代谢综合征风险、趋势预测和干预模型的构建

张佳伟 吉肖豪 陈旭宇

指导教师：数模教练组

复旦大学数学科学学院，上海200433

摘 要

代谢综合征现已成为危害我国民众健康的严重公共卫生问题，患者往往表现出代谢指标的异常，如中心性肥胖、血压及血糖水平增高、血脂异常等。随着代谢综合征发病率的不断升高，对我国公众生命健康的危害不断加大。如果能够对代谢综合征的发生和发展趋势进行预测，将会对代谢综合征的预防和治疗起到极为关键的作用，从而有效降低代谢综合征的发病率。

本文通过将生物信息学理论和现代数据科学技术相结合的方法来建立数学模型，进而根据研究对象的遗传信息来预测其罹患代谢综合征的可能性。主要内容分为以下三个部分：

第一节对从中国知网以及ECBI等数据库中收集到的代谢综合征的相关信息进行了梳理和整合，详细介绍了代谢综合征预测问题的研究背景和重要意义。

第二节对已有数据进行了解读和阐释，并对DNA数据进行了简化和预处理，将每个人的基因变异信息处理成向量。对处理后的DNA数据和RNA数据进行了相关性分析，再根据229个样本筛选出相关程度较高的DNA和RNA作为模型的输入参数。

第三节以支持向量机为基础建立了代谢综合征的预测模型，并与其他处理方法进行了比较。在运算过程中对各种可能出现的情况进行了分类讨论，对预测模型的准确性进行了一定的评估，并对模型的优点和可改进之处进行了一定的探讨，指出了预测模型对代谢综合征预防和治疗的积极意义。

**关键词：**代谢综合征；基因；相关性分析；支持向量机；组学研究

目录

[1 问题的背景和研究意义 3](#_Toc462829885)

[1.1 代谢综合征发病机理文献综述 3](#_Toc462829886)

[1.1.1 预测模型的研究背景 3](#_Toc462829887)

[1.1.2 代谢综合症基本特征 3](#_Toc462829888)

[1.1.3 常见发病人群分布[3] 4](#_Toc462829889)

[1.1.4 代谢综合征的主要致病因素 5](#_Toc462829890)

[1.2 组学数据研究和分析 5](#_Toc462829891)

[1.2.1 基本定义 5](#_Toc462829892)

[1.2.2 组学数据的处理方式 6](#_Toc462829893)

[1.2.3 组学数据的研究意义 6](#_Toc462829894)

[2 预测模型设计 7](#_Toc462829895)

[2.1 数据预处理 7](#_Toc462829896)

[2.2 预测模型的建立 10](#_Toc462829897)

[2.2.1 关联矩阵动态量化模型 10](#_Toc462829898)

[2.2.2 关键通路的寻找 11](#_Toc462829899)

[2.2.3 根据个人信息预测发病风险 16](#_Toc462829900)

[3 进一步研究方向 20](#_Toc462829901)

[参考文献 21](#_Toc462829902)

# 问题的背景和研究意义

## 代谢综合征发病机理文献综述

### 预测模型的研究背景

代谢综合征(MS)包括一簇心血管疾病危险因素,如中心性肥胖、血压及血糖水平增高、血脂异常。MS与糖尿病、心血管及肾脏疾病的发生以及心血管疾病死亡、全死因死亡的危险性增高相关。根据InterAsia结果提示,按照ATPIII的定义,我国35-75岁的成年人中6400万（13.7）的个体患有MS,而38%的男性和35%的女性至少拥有一种MS组分。根据IDF标准我国35-74岁的成年人中患有MS的人数上升至7700万(16.5％)。北方居民中MS的患病率高于南方居民,城市居民高于农村居民。这些结果表明MS已经成为我国的严重公共卫生问题。[1]经济的发展以及生活方式的改变或许可以解释这一现象。因此，对于MS的研究和预测自然成为了极为重要的问题。本论文通过对大量相关论文的解读，提炼出了相关方面最具价值的信息，从而为代谢综合征的预测模型提供了坚实的理论依据。

### 代谢综合症基本特征

根据国际糖尿病联盟(IDF)定义,要确认个体是否患代谢综合征,一般可以从以下两个个方面定义，若同时满足条件1,2，则可以定义此人存在代谢综合征[2]。

1. 中心性肥胖:  
   在欧洲裔人种中定义为男性腰围>=94cm女性腰围>=80cm,在其他人种中采用种族特异性的腰围切点见表1[2]。
2. 另加下列4因素中任意两项:
   * 甘油三酯(TG)水平升高>150mg/dl(1.7nmol/L),或已接受针对此脂质异常的特殊治疗
   * 高密度脂蛋白胆固醇(HDL-C)水平降低:男性<40mg/dl(1.03nmol/L),女性<50mg/dl(1.29nmol/L)或已接受针对此脂质异常的特殊治疗
   * 血压升高:收缩压>=130mmHg或舒张压>=85mmHg,或此前已被诊断为高血压而接受治疗
   * 空腹血糖升高:空腹血糖>=100mg/dl(5.6nmol/L),或已被诊断为2型糖尿病。如果空腹血糖>=100mg/dl(5.6nmol/L),则强烈推荐行口服葡萄糖耐量试验(OGCT);但是OGTT在诊断代谢综合征时并非必需。

注:国内外标准略有差异，根据2007年中国成人血脂异常防治指南，符合以下3项者即可诊断为代谢综合征:

* + - 腹部肥胖:男性腰围>90cm、女性腰围>85cm
    - TG（甘油三酯）≥1.7mmol/L
    - HDL-C（高密度脂蛋白胆固醇）<1.04mmol/L
    - 血压≥130/85mmHg
    - 空腹血糖≥6.1mmol/L、餐后2小时血糖≥7.8mmol/L或有糖尿病史

### 常见发病人群分布[3]

* 大于等于40岁者
* 有1项或2项代谢综合征组成成分但尚不符合诊断标准者
* 有心血管病、非酒精性脂肪肝病、痛风、多囊卵巢综合征及各种类型脂肪萎缩征者
* 有肥胖、2型糖尿病、高血压、血脂异常、尤其是多项组合或代谢综合征家族史者
* 有心血管病家族史

### 代谢综合征的主要致病因素

根据我们搜集的资料显示，当今医学界对MetS确切的病因与发生机制众说纷纭，尚未达成一致。它的发生是多个疾病与潜在危险因素交互作用的结果,目前学术界较为公认的发病学说有胰岛素抵抗学说,肥胖学说和炎症学说等等。目前,绝大多数的学者认为MetS发病机制的中心环节为胰岛素抵抗[3]。胰岛素抵抗是指研究对象的肝脏、骨骼肌以及机体的脂肪组织等对胰岛素不敏感,不能充分利用血浆中的葡萄糖。胰岛素抵抗是糖尿病发病的基础环节,此外它还刺激高血压症状的出现和血脂异常的发生,进而导致MetS的发病。

除以上的发病机制以外,MetS的发生也与遗传、生活环境、精神因素、药物、生活习惯、饮食结构、体育锻炼等的因素均密切相关。ChandolaT等人进行的一项针对10308名任职于伦敦政府机关的工作人员的前瞻性的研究发现,长期的工作压力是MetS发生的一个重要危险因素,研究得出,处于长期工作压力状态下超过14年是MetS发病的独立病因[5]。近期也有研究发现,现代生活的高脂高糖饮食能够刺激机体发生炎症反应和氧化应激,从而导致胰岛素抵抗以及MetS的发生[6]。缺乏体育锻炼、多食少动的现代生活方式是MetS近年来发病率上升的重要原因。

# 组学数据研究和分析

## 基本定义

在生物学的研究范畴中，组学数据常见的表现形式有基因组学数据，蛋白质组学数据等。组学数据表示的是生物体相关信息的完整数据库，所含的信息可谓海量。随着计算机技术的爆炸式发展，相应产生的生物信息学是一门将生物和计算机结合的学科。它的研究重点就在于组学数据，尤其是基因组学和蛋白质组学。生物信息学家们将整理好的相关生物信息录入庞大的数据，并运用数学和计算机技术对其进行分析和研究。在这一过程中形成的生物信息数据总和，就是我们通常所说的组学数据。人类基因组数据库就收录了人类的所有基因的各种信息，比如基因的位点，基因的碱基序列等。在当今热门的大数据分析和研究领域，组学数据为新的生物学发现打下了坚实的基础。

## 组学数据的处理方式

根据题目所给出的数据，以及之前我们在NCBI和EBI等数据库进行搜索的结果，我们对数据进行了简单的解读。

* 基因数据

在我们所得到的基因数据中，我们可以清楚地了解到基因的名称，以及基因的变异位点和变异类型，由于样本数量有限，我们对基因的相关信息进行了筛选，删去了一些无关的信息。至于具体操作过程我们后面也会详细讲到。

* RNA数据

我们所得到的RNA数据中用探针表示RNA，虽然与基因不同，但是RNA名称作为一个代号，并不影响我们的实际建模过程，相对于基因数据，RNA数据更容易处理成向量形式，从而为支持向量机的使用提供了许多便利。

* 蛋白组学和代谢组学数据

从上面的讨论我们可以轻松地类推到蛋白质组和代谢组的情况，在此不过多赘述，在题目给出相应数据后，我们可以类比基因数据的处理方式，对这类数据进行规约。

## 组学数据的研究意义

组学数据对于我们此次的研究意义重大，因为现代生物以及医学的新发现往往伴随着对组学数据的充分利用，组学数据能够帮我们跳出思维定势，从数据的关系之间寻找生物学的新发现。特别对于本篇文章要解决的问题来说，组学数据为问题的解决提供了数据基础。生物体的结构是非常复杂的，绝不能理解为简单的单向决定论。与传统的生物学研究不同，依赖大数据技术的组学数据生物信息学更倾向于把复杂的生物系统看作一个“黑箱”，侧重在数据的源头和末端利用计算机和数学手段分析数据之间的联系，从而再依据最终结果反馈，思考新的生物学理论和发现。

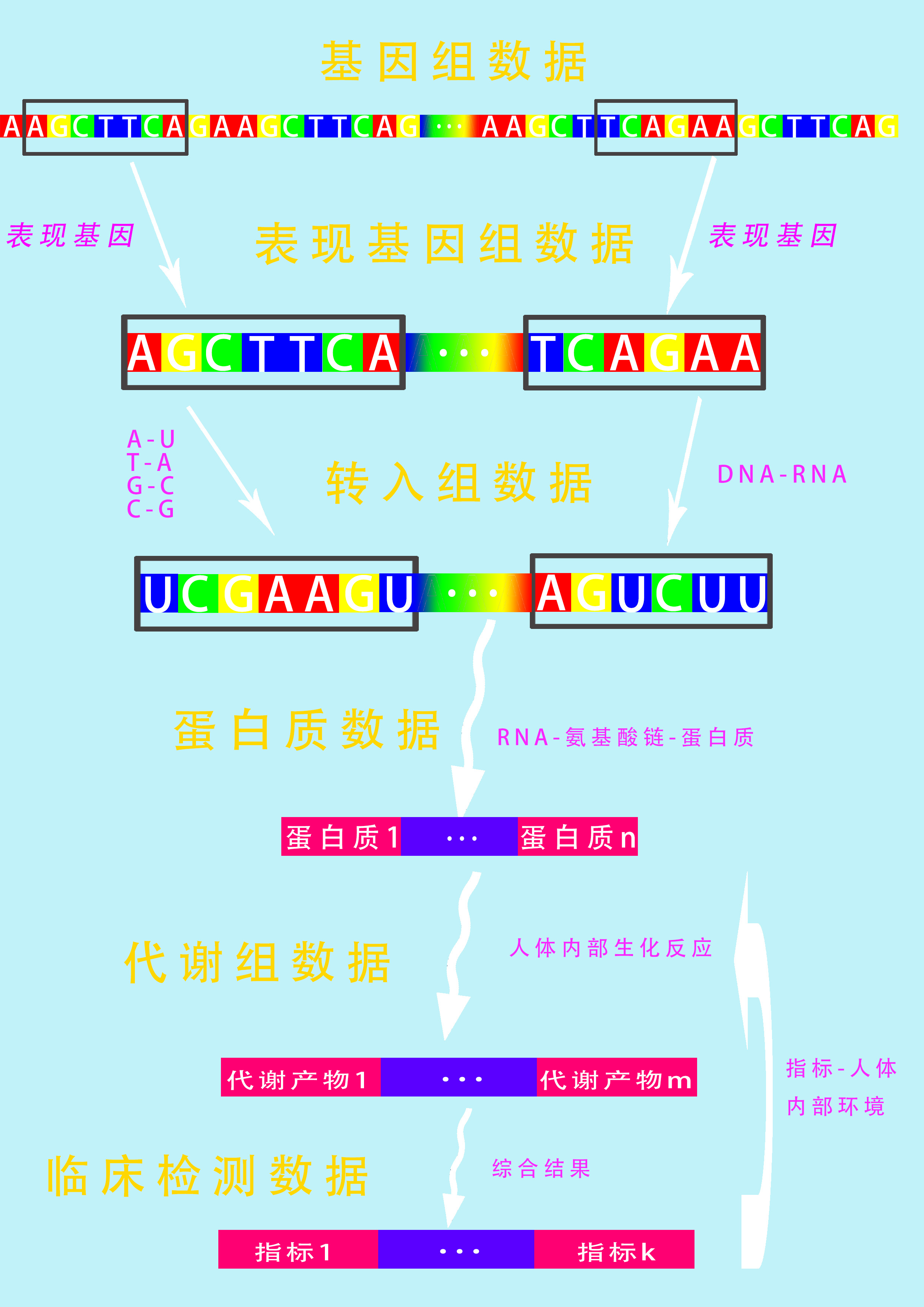
生物医学是当今大数据研究的主要方向之一，它的根基也就是内容丰富的组学数据库，我们建立的模型也以人类代谢综合征相关的组学数据为依托，在处理模型的过程中，暂时抛开既有的生物学框架，用数据分析的手段探索代谢综合征发病过程中的关键基因和蛋白，从而为代谢综合征的早期诊断和预防提供可能。

# 预测模型设计

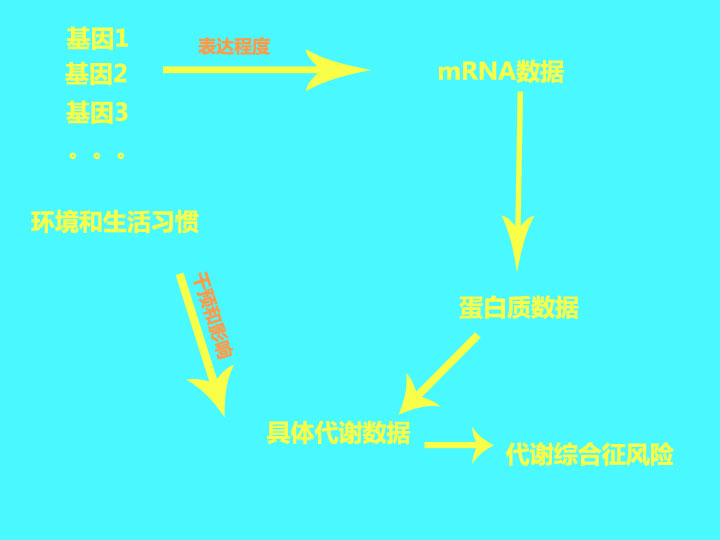
## 数据预处理

首先我们在前面已经讨论过，代谢综合征虽然有很复杂的疾病机理，但是丰富的组学数据为我们建立模型打下了坚实的基础，根据我们已有的基因和RNA数据，我们能够较好地建立模型估计代谢综合征的发病情况。但是，我们也注意到，样本的数量和单个样本蕴含的信息量并不相称，故而我们在建立模型的第一步应该是对我们已有数据进行一定的处理，使之符合我们建模的要求，接下来我们用我们处理好的基因数据和RNA数据，将数据按照测序批次，性别等特征做了多次分组，以便揭示数据的内在联系。另外，实践是检验真理的唯一标准，所以我们没有凭感觉去随便选择一种主流算法，而是在建模过程中，对照了神经网络和支持向量机的运算结果，并选择了效果较好的支持向量机方法。

我们用图表来说明模型的最终实现情况以及操作过程。首先我们用一张图表具体说明题目所给数据能对我们的模型取到的具体作用。即基因影响代谢的主要过程。

图基因影响代谢的主要过程

接下来是模型的建立过程，我们整个建模的思想和依据可以由下图给出，我们的预期目标在于，我们仅通过两个输入端即基因和生活习惯（从某种程度上来说，输入端的数据量要涵盖相关的DNA，所需要的数据量是很大的），来得出最后的输出端口。

图算法的主要结构和流程

那么本质说来，我们模型要涉及的主要内容有以下几点：

* **通过RNA和蛋白数据推测基因变异对表达量的影响**
* **综合基因表达量与基因类别判断不同组学数据对于代谢过程的影响（预测模型的核心，也是本文的重点）**
* **通过对于具体代谢数据的分析预测代谢综合征的风险**

本篇文章的目的即通过数学手段，解决量化模型的核心步骤，即不同组学数据对代谢数据的影响，其余几个部分需要借助进一步的资料和数据来进行模型的构建。接下来我们将具体阐释算法的实现过程。

## 预测模型的建立

### 关联矩阵动态量化模型

1. 数据之间的关系  
   任务a要求构建人类生命量化的动态模型。可以使用的数据有：
   * 临床检验数据
   * 基因组数据
   * 表现基因组数据
   * 转录组数据
   * 蛋白质组数据
   * 代谢组数据

参考图1，我们知道特定的基因对应着特定的氨基酸，而氨基酸的有序组合则构成了蛋白质。另外，转录组数据即为RNA数据。

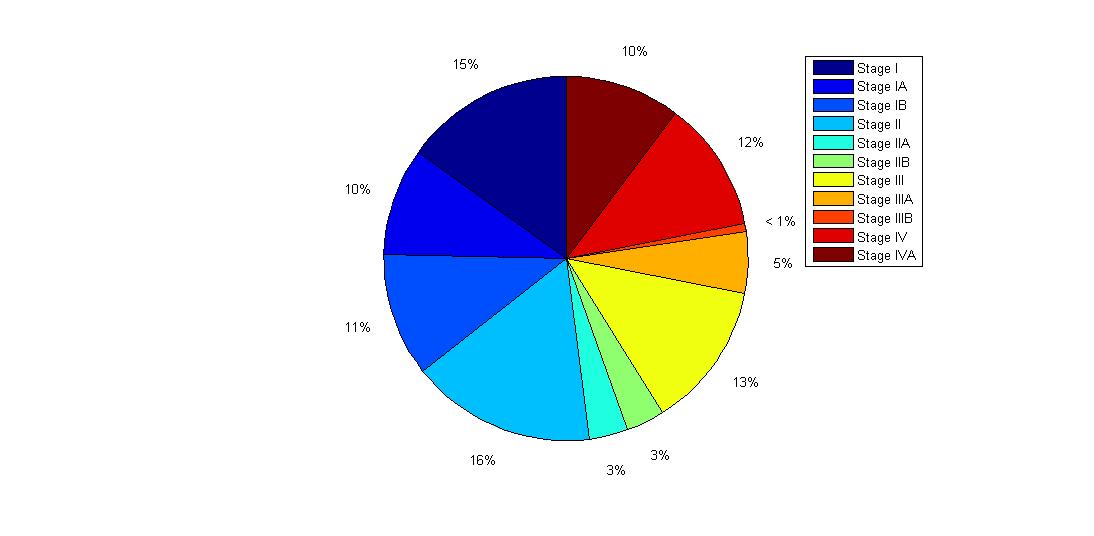
1. 符号规定
2. 建立动态模型
   * 首先，我们可以从一个确定的表现基因出发，根据基因和RNA以及蛋白质等的关联，借助关联矩阵的相关运算，我们可以比较轻易地求得它对所有临床指标的关联向量，我们可以将其记为。
   * 类似地，我们也可以从一个临床检验数据，回溯和该代谢指标相关基因以及RNA和蛋白，并通过关联矩阵，从而求得它对所有临床指标的关联向量，我们将其记为
   * 通过以上方法，任取两组数据，我们可以用类似地方法对关联矩阵操作，从而分析任意两组数据之间的关联度。事实上，经过我们的运算，我们得到的矩阵乘积中的每一个元素，即表示的相应数据之间的关联程度。从而我们用关联度矩阵的办法，构建我们的动态模型。当一个元素发生变化时，相应的关联矩阵也会发生变化，从而搭建了我们所需的动态模型。

### 关键通路的寻找

1. 任务描述  
   结合临床检测数据，哪些因素(数据特征或相互作用网络)是代谢综合症（一种临床诊断结论）的关键通路
2. 数据预处理

题目所给的数据可以分为三组，分别为样本信息，RNA数据以及DNA数据。以下将分别分析这三组数据的特点以及相应的处理方式。

* 1. 样本信息  
     样本信息包含了样本的编号，性别，患病情况以及实验批次。编号为1到229；其中男性107个，女性122个；实验批次分为批次0，111个样本和批次1，108个样本；患病情况分为四大类11小类。  
     根据得到的数据，在所有229个样本中，146个含有患病情况。因此在有监督学习的算法下实际可用的样本数为146。图3为患病情况的总体分布

**图3患病状况总体分布图**

* 1. RNA数据  
     批次0和批次1所记录的RNA数据量略有不同，批次0为20073个，批次1为200168个，两个批次相同的RNA数量为20028个。其中RNA以表达量记录在表格内。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ###probe | 133 | 137 | 73 | 132 | 135 | 136 | 134 |
| ?|100130426 | 0 | 0 | 0 | 0.3178 | 0 | 0 | 0 |
| ?|100133144 | 15.1383 | 8.9519 | 2.9848 | 11.1774 | 1.7573 | 23.7467 | 5.8305 |
| ?|100134869 | 47.5142 | 10.3642 | 3.101 | 17.1078 | 6.0058 | 18.4697 | 6.3991 |
| ?|10357 | 171.4605 | 157.2863 | 187.0512 | 272.5197 | 177.331 | 164.5646 | 206.2188 |
| ?|10431 | 998.3727 | 695.0602 | 703.7344 | 707.1293 | 1480.64 | 823.5488 | 920.2782 |
| ?|155060 | 53.7022 | 42.7486 | 66.39 | 63.5622 | 64.2222 | 137.2032 | 186.8073 |
| ?|26823 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ?|280660 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

* 1. DNA数据  
     相较于给出的RNA数据，DNA数据名目较多，从而需要预先对DNA数据进行一个转化和规约，以便使之适合模型的运算要求。

先对数据做最初的简化，原始数据中出现了一些诸如时间等数据，这些数据是无效的，从而直接进行删除操作，另外数据表格中出现的空白部分我们也将其删除。接下来我们继续分析了基因的数量和其他参数的数目，发现这个样本的数据量仍然较大。所以我们借助程序和自己的判断，对数据进行了进一步的处理，具体过程如下：

通过对原始数据的比对，我们发现，有一些数据对模型的建立是没有帮助，另有一些参数可以用其他参数表示，进而可以删去原始数据中的某些列，从而达到简化数据的目的。由于对于某一个给定的基因来说，它所在的染色体也是确定的，并且在有了基因名字之后，基因ID事实上也可以去掉。dbSNP\_RS列全部为空白，从而我们第一步将三列无关数据删除。

接下来找寻数据之间的联系。对于一个确定的基因，其变异种类往往是确定的，从而将变异种类作为重复数据，直接删除。然而事实上同一个基因发生变异的位点可能不同，从而变异位点应该作为考虑因素。但是观察后发现，同一个位点发生的变异情况是相似的，即在这一位点可能发生的碱基对替换和变异位点之间一一对应。那么我们可以将等位基因种类这一栏删除。

最后对数据进行程序化处理。目前得到的数据以文本为主，并不能很好地应用到模型的计算中，我们需要借助程序将其转化为适合模型预算的形式，类似于我们获得的RNA数据，我们期望将数据处理之后，能通过一个向量来表示样本的基因情况，若该基因正常，则这一分量记作0，则记录其变异位点开始位置。经过处理，我们将数据处理为如下形式。

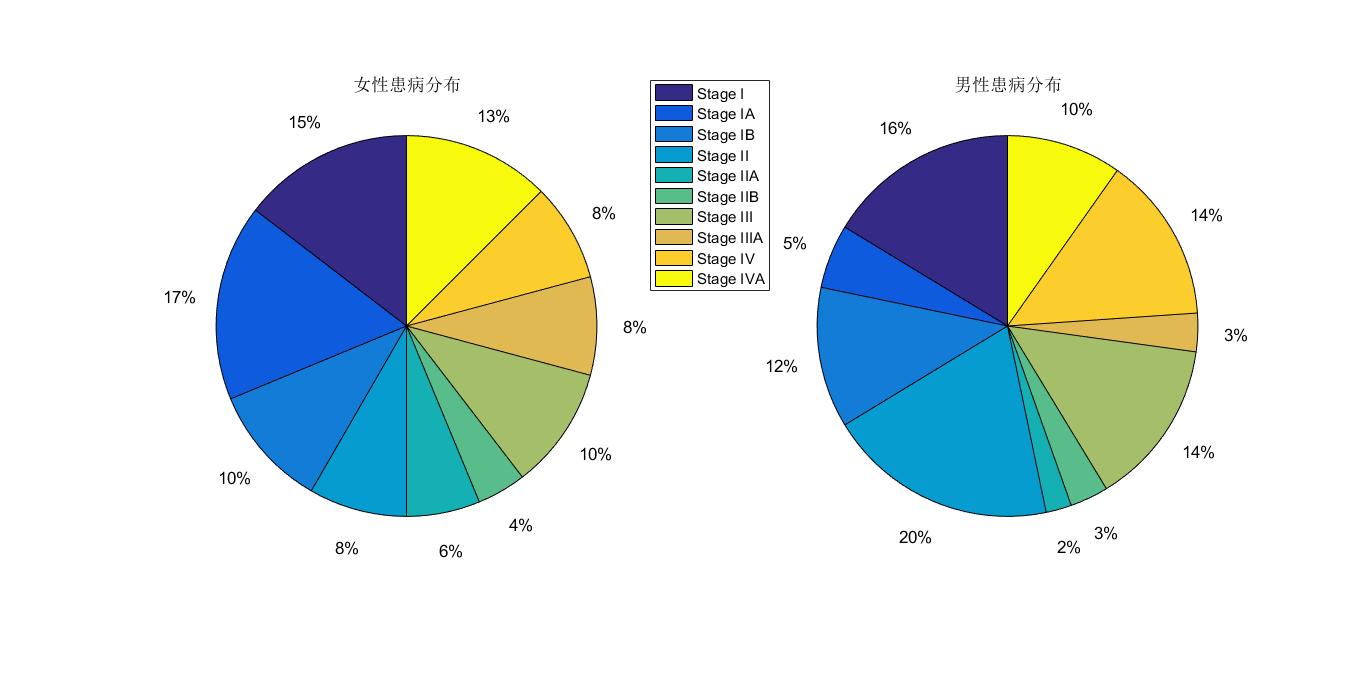
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 基因名称/样本 | 1 | 2 | 3 | 4 | 6 |
| A1BG | 0 | 0 | 0 | 0 | 0 |
| A1CF | 0 | 0 | 0 | 0 | 0 |
| A2M | 0 | 0 | 0 | 0 | 0 |
| A2ML1 | 0 | 0 | 0 | 0 | 0 |
| A4GNT | 0 | 0 | 0 | 0 | 0 |
| AACS | 0 | 0 | 0 | 0 | 0 |
| AADACL2 | 0 | 0 | 0 | 0 | 0 |
| AADACL3 | 0 | 0 | 0 | 0 | 0 |
| AADACL4 | 0 | 0 | 0 | 0 | 0 |
| AAGAB | 0 | 0 | 0 | 0 | 0 |
| AAMP | 0 | 0 | 0 | 0 | 0 |
| AANAT | 0 | 0 | 0 | 0 | 0 |
| AAR2 | 0 | 0 | 0 | 0 | 0 |
| AARS | 0 | 0 | 0 | 0 | 0 |
| AARSD1 | 0 | 0 | 0 | 0 | 0 |
|  |  |  |  |  |  |
| AASDHPPT | 0 | 0 | 0 | 0 | 0 |
| AASS | 121738842 | 0 | 0 | 0 |  |

0表示该基因正常，大于0指的是基因的变异位点。

1. 有效数据的分类

在建立模型的过程中，将只使用在三组数据中都有记录的样本。由于性别不同会造成基因方面较大的差异，而两个批次间检查的项目差异不大；为了确保建立的模型有足够的训练样本且保持良好的准确度，对数据按性别分类而不按照批次分类。

**图4不同组别患病状况分布图**

****

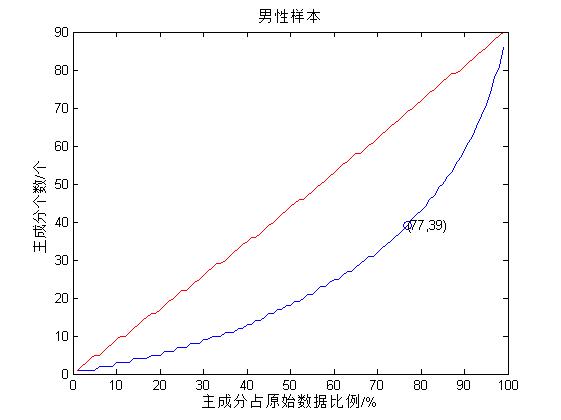
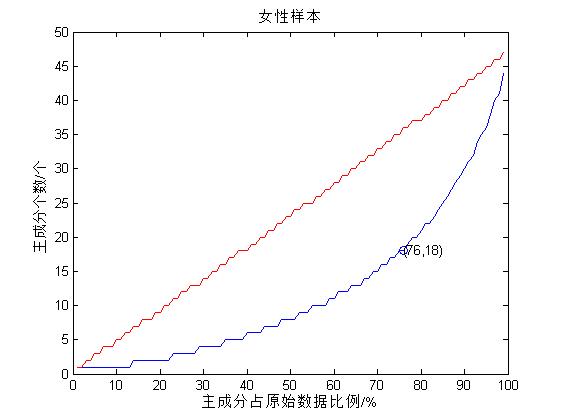
1. 数据的规约[18]  
   由于RNA为数值型数据，量级不同，因此需要对数据进行规约。

转换函数为：

1. 数据的主成分分析  
   对数据进行主成分分析的意义：
   * + 各项生物数据的参数数量很大，如果直接使用原始的参数，可能导致程序的运行速度变慢
     + 生物数据互相之间的关联性未知，使用主成分分析构造新的变量，可以使不同主成分之间的关联度降到最低，而主成分内的变量有较高的关联度，不仅有利于进一步的数据分析，也可以由主成分的构成来确定数据间的相互作用关系
2. 主成分分析的计算步骤(见代码5)[18]
   * + 设原始变量的观测n次数据矩阵为：
     + 将数据矩阵中心标准化，将标准化后的数据矩阵仍然记为X
     + 求相关系数矩阵的定义为：

其中,为第i列向量的均值

* + - 求R的特征方程的特征根
    - 确定主成分个数/,为主成分占原始参数的比例。  
      主成分的个数随着主成分占原始参数的比例而增加。一般取0.85到0.95之间。因为自变量的维数远超出样本数，因此主成分分析的个数收到了样本数的限制，为了确保主成分的选取能够在降维的同时反映数据的特性，作出主成分个数随比例的变化曲线，并与随机生成的一组同样大小的数据进行比较。



因为主成分占比越大，反应数据的特点就越全面；在相同占比下，主成分个数越少，对以后的建模越有利。因此计算随机样本所得曲线与患病样本所得曲线的垂直距离，取能使这个距离最大的点，以此确定主成分的比例。经过主成分分析后，女性样本得到了18个主成分，男性样本得到了39个主成分。

* + - 计算m个相应的单位特征向量：

,

* + - 计算主成分：

1. 关键通路的寻找  
   将处理后的DNA与RNA数据与样本的病程作相关性分析  
   根据数据的特性，使用Pearson相关系数  
   计算公式：  
   若，认为此生物数据i与临床指标j有重要关联

### 根据个人信息预测发病风险

1. 任务描述  
   若给定一个新的人类群体数据集（~10人），包含了每个个体的基因组、表观基因组、转录组、蛋白质组和（或）代谢组的部分测量，请问这些人有多大的代谢综合症风险？造成他们的代谢综合症风险的主要因素分别是什么？
2. 任务分析  
   部分测量的含义是指，存在一部分在任务b中发现和代谢综合征具有强关联的因素，没有被测量到，因而只能通过其他层的生物物质进行推断，补齐数据，用这些数据与临床检验数据的关联，推断代谢综合征的风险。引起代谢综合征的主要因素则可以通过同样的模型逆向去搜寻，下面的关键是采用何种方式来模拟生物物质之间的相互作用关系，而不仅仅是相关性。
3. 建模  
   人工神经网络在预测，分类方面有良好的表现。因此本文首先考虑使用人工神经网络对代谢综合征进行预测。

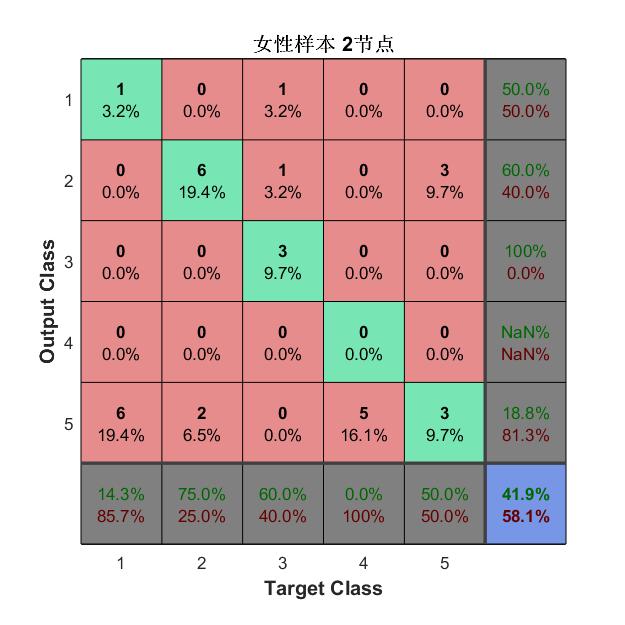
由于神经网络的训练需要将数据集分为训练集，验证集和测试集，因此如果某个疾病严重程度的样本数过少，神经网络将无法学习其中的特征，因此将样本数低于5的疾病程度去除。

在构建神经网络的过程中，隐层节点数目的确定非常重要。虽然目前没有一个确切的公式来确定具体的隐层数目，但是为确保模型的有效性,有以下两条经验原则需要遵守.

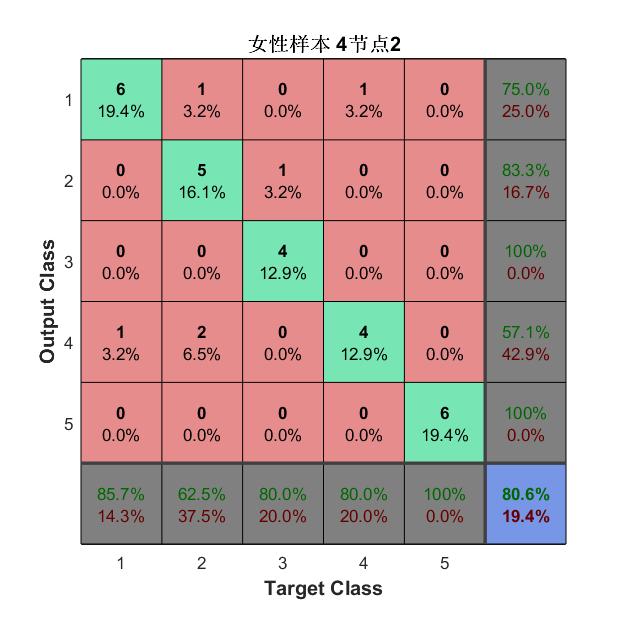
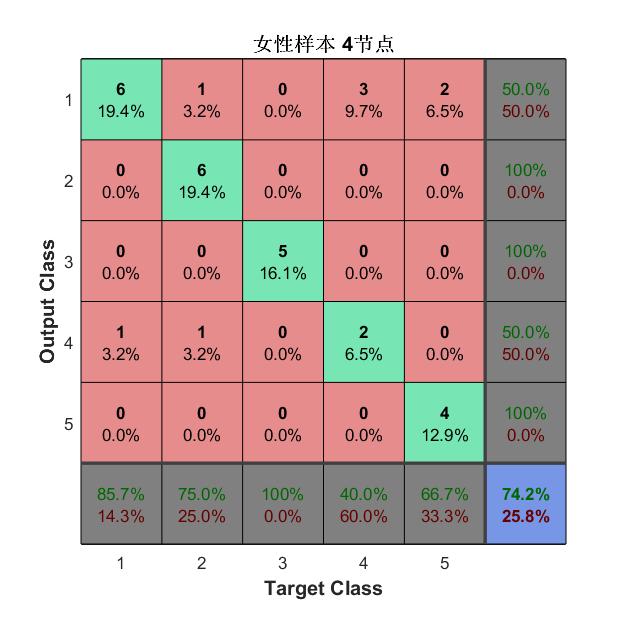
1. 输入节点个数小于样本数
2. 样本数不大于连接权数

设样本数为, 输入节点数为 ,隐层节点数为. 已知输出节点数为1，则上面两个经验等价于：

当输入节点数由主成分分析决定后，隐层节点的个数也大致确定。因为女性样本数为48，主成分个数为18，因此隐层节点数不大于2.当大于2时，由不同初值训练得到的模型不同，缺乏可靠性。



可以从右下角的绿色数字看出，虽然此时的神经网络模型的节点数符合要求，但是识别正确率过低。



采用4个节点后，虽然正确率有上升，但是模型不再稳定。

为了获得更好的正确率以及使模型更加稳定，需要更多的样本数进行训练。

综上所述，虽然神经网络在组内可以得到很好的分类正确率，然而由于样本数过少而需要分类数目过多，在不影响模型鲁棒性的前提下，可以选取的参数非常少，大量的数据只能被舍弃，无法在模型中体现生物系统中的相互作用。在目前小样本高维数的情况下，比较适合的模型为支持向量机。

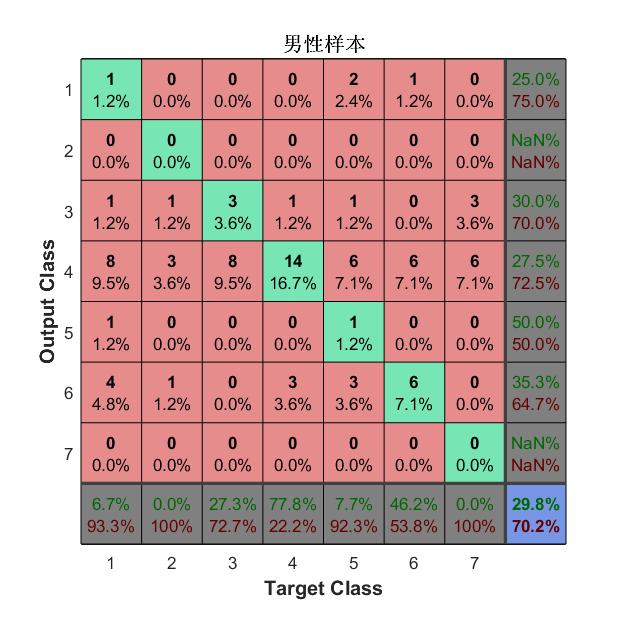
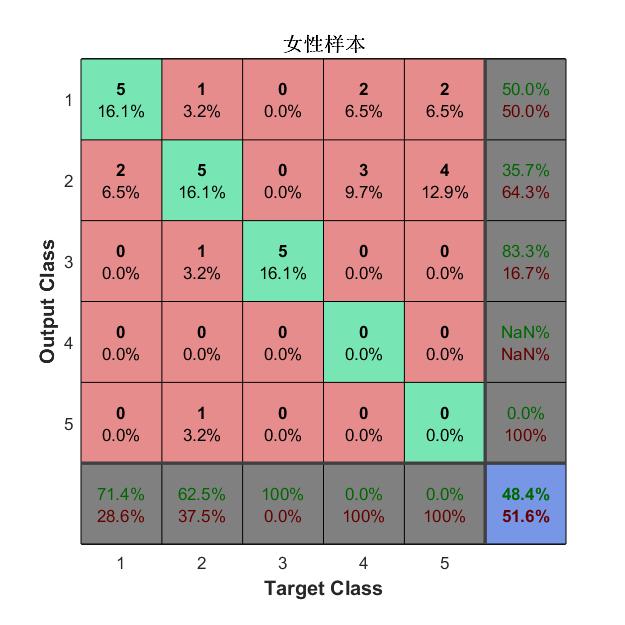
支持向量机通过特殊的核函数，将线性不可分的问题变为线性可分的问题。本文使用“一对一”的分类方法将一般支持二分类问题的支持向量机应用于多分类问题。

该方法在每两类间训练一个分类器，因此对于一个类问题，将有个分类函数。当对一个未知样本进行分类时，每个分类器都对其类别进行判断，并为相应的类别“投上一票”，最后得票最多的类别即作为该未知样本的类别。

使用Matlab中的templateSVM构造支持向量机模板，使用fitcecoc进行多分类的回归。支持向量机的核函数选用高斯函数。因为针对女性样本的疾病严重程度有5类（数据处理方式与神经网络相同），因此总共有10个支持向量机。

由于支持向量机对训练样本的拟合精确度几乎都可以达到100%，所以需要通过交叉验证来判断训练的优劣。具体做法如下：

* 依照90%的比例随机选取训练集，余下的样本作为测试集
* 在训练集上训练支持向量机，然后用测试集进行验证
* 重复以上两部10次，综合计算错误分类率



从图中可以看出，通过支持向量机分类的样本正确率不高。但是较之神经网络有6.5%的优势(48.4%-41.9%)

正确率不高的原因在于样本数量较需要分类的数目仍然过少，若能够在后期得到更多的样本，则可以更好地学习每一类的特征，得到更好的预测正确率。

由于目前的模型无法在识别疾病严重程度上做到很高的正确率。因此在判断一个人有多大风险患上哪一个阶段的代谢综合征只做理论推导。

已知代谢综合征共有11种严重程度，因此建立完全的多分类支持向量机模型就需要个支持向量机。将足量的男女样本代入模型后进行训练，得到 , , , , , , ,共55个支持向量机。现取一个新的样本，分别代入这55个支持向量机中, 得到其中的取值依照的特征为或者按照前文所述的投票方式，记录每个严重程度所获得的票数，将其除以55便得到获得这种严重程度的概率=

# 进一步研究方向

由于目前学界对于代谢综合征的发病机理没有定论，我们依据有限的数据得到的结果未必能较为直观地反映代谢综合征的具体发病机理，但是和前人的工作进行对比之后，我们认为我们仍然有一些可取之处。

1. 模型采用的数据分析手段多样，并且对问题有很好的针对性。我们在建模过程中采用的建模手段并不单一，而是综合了神经网络方法和支持向量机方法。这使得我们可以根据模型的输出结果选择更为贴近代谢综合征实际情况的模型，从而提升了模型的精确性与针对性。
2. 模型很好地满足了题目中对动态量化模型的要求，对于发病过程十分复杂的代谢综合征，我们通过对代谢综合征的解构，针对具体代谢数据进行研究，用关联矩阵直观地表现出了指标的相应变化。
3. 模型对于代谢数据的评估可以很好地在早期预测一个人某种代谢疾病的发病几率以及发病原因，通过基因分析和蛋白分析，可以在早期对于一个人某方面代谢数据异常做出较为准确的预测，从而对特定的个体实施个性化的预防。这应该是本模型最大的优势，模型的建立也为疾病的早期预测和干预给出了优秀的范例。
4. 模型具有较大的学习空间。由于模型的建立过程中数据来源是有限的，在获得更多数据的情况下，能不断地用更大的数据样本对模型进行训练，从而给出更精确的结果。另外，还有一点，模型可以加入时间序列的分析，从而除了早期预测之外，模型还可以针对个体进行长期跟踪，通过其代谢评估的变化趋势及时避免代谢综合征的发生。

通过对基因数据的分析，我们很好地对生物体的代谢过程做了量化评价，虽然目前的结果已经出乎我们在模型建立前的预料，但是我们必须要承认，模型的改进空间是无限的。有两方面原因，一是生物学中不断出现的个例会对既有的生物学理论做出很多修正，另一方面是我们需要参考的生物学因素是无穷无尽的，我们并没有获得病人生活习惯的直接数据，这也在很大程度上影响了我们建模的精确性。但是，麻雀虽小五脏俱全，我们在本篇论文中详细地展示了对于代谢调节量化分析的思考和分析。相信在获得进一步数据和资料之后，我们可以充分完善我们预期目标的各个分支，将其整合为一个完整的系统，并提供足够的数据以供模型学习，不断提升模型的精确度，为从数学角度分析生物问题提供一个新的方向和参考。

# 参考文献

1. IDF代谢综合征全球共识定义(2005)·中国成年人代谢综合征的患病率，中华糖尿病杂志2005年第13卷第3期
2. 宋秀霞，纪立农：国际糖尿病联盟代谢综合征全球共识定义，《中国糖尿病杂志》, 2005, 13(3).
3. 姚春莉,刘媛.代谢综合征病因及发病机制研究进展[J].实用医学杂志,2008,24( 20) :3453·3454.
4. HaagM, DippenaarNG.Dietaryfats,fatty acids and insulin resistance:short review of a multifaceted connection [ J ] . Med SciMonit ,2005,11(12):A359-A367.
5. ChandolaT,Brunner，E,MarmotM.Chronicstress at work and the metabolic syndrome : prospective study[ J] .BMJ,2006,332(7540) :521·525.
6. DandonaP,Aljada A, Chaudhuri A ,et al. Metabolic syndrome: a comprehensive perspective based on interactions between obesity, diabetes, and inflamation[J]. CirCulation,2005,1 1 l(1 1):1448-1454.

1. <http://www.ncbi.nlm.nih.gov/gene/?term=metabolic+syndrome>
2. http://www.ebi.ac.uk/ena/data/search?query=metabolic+syndrome
3. Laura J. Scott,1 Karen L. Mohlke,2，etc. :A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants.
4. Eleftheria Zeggini,1,2\* Michael N. Weedon3,4\* Cecilia M. Lindgren,1,2\*etc.: Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes.
5. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for Bio Medical Research :Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels.
6. Eleftheria Zeggini1,10, Laura J Scott2,10, Richa Saxena3–8,10 & Benjamin F Voight3–5,7,10, for the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium9Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.

1. <https://en.wikipedia.org/wiki/Genome>
2. "BMI Classification". Global Database on Body Mass Index. World Health Organization. 2006. Retrieved July 27,2012.
3. Applied Logistic Regression: Edition 3David W. Hosmer, Jr. Stanley LemeshowRodneyX.Sturdivant February 26, 2013,John Wiley & Sons
4. 张清浪，李先民，一种确定神经网络隐层节点数的新方法《吉首大学学报:自然科学版》, 2002, 23(1):89-91
5. Support Vector Machines for Pattern Classification, Shigeo Abe March 30, 2006Springer Science & Business Media

1. <https://en.wikipedia.org/wiki/Structural_equation_modeling#Advanced_uses>
2. 温正，精通MATLAB智能算法，清华大学出版社,2015.