第6讲 MATLAB数据挖掘项目实例

卓金武
MathWorks 中国

课程介绍

- **第1讲：MATLAB快速入门**
- MATLAB快速入门实例
- MATLAB实用操作技巧
- MATLAB数据类型
- MATLAB程序结构
- MATLAB编程模式
- MATLAB学习理念

- **第2讲：MATLAB数据挖掘基础**
- MATLAB数据挖掘的过程
- 数据的可视化
- 数据的预处理
- 数据的探索
- 假设检验
- 数据回归

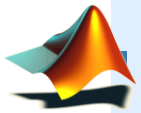
- **第3讲：MATLAB数据挖掘算法（上）**
- 回归算法
- 关联算法
- 聚类算法

- **第4讲：MATLAB数据挖掘算法（下）**
- 分类算法
- 预测算法
- 异常诊断算法

- **第5讲：MATLAB高级数据挖掘技术**
- MATLAB分类学习机
- 算法的高级使用方法
- 综合使用实例

- **第6讲：MATLAB数据挖掘项目实例**
- 故障诊断
- 生物信息学研究
- 量化投资

内容提要



故障诊断

- 生物信息学研究
- 量化投资

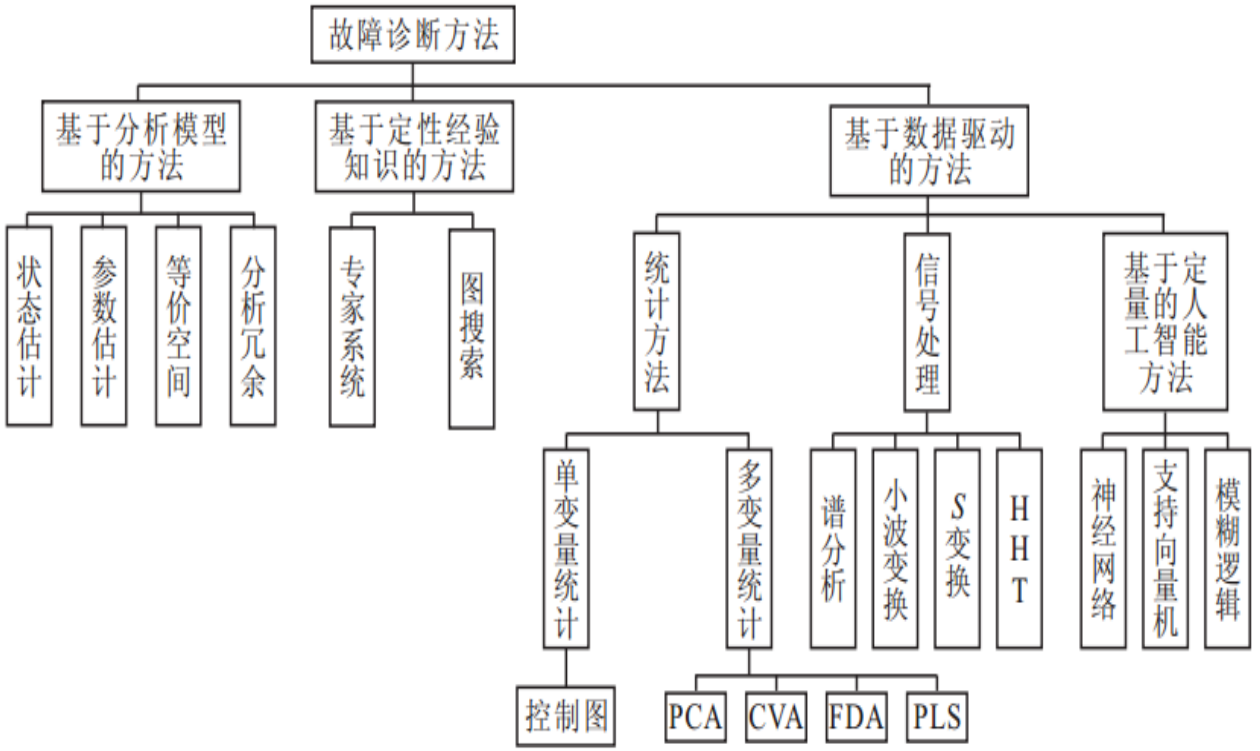
故障诊断的概念

故障诊断，又称为故障分析，是指为了确定故障原因以及如何防止其再次发生而收集和分析数据的过程。

◆ **基于分析模型的方法：**适用于能建模、有足够传感器的“信息充足”的系统，需要过程较精确的定量数学模型，而要建立过程的数学模型则必须了解过程的机理结构。

◆ **基于定性经验知识的方法：**适用于不能或不易建立机理模型、传感器数不充分的“信息缺乏”的系统。

◆ **基于数据挖掘的方法：**多数企业每天都产生和存储较多运行、设备和过程的数据，这些数据分为正常条件下和在特定故障条件下收集的数据。利用这些数据，利用数据挖掘技术实现故障建模及预测的方法。



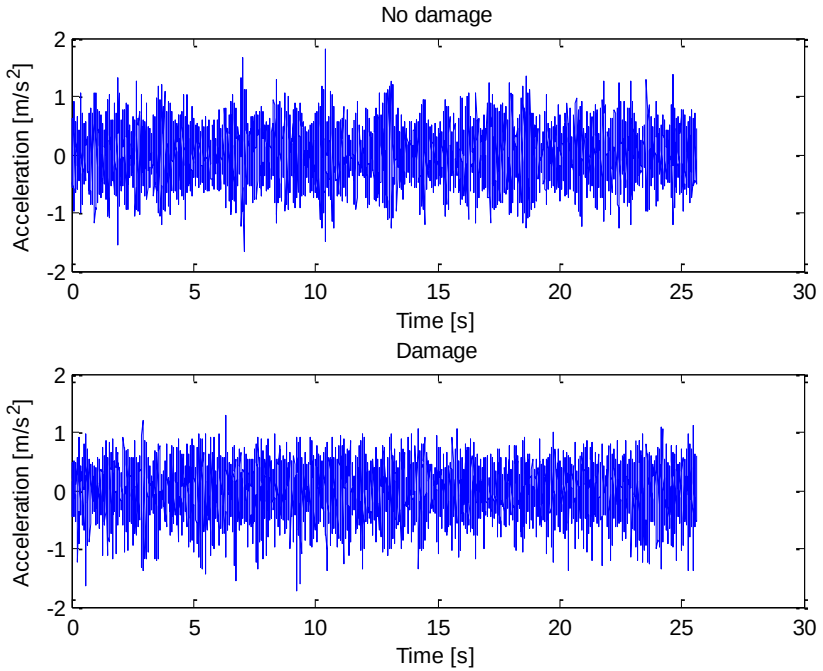
关于数据

案例中的数据，是关于170个设备的监控数据，每个设备有5个监测位（频道），每个位置都有一段时间的某个指标的测量数据。

探索数据

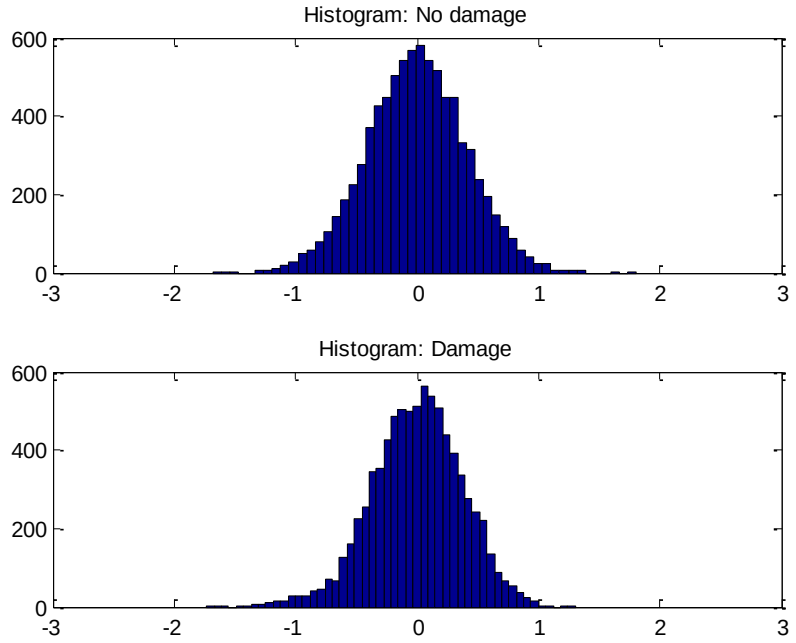
（1）显示时间序列

绘制设备监控数据的时序图（见下图），很难发现它们有什么不同。



（2）显示柱状图

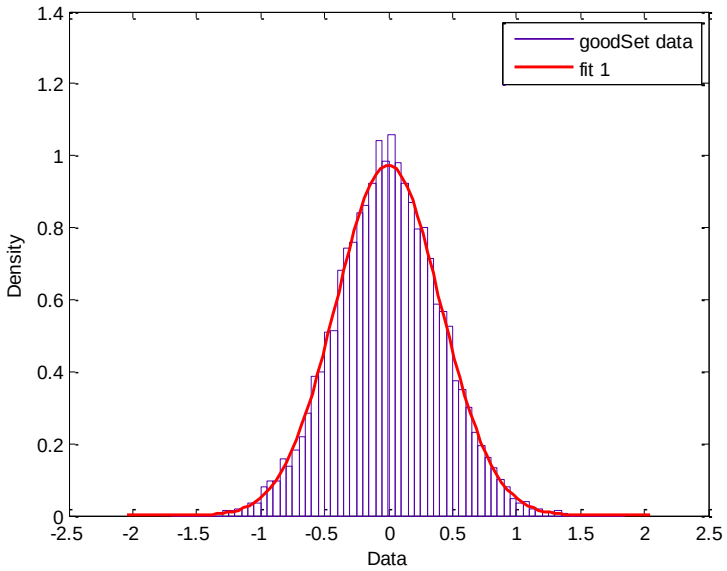
绘制这些时序数据的频次图（见下图），发现好坏设备的图形有差异。



探索数据

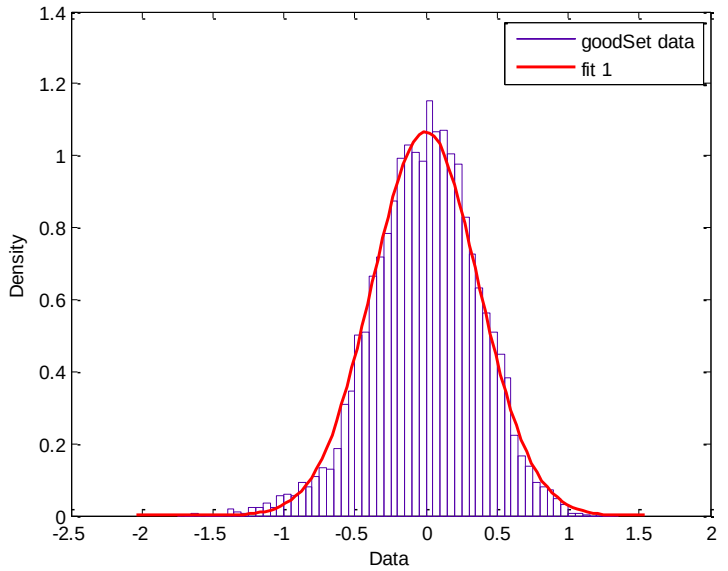
（3）量化柱状图信息

对频次图进行分布拟合，可以得到正常设备（左图）和有缺陷设备（右图）的分布曲线及对应的参数。



参数：

Normal distribution
mu = -0.00313008 [-0.0119913, 0.00573117]
sigma = 0.409145 [0.402975, 0.415509]

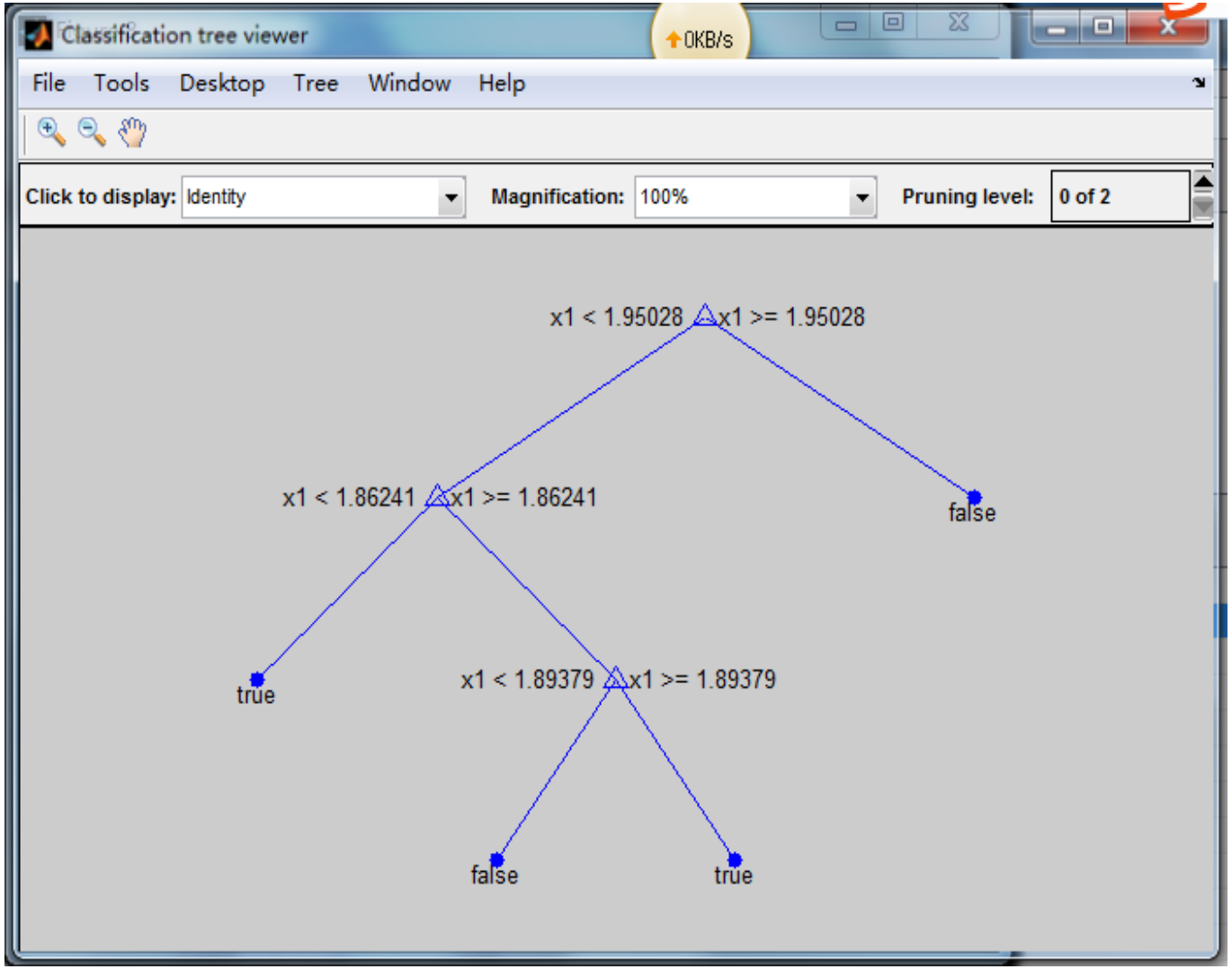


参数：

Normal distribution
mu = -0.00331835 [-0.0114208, 0.0047841]
sigma = 0.37411 [0.368468, 0.379928]

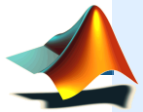
训练模型

在所有的分类方法中，对于特征变量的识别，决策树算是一种比较好的方法，所以先用决策树方法来训练分类模型，所得的决策树模型如右图所示。



内容提要

- 故障诊断



生物信息学研究

- 量化投资

生物信息学的内容

生物信息学是将计算机科学和数学应用于生物大分子信息的获取、加工、存储、分类、检索与分析，以达到理解这些生物大分子信息的生物学意义的交叉学科。

这一定义主要包含三层意思：

- 需要利用计算机及其相关技术来进行研究；
- 对海量数据进行搜集、整理；
- 分析这些数据，并从中发现新的规律。

另外，从基因分析角度来讲，生物信息学主要是指核酸与蛋白质序列数据、蛋白质三维结构数据的计算机处理和分析。

生物信息学近几年获得突破性进展，随着基因组研究的进展，积累了各种大量的生物数据，提供了揭开生命奥秘的数据基础。而随着生物数据种类的丰富以及数据量的增大，如何更有效地的处理、挖掘、分析和理解这些数据日益迫切。

数据挖掘技术在生命科学中的作用

（1）蛋白质结构预测

蛋白质结构预测主要包括二级结构预测和三级结构预测。近年来，神经网络和支持向量机在蛋白质二级结构的预测中有较好的效果。遗传算法在三级结构中应用较多。

（2）微阵列数据分析

数据挖掘在微阵列数据分析中的主要应用有：①基因的选取，即如何从成千上万个基因中选择与需要分析的任务最相关的基因；②分类和预测，即根据基因的表达模式对疾病进行分类；③聚类，即发现新的生物类别或对已有的类别进行修正。

（3）DNA序列相似搜索与比对

DNA序列间相似搜索与比对是基因分析中最为重要的一类搜索问题。这个研究主要是搜索、比对来自带病组织和健康组织的基因序列，比较出两者的主要差异。主要过程是，首先，从两类基因中检索出基因序列，然后找出每一类中频繁出现的模式。

（4）生物数据可视化

可视化应用的主要需求有以下三点：①进行序列操作和分析的图形用户界面，通过便捷的桌面工具进行数据的浏览和与数据间的互动；②专门的可视化技术，灵活运用图形、颜色和面积等方法对大量的数据进行描述，最大限度地利用人类的感官对特征和模式进行挑选；③可视编程，属于特殊的、高级的、领域专有的计算机语言的图形描述算法。

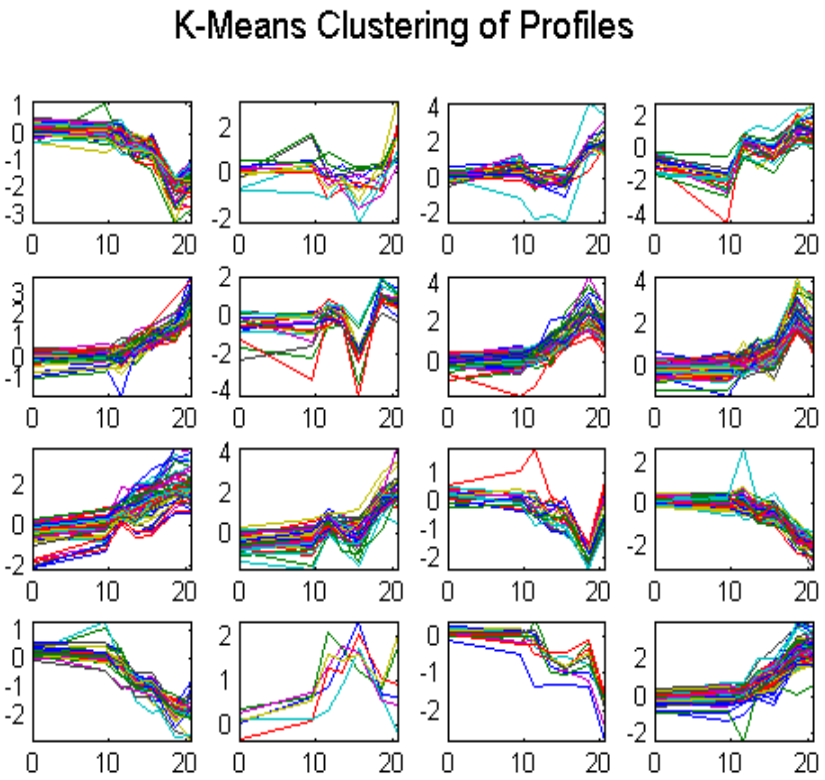
关于数据

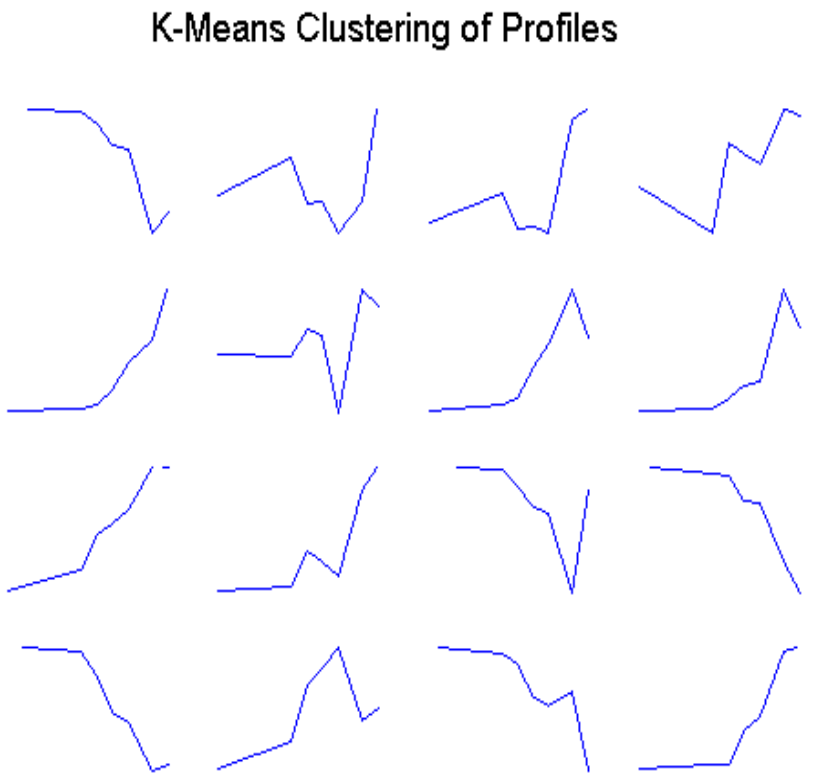
所用的数据是1997年DeRisi在研究酵母基因表达时所得的数据。当时，DeRisi用DNA微阵列研究酵母在新陈代谢过程中的临时基因表达，并在酵母的生长过程中的7个不同时间点对基因表达水平进行测量，从而得到这些数据。

这些数据可以在基因表达的综合研究网站下载，网址为：
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28。

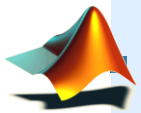
基因聚类

用K-means方法得到的模式图





内容提要

- 故障诊断
- 生物信息学研究
-  量化投资

量化选股的概念

量化选股，简言之就是所有通过计算机软件程序进行买卖股票。

优点：

（1）计算机能够持续稳定、精确严格地按原则工作，能够大规模地进行数据处理，而人灵活有余、原则不足且不能长时间地机械操作。

（2）贪婪、恐惧等是人的天性，犯了错误也不愿意纠正，而计算机会按照既定的规则去处理错误信号发出的指令和生成的持仓。

（3）市场有着无可比拟的高效率和丰富的市场机会（短线、中线、长线甚至T+0），由于对行业和品种认识的局限性，自然人不能精通每一个品种，而每个品种都有活跃期和萎靡期，只有选择在活跃期跟踪交易这个品种，我们才能取得良好收益，有了捕捉市场趋势的程序就能很好地解决这一问题。

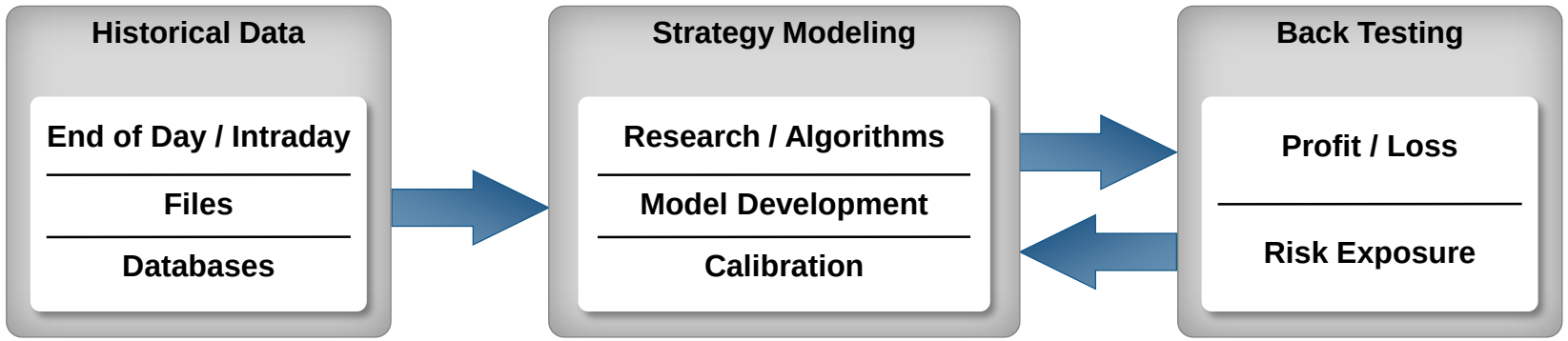
缺点：

（1）大部分量化选股系统都是为了追随趋势而编写的，比较注重技术分析，但技术分析一般是滞后于价格变化的，这样就会导致在区间震荡行情中如果进行频繁交易则就可能会出现连续亏损的现象。

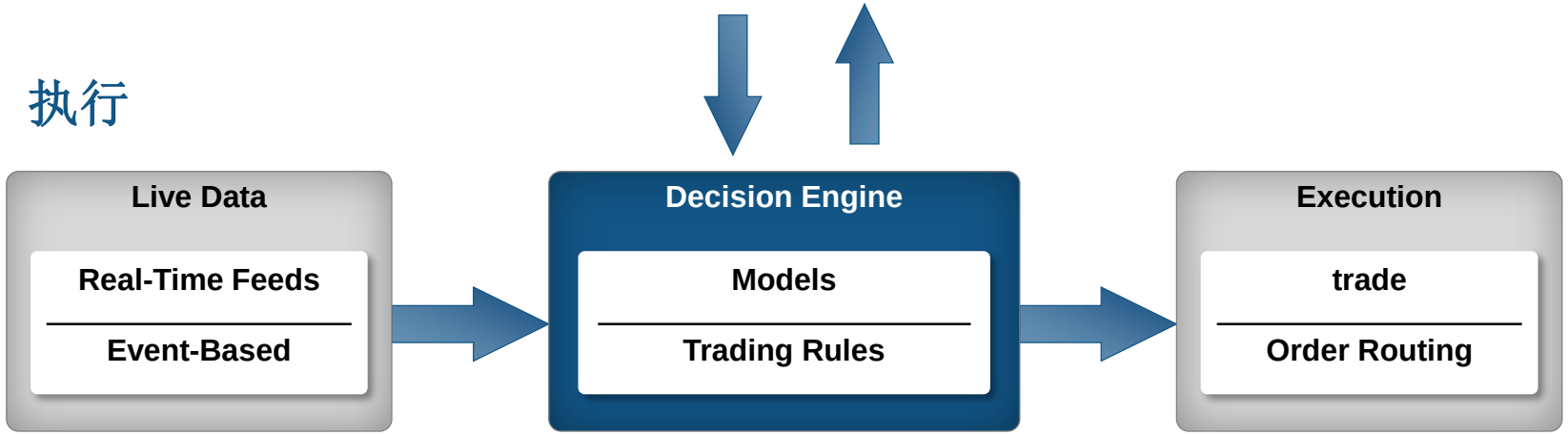
（2）难以确定头寸规模的大小。

量化选股实现过程

开发和测试



执行

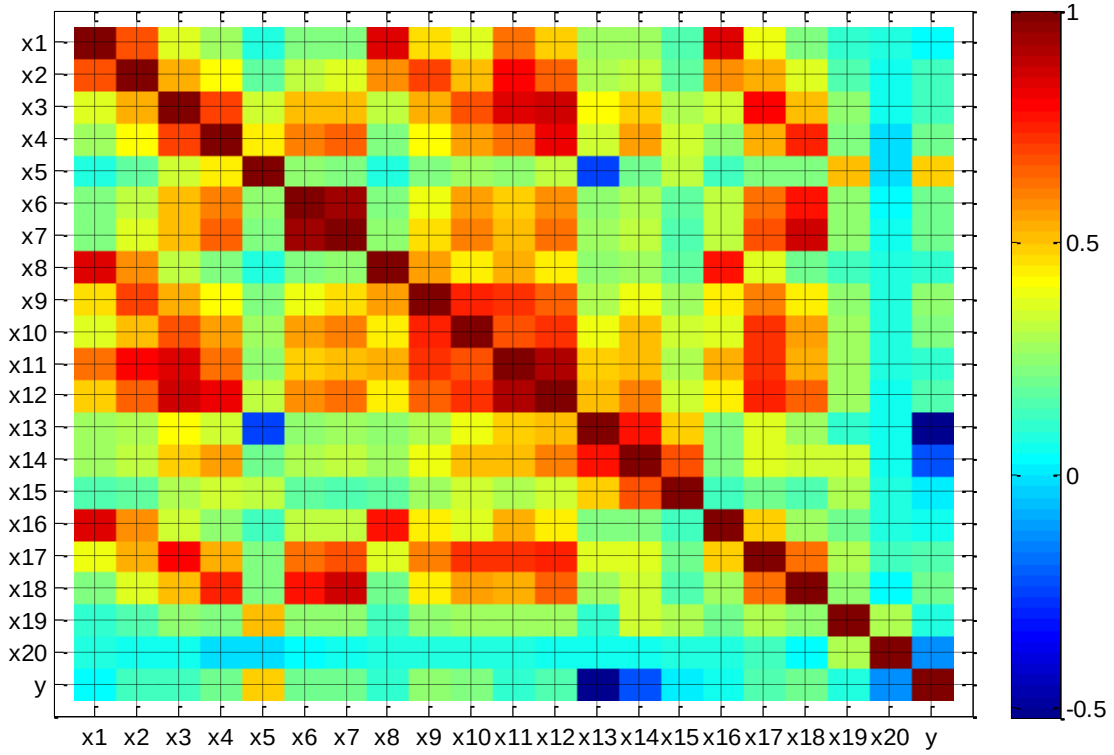


计算指标

指标标记	指标名称	计算方法
s_x1	当日涨幅	(当日收盘价 - 前第n日收盘价)/ 前第n日收盘价 ×100%
s_x2	2日涨幅	
s_x3	5日涨幅	
s_x4	10日涨幅	
s_x5	30日涨幅	
s_x6	10日涨跌比率ADR	10日内股票上涨天数之和 / N日内股票下跌天数之和
s_x7	10日相对强弱指标RSI	RSI = 100 × RS(1 + RS) RS = n日的平均上涨点数/n日的平均下跌点数
s_x8	当日K线值	(收盘价 - 开盘价)/(最高价 - 最低价)
s_x9	3日K线值	(收盘价 - 3日前开盘价)/(3日内最高价 - 3日内最低价)
s_x10	6日K线值	(收盘价 - 6日前开盘价)/(6日内最高价 - 6日内最低价)
s_x11	6日乖离率（BIAS）	乖离率 = [(当日收盘价 - 6日平均价)/ 6日平均价] ×100%
s_x12	10日乖离率（BIAS）	乖离率 = [(当日收盘价 -10日平均价)/10日平均价] ×100%
s_x13	9日RSV	(n日收盘价 - n日最低价)/(n日最高价 - n日最低价) ×100%
s_x14	30日RSV	
s_x15	90日RSV	
s_x16	当日OBV量比	n日OBV / 5日OBV
s_x17	5日OBV量比	
s_x18	10日OBV量比	
s_x19	30日OBV量比	
s_x20	60日OBV量比	
s_y	分类指标	根据未来1日与3日涨幅来确定s_y为1或-1

变量筛选

执行程序，会得到变量间的相关系数矩阵及相关系数图，从该图可以看出，x1-x20与y的相关性有显著差异。

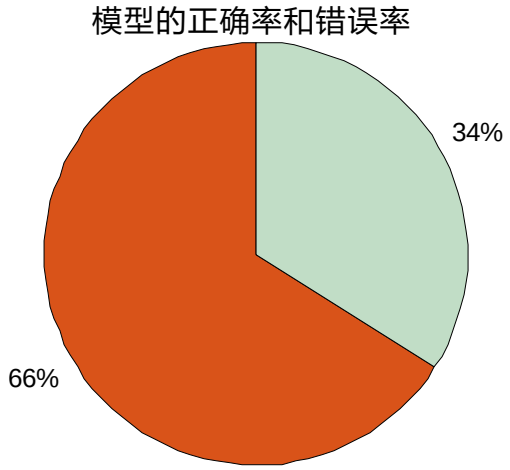


设定一个相关系数阈值，可由这个阈值来确定选哪些变量，这里取0.2，这样相关系数的绝对值大于0.2的变量都被选中。

10	0	0.069164	0.302117	0.173961	0.285799	0.19314	0.128049	0.063319	1
16	0.737536	0.612279	0.361291	0.341493	0.460639	0.706977	0.453238	0.62247	1
17	1	1	1	1	0.336484	0.417322	0.611387	0.754004	1
18	0.642994	0.573002	0.77472	0.844089	0.78096	0.84974	0.426179	0.380197	1
25	1	1	0.77472	0.844089	0.673976	0.843547	0.539033	0.662783	1
25	1	0.923603	0.77472	0.844089	0.459012	0.589257	0.40598	0.529239	1
25	0.824973	0.930236	0.568177	0.676557	0.854067	0.709632	0.311905	0.40163	1
28	0.776759	1	0.444087	0.509025	0.636303	0.807272	0.44011	0.627117	1

模型的训练及评价

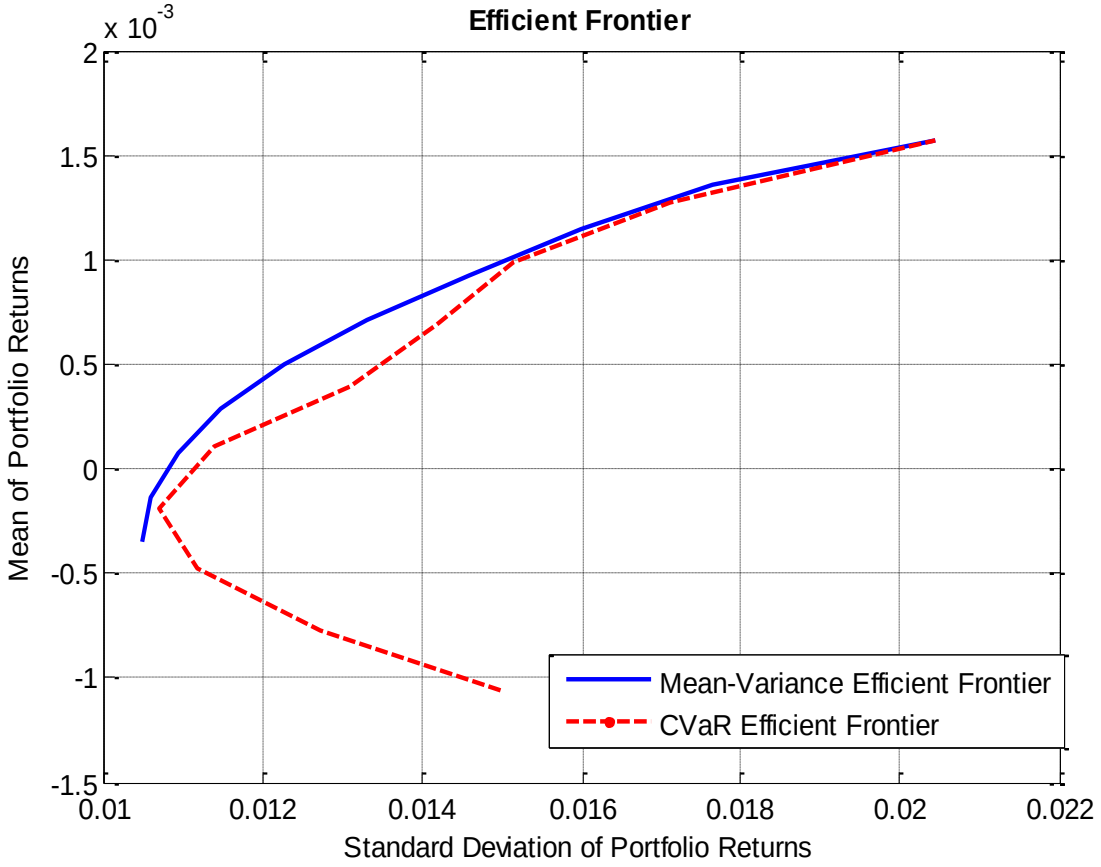
模型分类的正确率和错误率

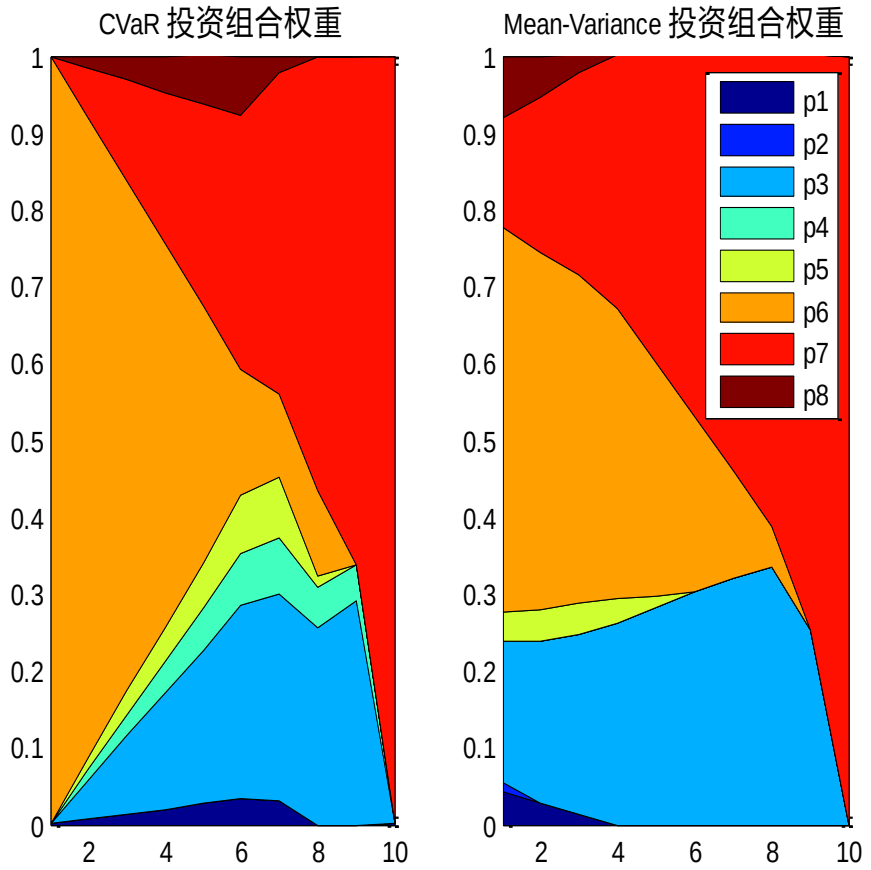


这个结果的作用的是，在实际股票买卖过程中，我们可以选择排名靠前的股票买入，反之卖出，这就提供了量化选股中买入和卖出的条件。

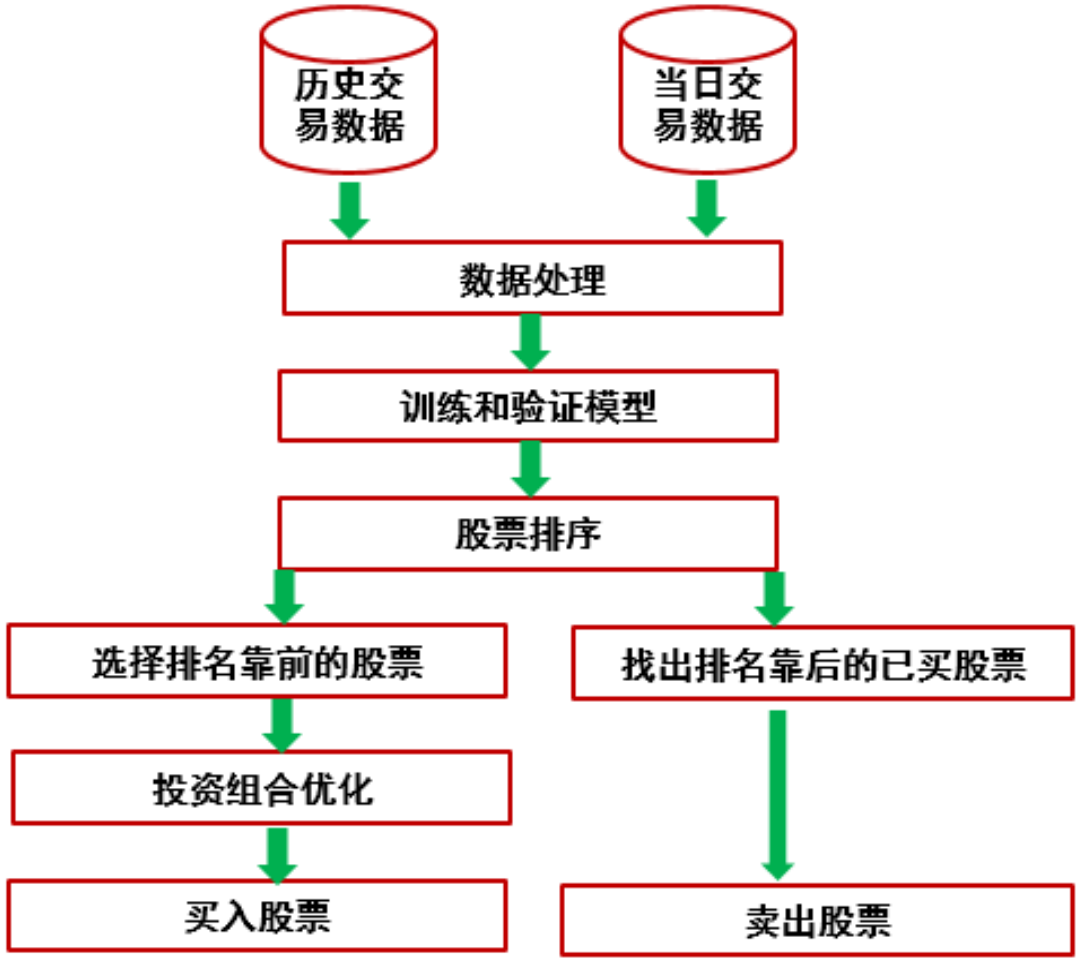
65	1	1	1	1	0.217464	0.689387	0.615622	0.933314	1.076462
802	0.649562	0.714952	0.590378	0.669138	0.533305	0.493489	0.119175	0.450005	0.995385
985	0.489474	0.388007	0.219643	0.032438	0.289402	0.922103	0.458649	0.370715	0.985637
582	0.350914	0.507703	0.590378	0.669138	0.58377	0.410922	0.118595	0.226798	0.940392
66	0.846695	0.593295	0.590378	0.669138	0.551252	0.670699	0.293865	0.605941	0.885136
751	1	1	0.87818	0.881371	0.332703	0.595813	0.626997	0.948292	0.88133
707	0	0.650724	0.302097	0.244671	0.699561	0.544556	0.236814	0.403214	0.830667
819	1	0.888569	0.87818	0.881371	0.613117	0.822664	0.666953	0.978776	0.826818
522	0.343439	0.942634	0.417467	0.456905	0.029334	0.000374	0.035146	0.607885	0.778539
521	0.710836	1	0.302097	0.244671	0.396943	0.315258	0.393372	0.913728	0.75364

组合投资的实现


Efficient Frontier
x 10^-3
Mean of Portfolio Returns
Standard Deviation of Portfolio Returns
Mean-Variance Efficient Frontier
CVaR Efficient Frontier


CVaR 投资组合权重
Mean-Variance 投资组合权重
p1
p2
p3
p4
p5
p6
p7
p8

量化选股的实施


历史交易数据
当日交易数据
数据处理
训练和验证模型
股票排序
选择排名靠前的股票
找出排名靠后的已买股票
投资组合优化
买入股票
卖出股票

MATLAB 学习资源

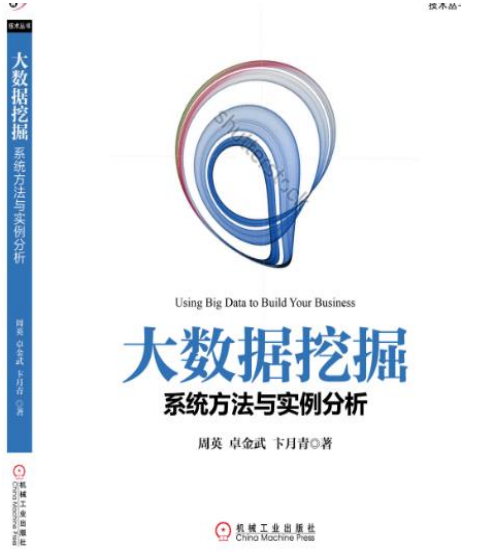
- **www.mathworks.com**
  - 录制的讲座
  - 行业解决方案
  - MATLAB central
- **www.ilovematlab.cn**
  - 问题交流
  - 图书

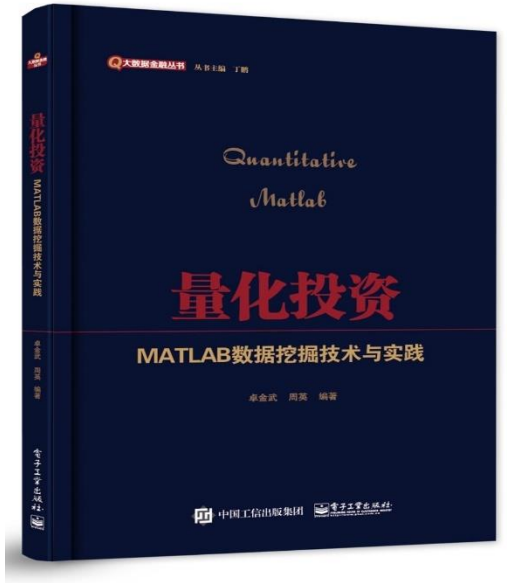
  《大数据挖掘：系统方法与实例分析》

  《量化投资：MATLAB数据挖掘技术与实践》
- 购买正版MATLAB

  电话：010-59827000
- 答疑方式

  邮箱： 70263215@qq.com





关注MATLAB微信公众号，获取更多官方资讯！

关注MATLAB官方微信平台，发送你感兴趣的关键词，即可查看**MathWorks在线资源**。

MATLAB

实践与资源

第1讲：数据与程序	http://pan.baidu.com/s/1boGzSwn
第2讲：数据与程序	http://pan.baidu.com/s/1dELf87f
第3讲：数据与程序	http://pan.baidu.com/s/1c1Fcu5M
第4讲：数据与程序	http://pan.baidu.com/s/1jIbll2Y
第5讲：数据与程序	http://pan.baidu.com/s/1b8vCAE
第6讲：数据与程序	http://pan.baidu.com/s/1jI8vRoi

谢谢大家！