems[J]. Department of Civil and Environmental Engineering, Duke University, 2019, 19.