## 问题分析

（接上次）鉴于自体Treg细胞分离培养十分困难，另外，动物实验与B细胞共培养实验也难以进行，更为重要的是，实验使用的Jurkat细胞系十分特殊，对于上述问题的建模也十分局限。因此，建模必须在几乎不依靠实验数据的基础上来进行，从合成生物学角度来讲，对于合成生物系统的建模中，通用的方法是工程化的，一般分析合成生物系统的性能，主要有以下几个主要指标：（1）稳定性（2）鲁棒性（3）相应快速性；其中稳定性和鲁棒性（又称稳健性）较为重要，我们将着重选取其中的稳定性指标进行重点分析。

系统的稳定性对于我们的系统是十分重要的，因为我们的系统是用于治疗的一种尝试，首先考虑的应该是安全性的问题，通过实验测定黑箱系统的工作参数，以此设定治疗方案或优化系统设计，这对该项目设想将来的发展具有重要的参考价值。

在考虑了系统的安全性后，下一个重要指标也就必须列入考虑范围内：有效性，有效性和安全性对于一个系统而言就如数学证明中充分性与必要性一样，一个治疗系统在保证安全性的条件下，首先要考虑的应该是有效性，有效性的评价方法是多样的，一般可通过与对照组的实验得出，在这种情况下，对照组通常是有效性得到确定评价的，但由于我们项目的特殊性，这一方法难以实行。这里，我们考虑单纯从验证系统中各组件运行状态出发，先通过模拟建模，再以实验数据进行模型的修正，计算系统在各参数下运行的状态。

## 模型建立

### Part 1: 基于机器学习的实验数据挖掘与参数计算

##### 机器学习概要

Machine Learning(机器学习)是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

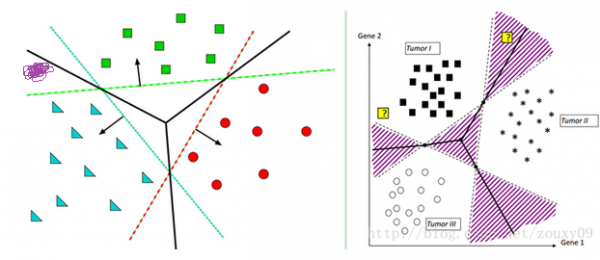
目前存在几种不同类型的学习算法。主要的两种类型被我们称之为监督学习和无监督学习。

1. 采用机器学习中什么方法，具体介绍
2. 如何用此方法进行数据挖掘和参数计算
3. 所需的实验数据
4. 对于方案存在的问题进行讨论（模型优缺点），可行性等

1. 采用机器学习中什么方法，具体介绍

（1）无监督学习

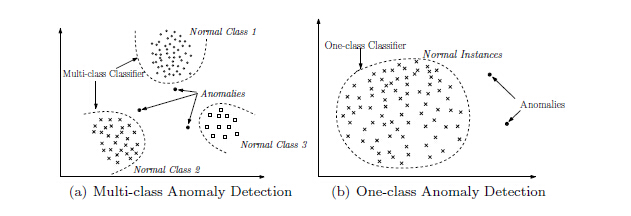
无监督学习最重要的用途应该就是用来研究分类问题，通过对数据进行聚类分析，主成分分析，最终找出数据中的关键因素，以及找出数据之间潜在的类别。这些在经典的统计学中是很难做到的，但是借助现代计算机的高速运算能力，我们可以让计算机通过对数据进行“学习”，从而对数据进行分类。



【无监督学习二维空间分类示意图】

（2）异常值检测

异常值检测是基于多元高斯分布的一种机器学习算法，它可以允许我们使用训练集对模型进行训练，并通过交叉测试集对模型进行改进提高预测的准确性，降低误报率，最终可以利用测试集对模型进行评估，从而实现对系统的安全性和稳定性进行评估。



【异常值检测示意图】

2. 如何用此方法进行数据挖掘和参数计算

（1）利用无监督学习找到系统潜在的稳态

对于这样一个复杂体系，难免会出现很多稳态，但是发现这些稳态是比较困难的，因此可以使用机器学习中的无监督学习对系统的稳态进行挖掘。通过聚类分析，可以讲数据点聚成多维空间中的一些类，这些类很可能就是一些潜在的系统稳态，因为如果数据能够成簇出现，那么这个类中心所对应的动力学参数大多是稳定的（可以用微分方程的向量场理论解释）。因此用非监督学习对数据分类可以指导后期进行动力学建模。

（2）利用异常值检测方法评估系统的稳定性

异常值检测方法实际在医学中应用广泛，比如我们用一些疾病的患者和正常人血常规的数据训练对模型进行训练，然后有病人的新的数据输入后，模型可以指示此病人的血常规是否正常以及可能患有什么疾病。实际医生看病的原理和这个差不多，只是我们这里不是训练一个医生，而是训练一个机器。

我们通过异常值检测模型，可以预测我们的系统是否工作正常，以及如果不正常是出了哪方面的问题。

3. 所需的实验数据

在不同IL-17A浓度梯度和时间梯度下的Foxp3,USP7。以及Foxp3,USP7的mRNA的QPCR数据（非必须）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | t1 | t2 | t3 | t4 | tn |
| C1 |  |  |  |  |  |
| C2 |  |  |  |  |  |
| C3 |  |  |  |  |  |
| C4 |  |  |  |  |  |
| Cn |  |  |  |  |  |

【时间梯度加浓度梯度数据示意表】

4. 对于方案存在的问题进行讨论（模型优缺点），可行性等

（1）优缺点

对于这样的一个模型，我们的优点是模型的算法较为先进，缺点是我们的数据可测量难度大，数据量小。由于这是一个生物复杂系统，我们很难像测量一个机械系统那样做到数据量大准确。我们能测量的数据很少，并且每个数据的测量代价都非常大。因此在这种情况下经典统计学的方法都失效了，因此我们通过建模的分析想找到一种可以用少量数据就可以达到较好的效果的方法，因此决定采用机器学习的方法，机器学习在处理复杂系统方面非常有优势，因为它的算法非常类似人类的思考方式。

当然，再好的算法如果没有数据也是不行的，因此我们再保证能够完成的情况下尽可能多的数据点，这样模型会更加精确。

（2）可行性分析

我们数模组成员之前都学习过一些机器学习的知识，也做过一些真实的案例。实验方面量也不是很大，在保证完成主要实验的基础上可以完成，当然这些实验数据对于我们的课题也是非常有必要的。

### Part 2 : 外部噪声诱导的状态切换