基于聚类思想的特征提取以及贝叶斯网络的实时股票波动预测模型开发文档

1. **需求分析**

通过当天股价、资金的进出情况，通过公司公布的历史数据，如利润表、资产负债表、现金流表、业绩预告信息、财务审计信息等，对个股是否有涨幅进行概率计算。

1. **概要设计**

将聚类方法与贝叶斯算法相结合，先将股票分类，再进行预测等数据处理，期望得到 满意的结果。

首先通过第三方数据平台获取数据，并观察数据类型以及噪声点、查看是否有缺失值等，确定下一步的数据清洗的大体工作。

使用pandaSQL包以及pandas包对原始数据进行清洗，将噪声点、缺失值、错误值、重复值等全部处理干净，以便于下一步的数据特征提取与聚类算法对特征进行分类。

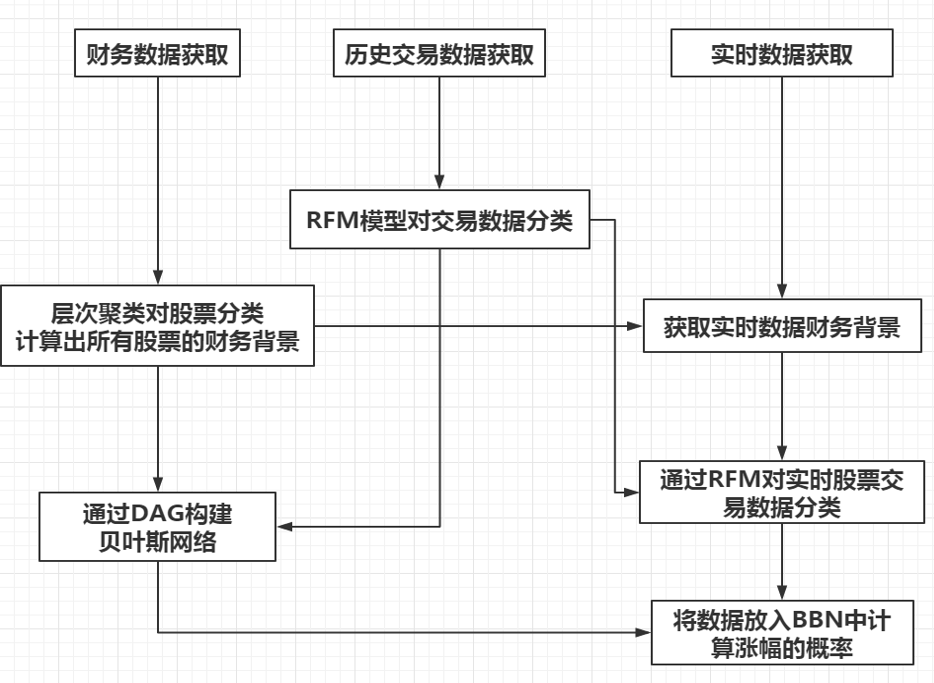
基于 k-means 以及层次聚类规则，对历史数据进行分类筛选，通过部分手工标注，提取相关特征并对股票进行分类。将离散型分布的特征转换成多项式分布，以便于下一步通过贝叶斯网络进行数据的建模。

基于业务构建有向无环图（DAG），通过pybbn、network实现贝叶斯网络的构建，并且输入原始数据，拟合贝叶斯网络。

通过对当天实时（t+0）数据进行测试，预测当天（t+0）是否涨幅能超过 5%。

1. **系统设计**

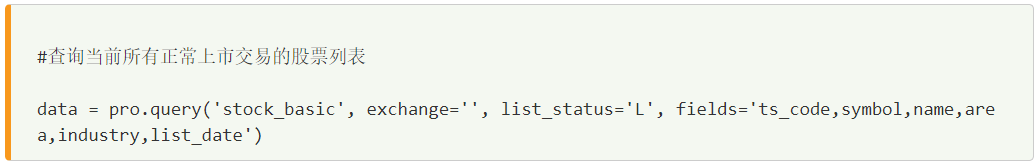
整体开发流程设计：



1. 数据获取

通过第三方开源股票数据平台tushare获取原始数据。Tushare的数据大多来自于新浪财经，可靠性有一定保障，并且是开源、免费版本，对于新手以及入门量化开发非常友好。

数据获取演示：



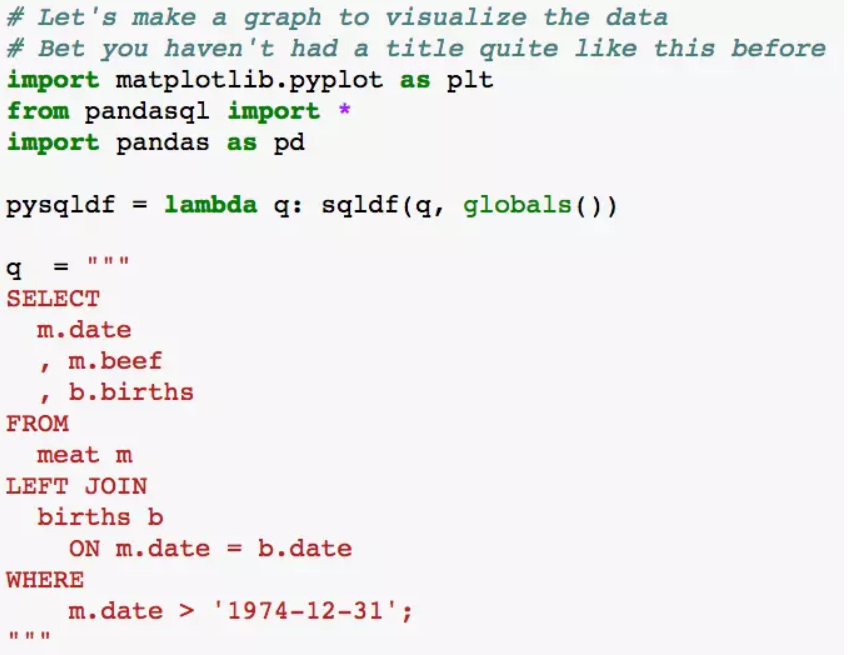


1. 数据清洗

由于第三方数据平台的不稳定性，数据存在一定的漏报、缺失、重复、空值等，在处理数据时，主要依赖的格式是DataFrame格式，所以使用pandas的自带的函数以及pandasql包进行数据清洗。

其中pandasql是一个可以在DataFrame中使用sqlite进行SQL增删改查操作的第三方包，本文中主要使用SQL进行数据清洗。

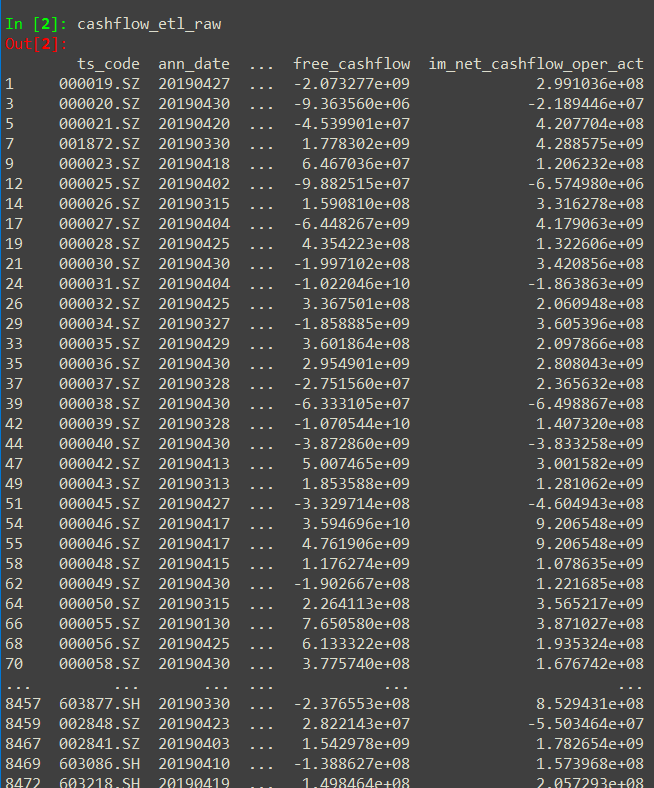
使用方法如下：





1. 数据预处理

部分原始数据如下：

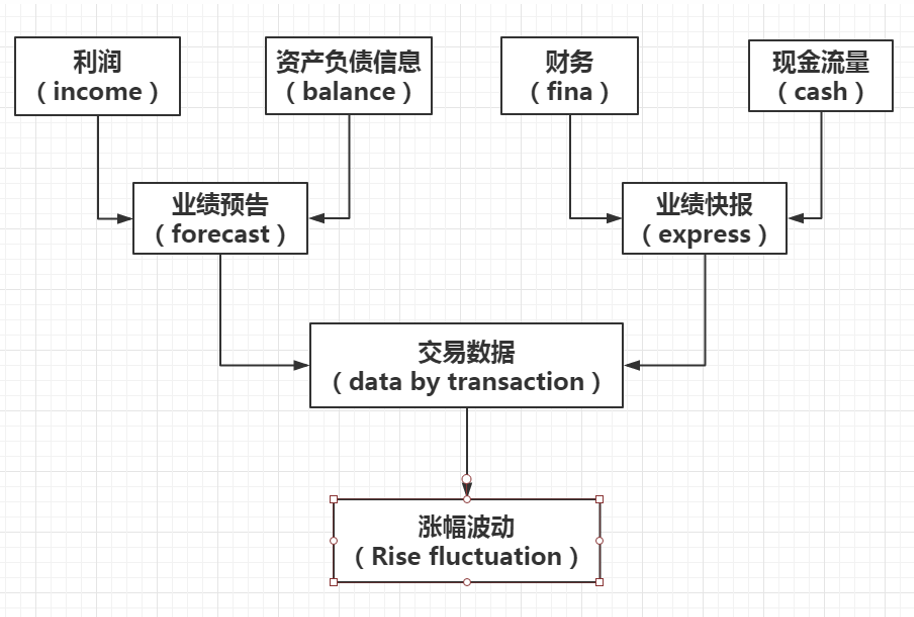


通过观察数据发现，数据特征离散程度较高，但是无法判断是否属于高斯分布，或者泊松分布，所以默认为是multi-Bernoulli分布。

由于贝叶斯网络的特性，需要计算不同情况下出现的概率，在此将数据特征从multi-Bernoulli分布转换成多项分布，所以运用多项式贝叶斯信念网络。在处理成多项式贝叶斯的时候，通过聚类方法以及RFM分组划定类别，将形成的类别看作多项式，进行贝叶斯信念网络的构建。

4、数据建模

依据数据特征建立有向无环图（DAG）

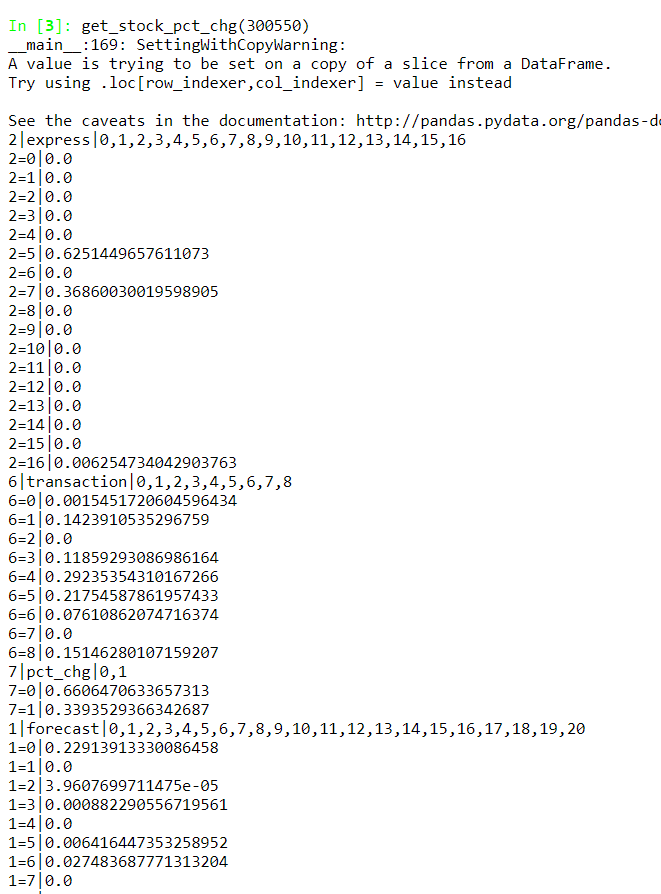


有向无环图建立完成后，通过第三方模块pybbn，完成贝叶斯信念网络的搭建

1. 预测结果

调用函数get\_stock\_pct\_chg(code) 例如下图

函数会计算并且返回各个类别的概率，这使得投资者们可以根据这一概率进行投资配置，以完善最佳的投资组合。



**四、详细设计**

1、（1）财务数据获取 范景泉

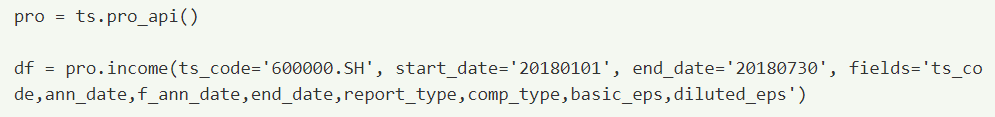
通过tushare数据包获取数据

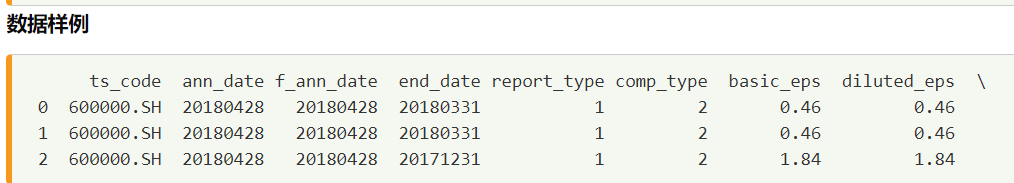
Tushare是一个免费、开源的python财经数据接口包。主要实现对股票等金融数据从数据采集、清洗加工 到 数据存储的过程，能够为金融分析人员提供快速、整洁、和多样的便于分析的数据，为他们在数据获取方面极大地减轻工作量，使他们更加专注于策略和模型的研究与实现上。考虑到Python pandas包在金融量化分析中体现出的优势，Tushare返回的绝大部分的数据格式都是pandas DataFrame类型，非常便于用pandas/NumPy/Matplotlib进行数据分析和可视化。

下载安装tushare:

安装Python安装pandaslxml也是必须的，正常情况下安装了Anaconda后无须单独安装，如果没有可执行：pip install lxml

方式1：pip install tushare，如果安装网络超时可尝试国内pip源，如pip install tushare -i https://pypi.tuna.tsinghua.edu.cn/simple 方式2：访问https://pypi.python.org/pypi/tushare/下载安装 ，执行 python setup.py install方式3：访问https://github.com/waditu/tushare,将项目下载或者clone到本地，进入到项目的目录下，执行： python setup.py install 使用实例：



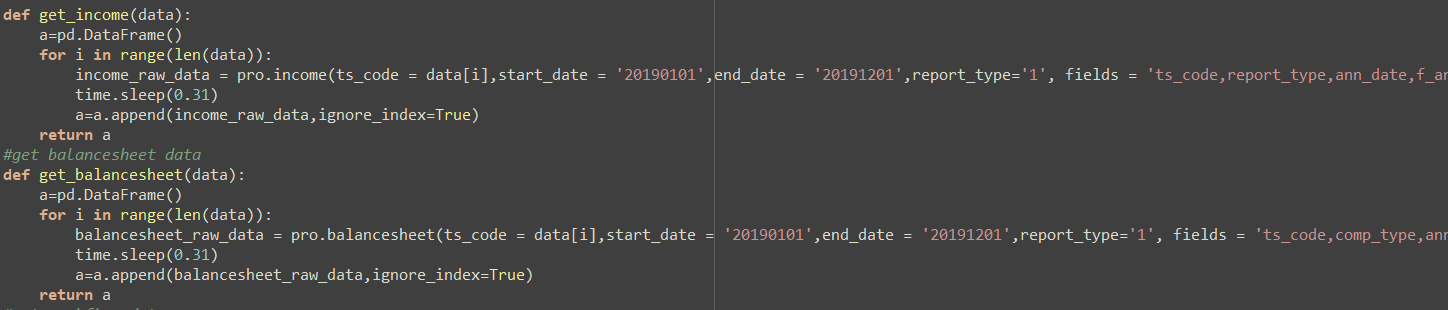


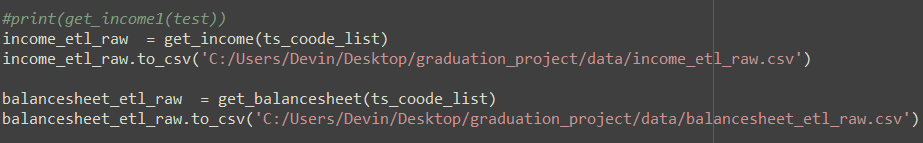
由于财务数据有很多种，在这里并不是获取的全部财务数据，比如在利润额相关信息表中，我们只获取了部分数据，



类似的还有很多，不过函数都已经封装好了，

循环调取函数，因为数据接口权限问题，0.31s获取一次

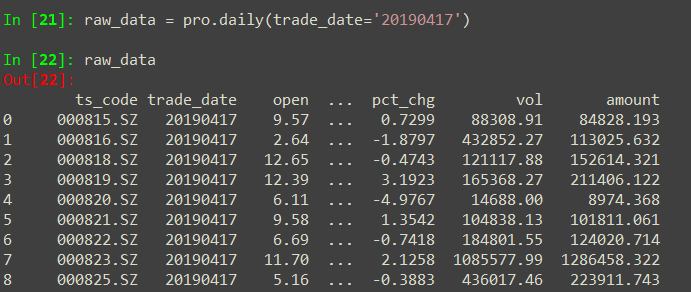




（2）历史交易数据获取 范景泉

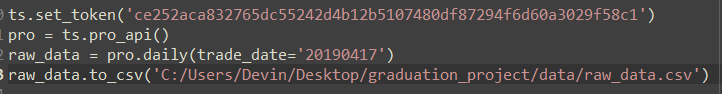
通过tushare包中内置的daily函数，可以获取某一天的历史交易数据

用法实例如下：



本次采用这天的日期作为原始数据，将其中的ts\_code保存为股票列表，当然，由于股票的停盘复盘问题，这里的股票列表并不是所有的上市公司股票列表，包含了一部分没有复牌的公司。

这是将实时数据的股票代码列表保存为股票列表：

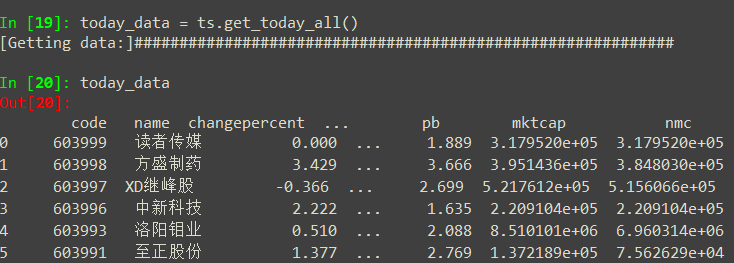


（3）实时数据获取 范景泉

通过内置的get\_today\_all() 函数获取当天的实时数据，需要注意的是，本文其他地方用的是tushare的PRO版本，但是在实时数据的获取用的是tushare的老版本，因为Pro版本没有实时数据的获取功能。

需要注意的是，但是在使用老版本时，tushare获取到的股票代码是不带交易所的，仅有六位数股票代码，二在Pro版本中，所有的ts\_code（股票代码）包含了那个交易所的信息，即.SH = 上海，.SZ = 深圳。

并且在算法中，每次构建bbn网络时，都会重新获取实时数据，因为数据被读到内存中，运行算法后会改变贝叶斯网络的节点值，为了使用原来的模型进行预测，所以必须在每次重新运行文件的时候重新读取。



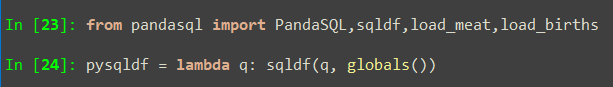
2、（1）数据清洗 张剑维

使用第三方包pandaSQL，通过其内置的sqlite 对DataFrame 进行数据清

Panda-Sql是一个Sql编辑器，Pl Sql和一个用于Oracle，Mysql，MsSql（sql server），PostgreSQL和SqLite数据库 的浏览器 。它是一个开源软件，轻量级（13MB），可用 法语和英文，使用简单直观。 它适用于 Window 和 Linux。

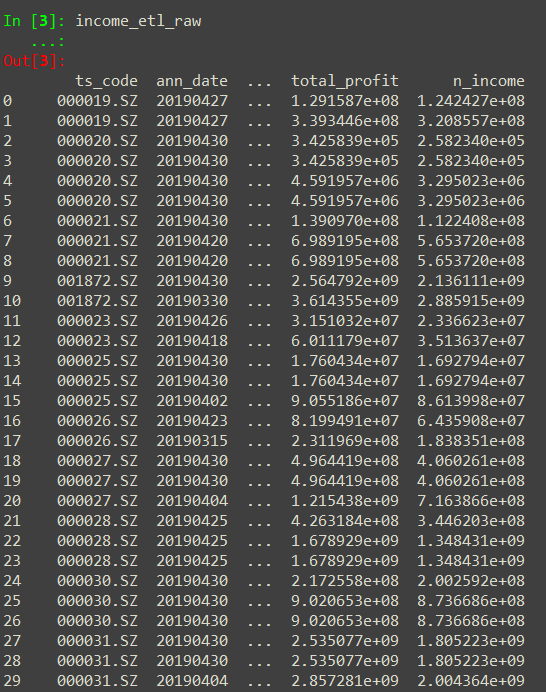
该程序的开发是为了便于数据库之间的数据操作和交换。它主要致力于数据库管理员，开发人员 ，数据采集员和支持技术人员。对于密集使用，仅保留了必要和常用功能，简单性和速度。

初始化函数如下：

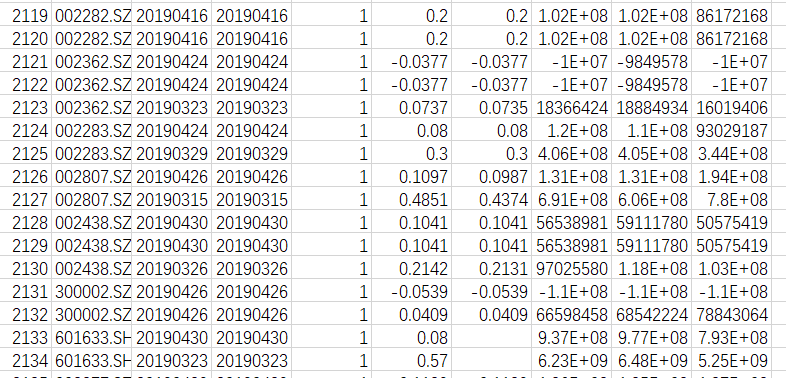


原始数据中存在着大量的噪声、包括不限于重复、数值缺失、空值等等。

原始数据：



为了能够更详细的观察数据，我把它存入到Excel中，用过Excel观察他的数据分布、数据的特点。



可以看到，同一只股票，存在两个不一样的日期，也就是说有两条记录。也有的是同一只股票，两个日期相同，但是数值不一样。还有就是两条记录重复了。以及有空值存在。

对上述的数据异常分布，我们使用SQL进行数据清洗。

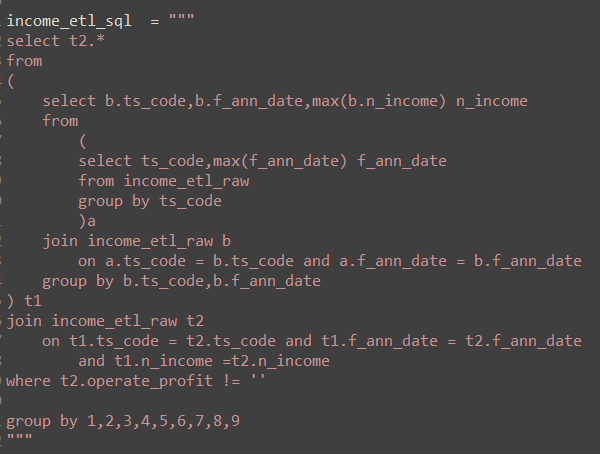
对与两个不一样的日期，取最近的日期。且以f\_ann\_date（实际公告日期）为准。

对于重复记录，我们删除了重复的记录。

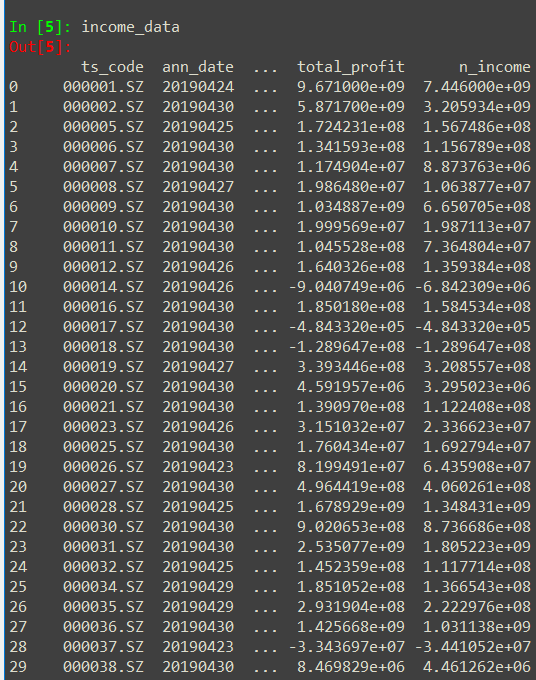
对于日期相同，但是数值不相同的数据。我们采用利润最大的那一条数据。

对于空值的情况，我们采用了零填充。

Sql ETL 如下：

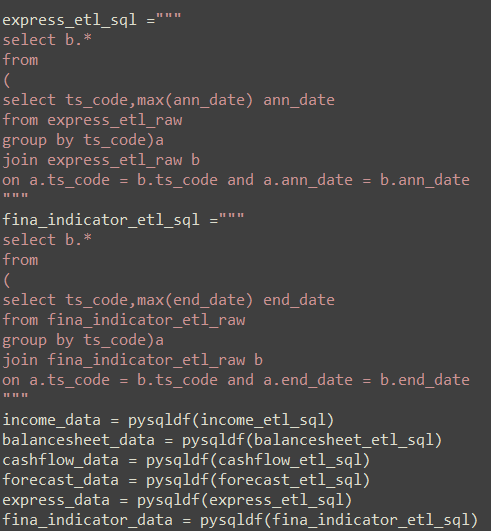


处理后的结果如下：



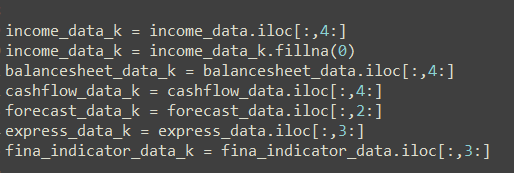
可以很明显的看到，重复值、错值、噪声都很好的剔除了。缺点是运行速度有点慢。

采用同样的方法对其他数据集进行处理。这里只列出了两个例子。

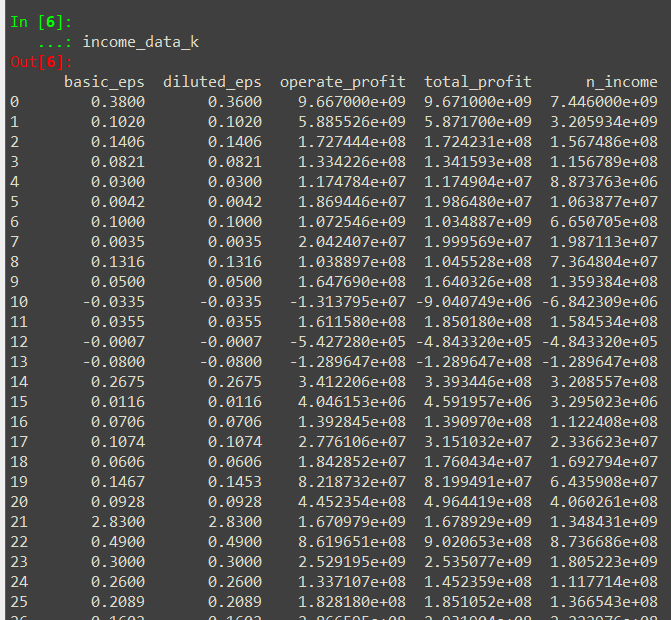


（2）指标构建 张剑维

这里的指标构建是指，将数据集切片。分割成便于进行聚类的数据结构，及全是数值型的数据结构，将日期、股票代码全部剔除，只保留部分特征信息。



指标构建后的数据例子：



3、（1）采用k-men和层次聚类方法对数据进行特征建模 范景泉 张剑维

所谓聚类，就是将相似的事物聚集在一 起，而将不相似的事物划分到不同的类别的过程，是数据分析之中十分重要的一种手段。比如古典生物学之中，人们通过物种的形貌特征将其分门别类，可以说就是 一种朴素的人工聚类。如此，我们就可以将世界上纷繁复杂的信息，简化为少数方便人们理解的类别，可以说是人类认知这个世界的最基本方式之一

K-means算法是聚类分析中使用最广泛的算法之一。它把n个对象根据他们的属性分为k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小

缺点：聚类中心的个数K 需要事先给定，但在实际中这个 K 值的选定是非常难以估计的，很多时候，事先并不知道给定的数据集应该分成多少个类别才最合适

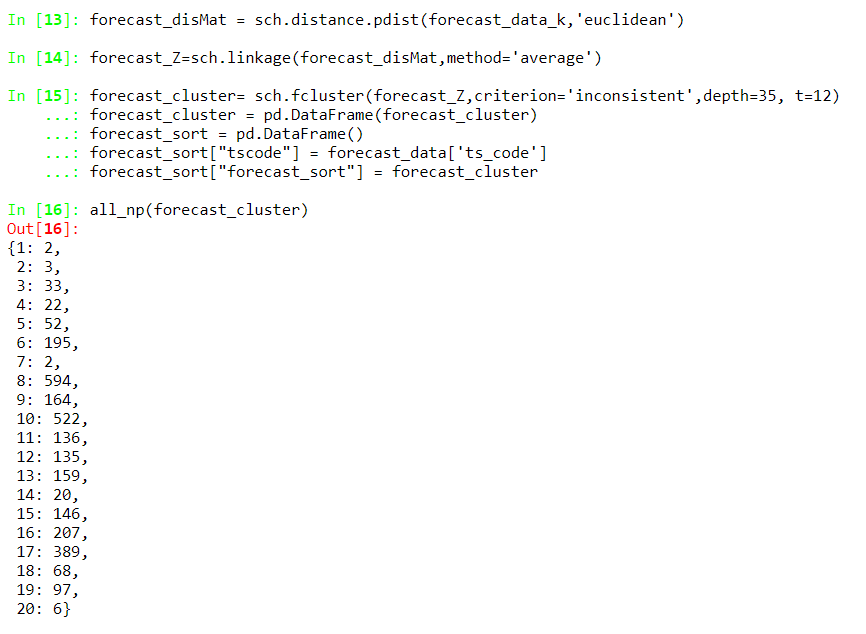
层次聚类，是一种很直观的算法。顾名思义就是要一层一层地进行聚类，可以从下而上地把小的cluster合并聚集，也可以从上而下地将大的cluster进行分割。似乎一般用得比较多的是从下而上地聚集。所谓从下而上地合并cluster，具体而言，就是每次找到距离最短的两个cluster，然后进行合并成一个大的cluster，直到全部合并为一个cluster。整个过程就是建立一个树结构。

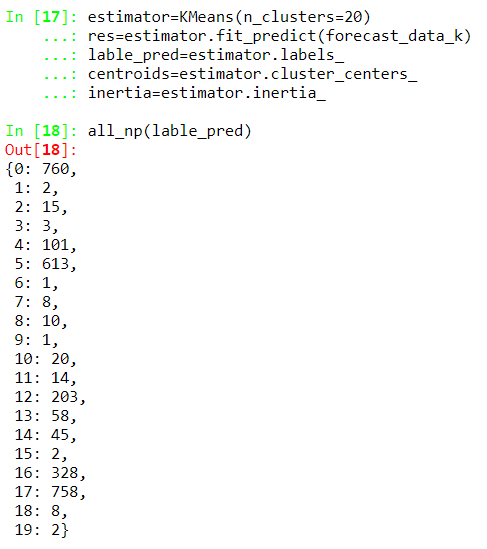
层 次聚类较大的优点，就是它一次性地得到了整个聚类的过程，只要得到了上面那样的聚类树，想要分多少个cluster都可以直接根据树结构来得到结果，改变 cluster数目不需要再次计算数据点的归属。层次聚类的缺点是计算量比较大，因为要每次都要计算多个cluster内所有数据点的两两距离。另外，由 于层次聚类使用的是贪心算法，得到的显然只是局域最优，不一定就是全局最优，这可以通过加入随机效应解决，这就是另外的问题了。

通过处理后的数据，经过指标构建之后，进行无监督的数据分类。

下图是k-men和层次聚类的效果图

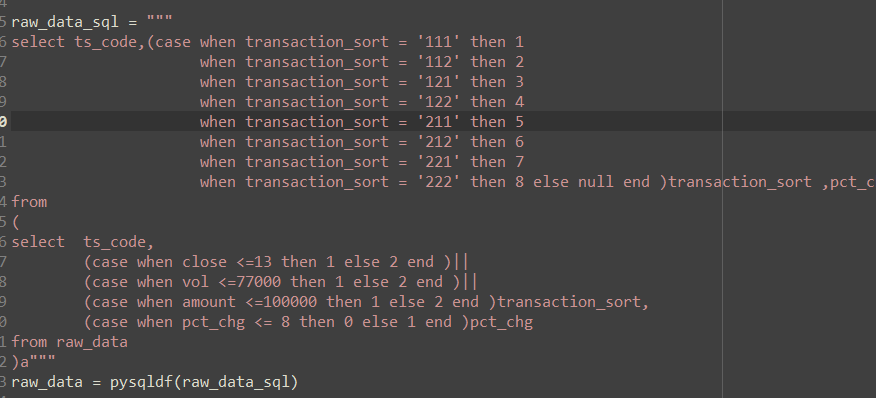
比较k-mens聚类以及 层次聚类之后，层次聚类的效果更好。





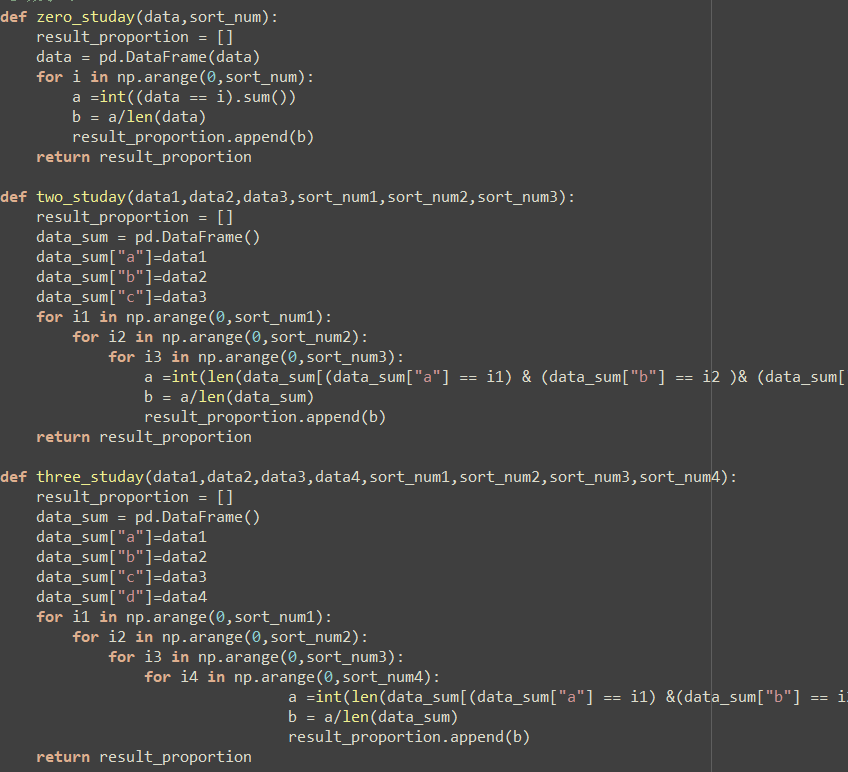
（2）采用RFM模型变体，通过专家法进行交易数据建模 范景泉 张剑维

通过阅读文献，咨询等方式，确定划分界限。通过SQL的etl过程区分类别，这里为了程序运行速度，牺牲了分组的可能性。



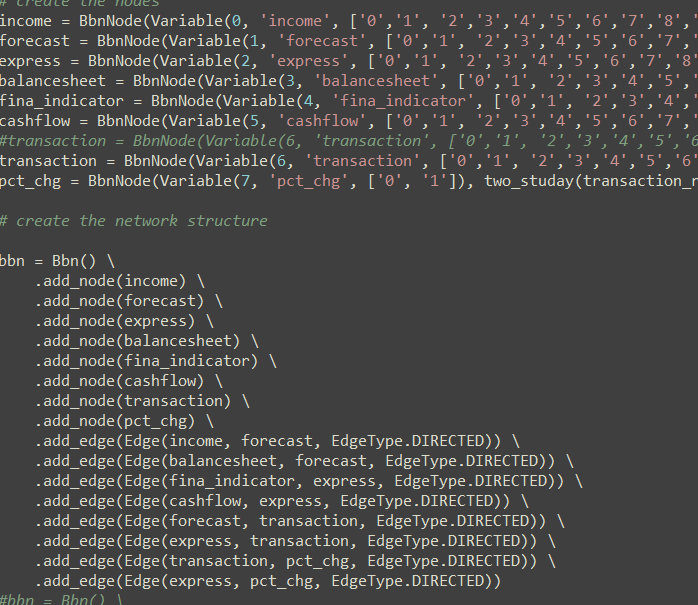
4、（1）设计贝叶斯有向无环图 范景泉 张剑维

（2）建立贝叶斯网络参数学习模型 范景泉



（3）建立贝叶斯网络模型 范景泉

建立点模型，构建贝叶斯网络。



5、（1）获取实时数据

输入并清洗数据



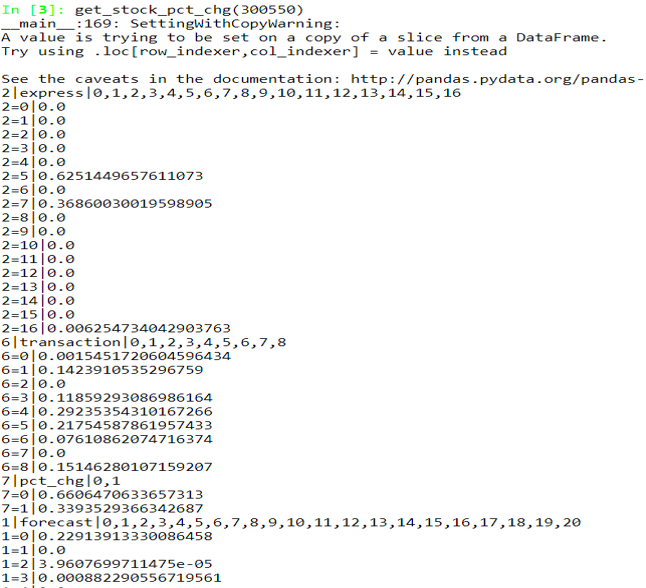
（2）输入实时数据进行预测

定义函数，插入数据，改变贝叶斯网络的概率，更新贝叶斯网络，计算概率，对数据进行预测



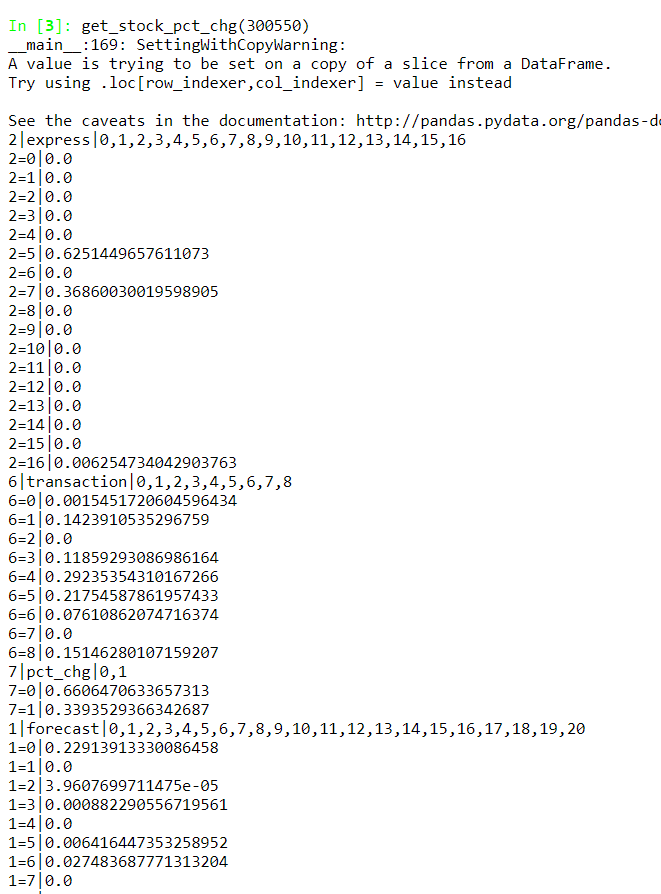
**五、软件测试**

运行结果如下，计算可得各条件发生的概率。



通过在线平台以及本机进行调试

输入结果函数，获取当前股票的涨幅概率



**六、软件完成后的总结汇报**

数据清洗过程中还是需要更快速的过程

功能基本可以实现，数据源有问题，原始数据缺失严重。

指标构建不精确。

聚类类别调参不细致，可以用手肘法进行调试

贝叶斯网络训练时间较长，pc端计算能力不足，只能把他拆分

参数学习中，尽量调用机器学习包

**七、参考文献**

数据清洗过程中还是需要更快速的过程

References

Blog.51cto.com. (2019). *基于离散型贝叶斯网络的模型平均的预测-风之祭-51CTO博客*. [online] Available at: https://blog.51cto.com/822863/163930 [Accessed 18 April 2019].

Blog.csdn.net. (2019). *机器学习笔记六：K-Means聚类，层次聚类，谱聚类 - 谢小小XH - CSDN博客*. [online] Available at: https://blog.csdn.net/xierhacker/article/details/61632529 [Accessed 28 April 2019].

Bunnak, P., Thammaboosadee, S. and Kiattisin, S. (2015). Applying Data Mining Techniques and Extended RFM Model in Customer Loyalty Measurement. *Journal of Advances in Information Technology*, pp.238-242.

Chen, D., Sain, S. and Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), pp.197-208.

Cnblogs.com. (2019). *伯努利分布、二项分布、Beta分布、多项分布和Dirichlet分布与他们之间的关系，以及在LDA中的应用 - stardsd - 博客园*. [online] Available at: https://www.cnblogs.com/sddai/p/9825000.html [Accessed 28 April 2019].

Honchar, A. (2019). *Financial forecasting with probabilistic programming and Pyro*. [online] Medium. Available at: https://medium.com/@alexrachnog/financial-forecasting-with-probabilistic-programming-and-pyro-db68ab1a1dba [Accessed 28 April 2019].

Python - 伯乐在线. (2019). *用 Python 进行贝叶斯模型建模（0） - Python - 伯乐在线*. [online] Available at: http://python.jobbole.com/85954/ [Accessed 8 April 2019].

waditu (2019). *Try to fetch data using my own token via http restful api followed the offical sample, but failed · Issue #799 · waditu/tushare*. [online] GitHub. Available at: https://github.com/waditu/tushare/issues/799 [Accessed 10 April 2019].

Xie, Y. and Song, T. (2018). Bearings-only multi-target tracking using an improved labeled multi-Bernoulli filter. *Signal Processing*, 151, pp.32-44.

yhat (2019). *yhat/pandasql*. [online] GitHub. Available at: https://github.com/yhat/pandasql/ [Accessed 5 April 2019].

张, 洋. (2019). *算法杂货铺——分类算法之贝叶斯网络(Bayesian networks) - T2噬菌体 - 博客园*. [online] Cnblogs.com. Available at: https://www.cnblogs.com/leoo2sk/archive/2010/09/18/bayes-network.html [Accessed 28 April 2019].

张, 洋. (2019). *算法杂货铺——分类算法之贝叶斯网络(Bayesian networks) - T2噬菌体 - 博客园*. [online] Cnblogs.com. Available at: https://www.cnblogs.com/leoo2sk/archive/2010/09/18/bayes-network.html [Accessed 11 April 2019].

机器之心. (2019). *概率编程：使用贝叶斯神经网络预测金融市场价格*. [online] Available at: https://www.jiqizhixin.com/articles/03134 [Accessed 1 May 2019].

杜, 忠. (2018). Electricity Customer Segmentation Study Based on Payment Behavior. *Smart Grid*, 08(03), pp.234-243.

高, 阳. (2019). [online] Cs.nju.edu.cn. Available at: https://cs.nju.edu.cn/gaoyang/ai2016/5.pdf [Accessed 08 April 2019].