

知乎

首发于  
川流不息

## 用 K-means 聚类做市场状态分析：大阳线之后更危险？



石川

量化交易 话题的优秀回答者

已关注

29 人赞同了该文章

### 1 无监督聚类

**无监督学习 (unsupervised learning)** 是机器学习中的三大类问题之一，另外两类分别为**有监督学习 (supervised learning)** 和**强化学习 (reinforcement learning)**。在无监督学习问题中，对于给定的观测数据无需（也没有）已知的响应 (response)，而是希望分析出观测数据本身的结构。在无监督学习中，**聚类 (clustering)** 和**降维 (dimension reduction)** 是主要的两大应用场景。

**无监督聚类的目的是将观测点按照它们的特征分成若干个子集——这些子集又称为簇 (cluster)——以使得每一簇内的观测点有相似的特性，而不同簇之间的观测点有不同的特性。**聚类分析的算法有很多；其中一个常见且有效的算法是 **K-means 聚类** (译作 **K-均值聚类**)，其中 K 代表簇的个数。

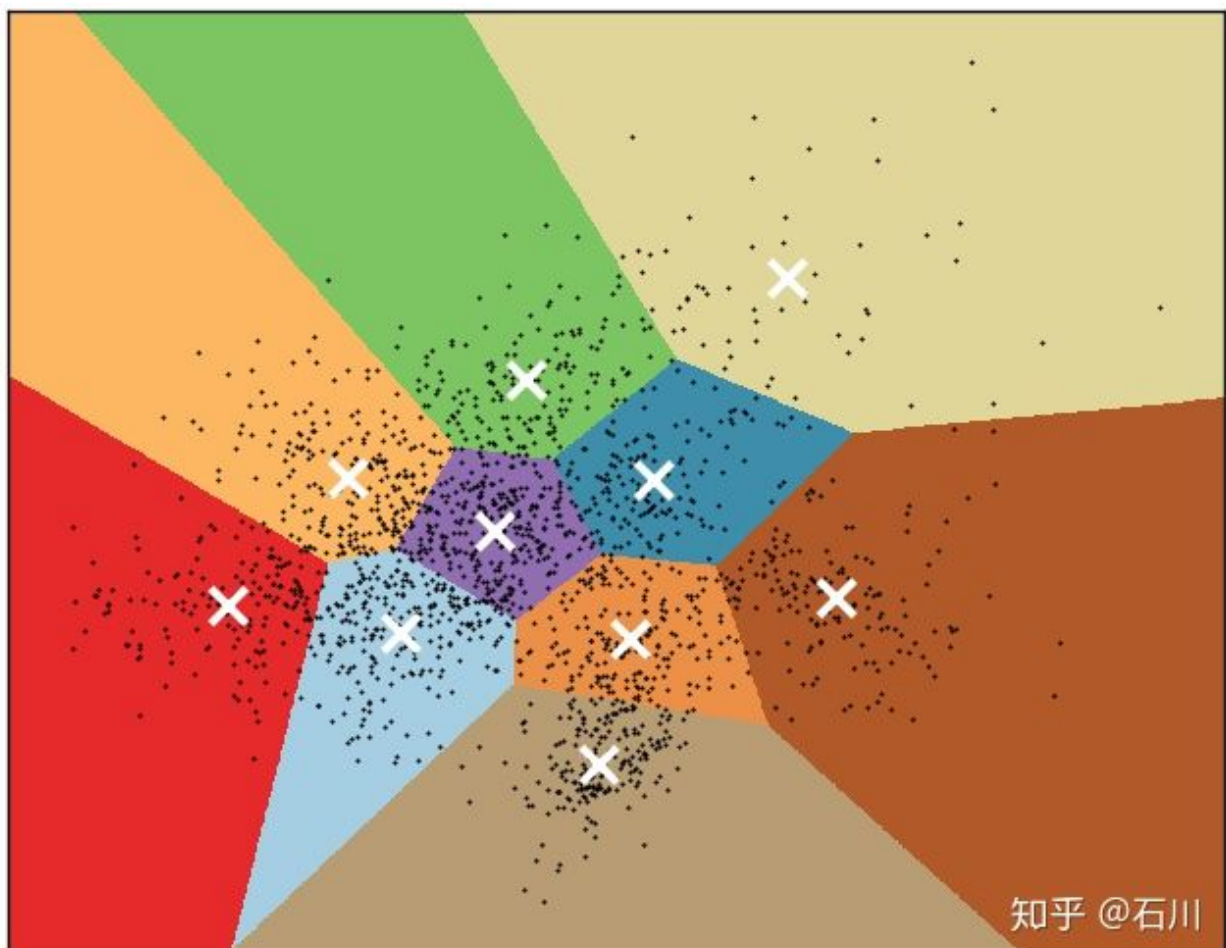
今天我们就来说说 K-means 聚类，它在量化投资领域有很多应用。为了说明这一点，本文除了介绍该算法外，还会以上证指数的价格数据为例说明如何利用该算法进行市场状态监测 (regime detection)。

## 2 K-means 聚类

**K-means 聚类是一种硬聚类 (hard clustering) 算法。**所谓硬聚类就是说每一个样本点都必须“非此即彼”的被分到某一个簇中。与硬聚类对应的是**软聚类 (soft clustering)**。针对每一个样本点，软聚类算法计算该点属于不同簇的概率，这是一种模糊 (fuzzy) 的概念，它不要求样本点和簇之间“非此即彼”的映射，而是允许样本点以不同的概率所属于不同的簇。

假设  $n$  维空间中共有  $N$  个观测数据。在数学上，硬聚类意味着  $K$  个簇将该  $n$  维度空间划分为  $K$  个互斥的区域，每个观测点属于且仅属于这  $K$  个簇中的某一个。令  $x_i$  代表簇  $k$ ， $k$  属于，不同的簇之间满足如下关系：

这两个式子说明硬聚类对空间的划分满足 **MECE 原则**，即 **Mutually Exclusive** (上面第一个式子)，**Collectively Exhaustive** (上面第二个式子)。下图是一个 K-means 聚类的示意图。图中不同的颜色代表着 10 个簇；每一个黑点代表一个观测点。每个簇内的白色叉子代表该簇的质心。这个图的意思是，如果我们有下图中的那些观测点，想采用 K-means 聚类将它们分为 10 个子集，那么就会得到如下的结果。



该簇质心的距离的平方和，因此簇内差异又称为簇内平方和 (within-cluster sum of squares)。由于质心代表着均值，这也是 K-means 聚类名字中 mean 一词的含义。

在数学上，该优化问题可以表示为：

其中， $\mu_k$  代表簇 k 的质心向量，它和观测点一样是 n 维向量。表达式  $d(x_i, \mu_k)$  代表簇 k 内的第 i 个点到质心的欧氏距离。

在欧几里得空间中，两个 n 维向量  $x$  和  $y$  的欧氏距离 (Euclidean distance) 定义如下：

对该优化问题求解，就可以得到最优的划分。**不幸的是，寻找该问题的全局解 (global optimum) 是 NP-hard (简单的理解就是复杂度太高，让计算机硬来也算不出来)**。所幸的是，可以使用启发式算法找到局部解 (local optimum)。该启发式算法分为两部，思路如下。

**第一步：随机的将每个观测点划分到一个簇 k；**

**第二步：重复本步骤中的过程，直到聚类结果收敛：**

**1. 根据当前的聚类结果，计算每个簇的质心**

**2. 根据最新的质心，计算每个观测点到这些质心的欧氏距离，将该点重新划分到距离它最近的质心所处的簇内。**

值得一提的是，局部解十分依赖于求解过程的初始值。且由于不知道全局解是什么，我们没法证明局部解就是最优的。为了尽可能降低这个问题的影响，可以多次使用该启发式算法找到不同的局部解，然后从它们中间找到最小的，作为最终的解。

在 python 的 sklearn 包里，有实现 K-means 算法的类 `sklearn.cluster.KMeans`。它的输入参数中，有一个 `n_init` (默认值为 10)，它就决定了求解局部解的次数。该算法会在求出的所有局部解中找到最优的，作为最终的解。

### 3 K-means 聚类算法的不足

在将 K-means 聚类应用于量化投资之前，有必要知道它的不足。具体来说，特别是针对金融数据，它有以下四点不足之处：

1. 金融数据信噪比太低，这意味着价格序列中有很多噪声。由于 K-means 是硬聚类，因此每个观测点都被迫分到一个簇中，因此噪声对聚类结果的影响不可忽视。
2. 金融数据中存在异常值 (比如黑天鹅事件造成的大跌，或者因为乌龙指造成的价格大幅震荡)。K-means 会把它们当作普通样本处理。因此这些异常值会对聚类结果产生影响。
3. K-means 对训练集的数据比较敏感。举例来说，如果将历史数据分为两份，分别进行聚类。

知乎

首发于  
川流不息

该算法对样本数据敏感。当样本点不足的时候，这个问题尤其严重。

4. K-means 对 K 的取值（即簇的个数）非常敏感。如果 K 的取值不当，便很难从聚类的结果中得到有益的推断。下一小结的例子就说明这一点。

## 4 K 的取值至关重要

聚类分析是为了挖掘观测数据自身的结构。如果我们在事前从业务的角度对数据的结构有一个认知、并以此来选取簇的个数，那么聚类分析的结果将会更有意义。反之，如果我们对待分析的数据一无所知，盲目的选择 K 的取值，那么得到的很可能是无意义的分析结果。

下面通过一个例子说明正确选取 K 值的重要性。

假设我们有 3 个二元正态分布，它们的均值向量、协方差矩阵分别如下所示：

$$\mu = \begin{pmatrix} 6 \\ 12 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 8 \\ 7 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2.5 & 0.0 \\ 0.0 & 5.0 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 2 \\ 6 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2.5 & 0.0 \\ 0.0 & 3.5 \end{pmatrix}$$

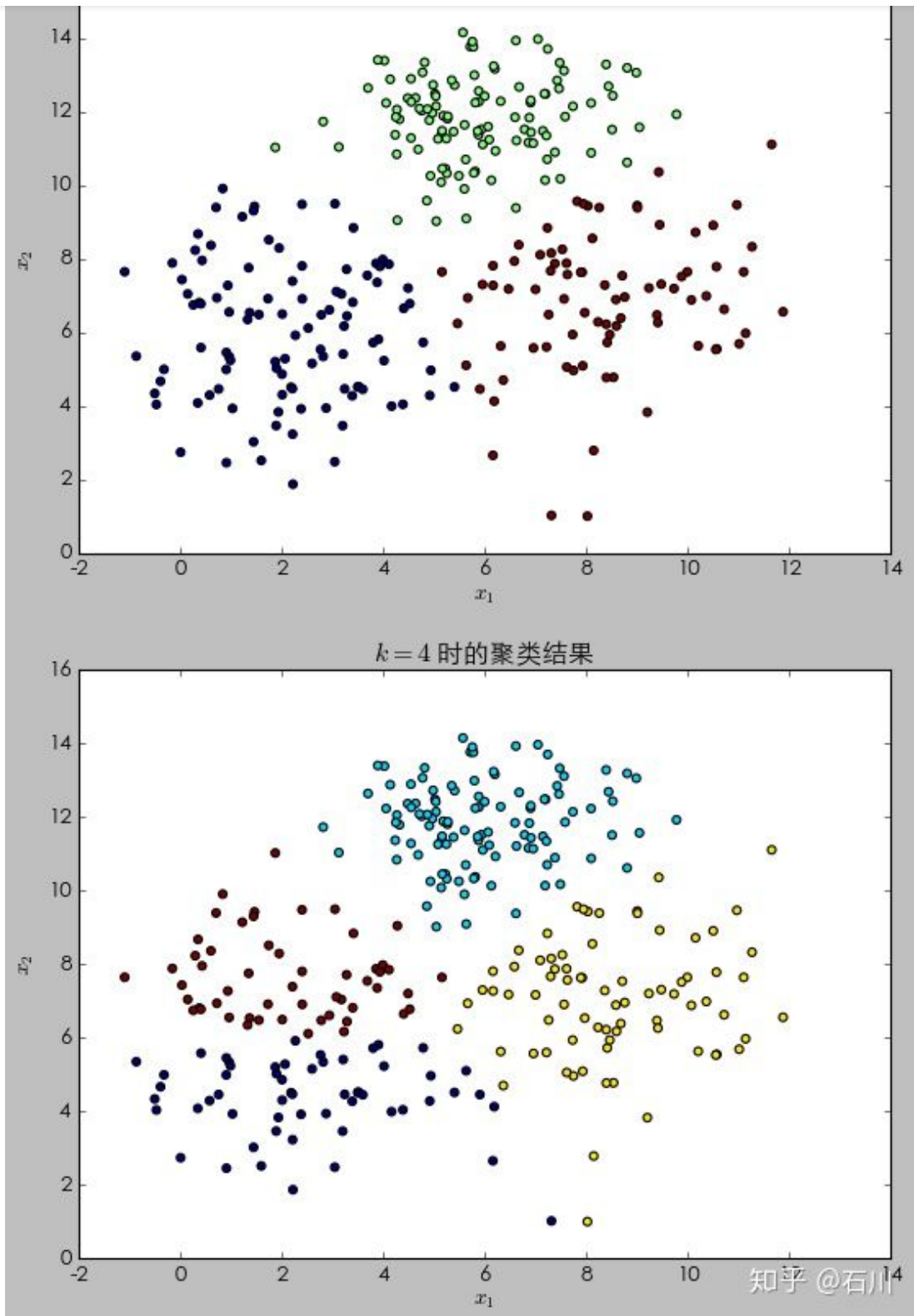
知乎 @石川

使用这 3 个二元正态分布在二维空间内各随机生成 100 个观测点（即一共有 300 个点），然后使用 K-means 聚类对他们进行划分。由于在这个例子中，我们知道这些点来自 3 个不同的二元正态分布，因此簇数 K 的正确取值应该为 3。为了比较，我们同时考虑 K = 4 的情况。下图展示了 K = 3 和 K = 4 时的 K-means 聚类结果。





知乎

首发于  
川流不息

当  $K = 3$  时，这 300 个观测点被分为了 3 簇。它们的质心基本位于  $(2, 6)$ 、 $(8, 7)$  以及  $(6, 12)$  这三个点附近——即这三个二元正态分布的均值点。由于  $K = 3$  和这些点的内在结构吻合（因为在这个例子中我们知道这些点是来自这 3 个不同的二元正态分布！），所以聚类挖掘出了有效的信息。

当  $K = 4$  时，这 300 个观测点被分为了 4 簇。比较两个聚类结果可知，来自于均值向量  $(2, 6)$  协方差矩阵  $(2.5, 0; 0, 3.5)$  这个二元正态分布的样本点被进一步细分为两个不同的簇（这是因

这个例子说明，当  $K$  的取值不当时，我们有可能从聚类的结果中得出错误的推断。因此，在使用 K-means 聚类之前，如能对待分析的数据有一定的了解，并能从业务的角度判断出合适的簇数  $K$ ，将大大提高聚类分析结果的可靠性。

## 5 用 K-means 聚类进行市场状态监测

本节使用一个简单的例子将 K-means 聚类应用于量化投资领域。我们使用上证指数日线的开盘、最高、最低、收盘价（即 OHLC 数据）来描述市场所处的（未知）状态，通过聚类将不同的交易日划分到不同的市场状态中，并在聚类的结果上进行进一步的推断。交易日的时间跨度为过去 5 年（本文写作于 2017 年 7 月，因此实证结果仅到写作之前）。

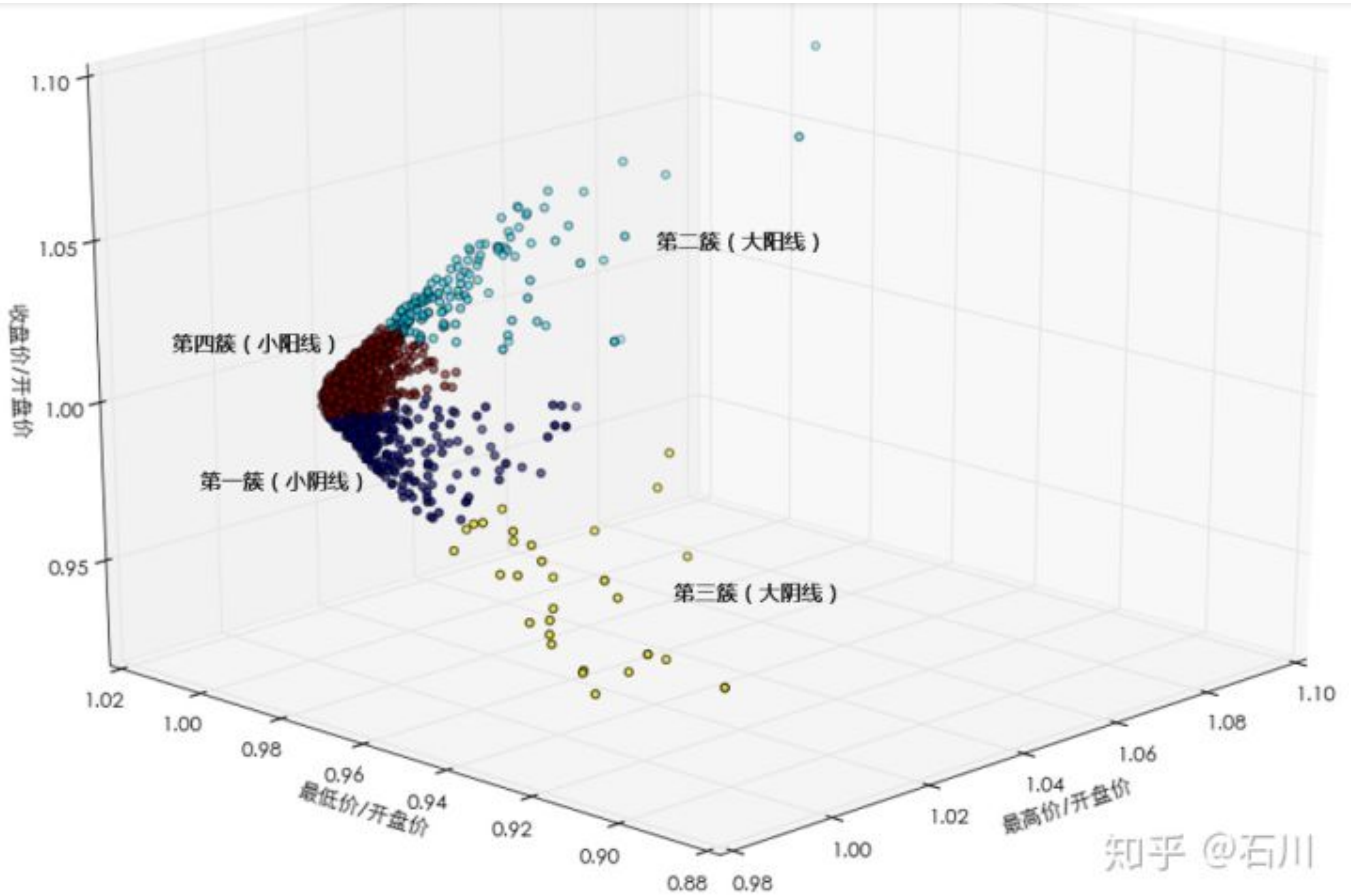
在这样的设定下，每一个交易日的 OHLC 数据就是一个观测点。**为了不同的交易日的价格数据有可比性，有必要进行标准化处理。**为此，使用每日的开盘价对其他三个价格进行标准化，得到  $H/O$ ， $L/O$ ， $C/O$ ，即最高价和开盘价之比、最低价和开盘价之比、以及收盘价和开盘价之比。标准化后，每一个观测点实际上是一个三维向量。

接下来就是确定簇数  $K$  的取值。在本例中，每一簇便代表了市场的一种状态。从这个角度出发，我们假设  $K$  的取值为 4 —— 即市场存在 4 种状态。

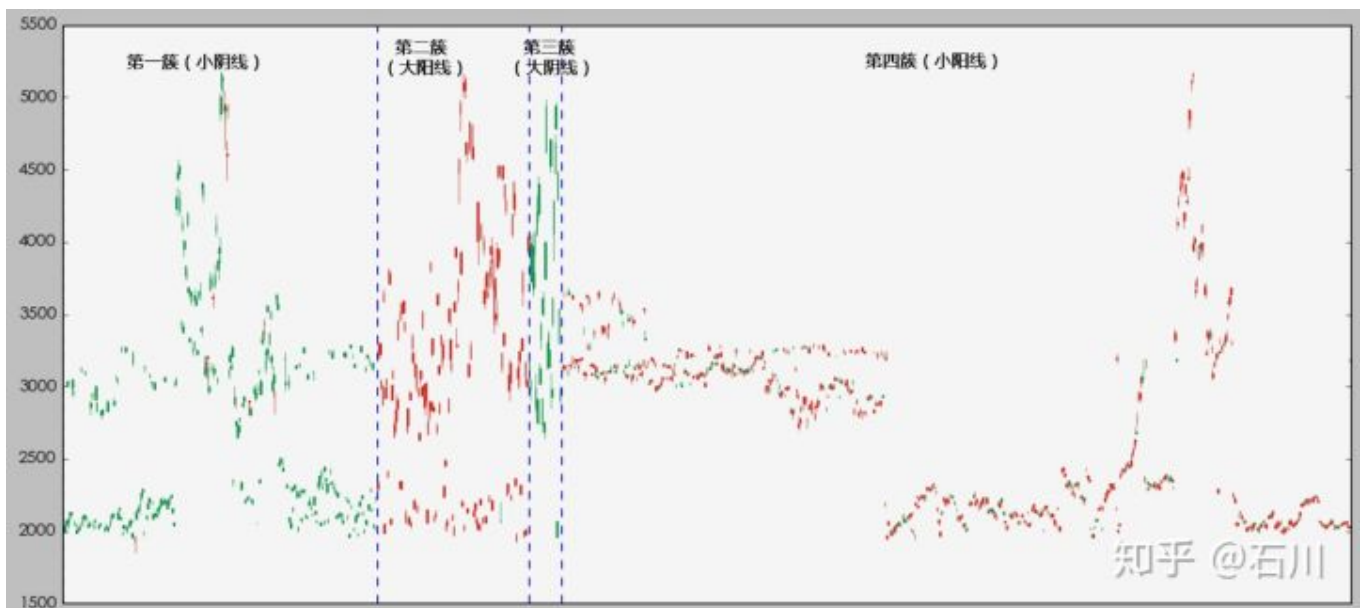
这里取 4 并没有什么特别的含义，作为读者的你也尽可以发挥想象来解读这个取值。从聚类的结果来看，由于我们用的是标准化后的 OHLC 数据，这 4 类市场状态对应的基本上是大阳线、大阴线、小阳线和小阴线。

由于观测点都是三维的，因此可以方便的在三维空间画出聚类的结果。以不同颜色表示不同的簇，这 4 簇的聚类结果如下图所示。大部分观测点都围绕在  $(1.0, 1.0, 1.0)$  附近，它们构成了两簇 —— 小阳线和小阴线；少量的观测点在远离  $(1.0, 1.0, 1.0)$  的位置，构成另外两簇 —— 大阳线和大阴线。





如果我们按照簇把每个交易日的 K 线画出来，则可以更清晰的看出簇与簇之间交易日 K 的差异（下图）。



从这个图中可以看出：

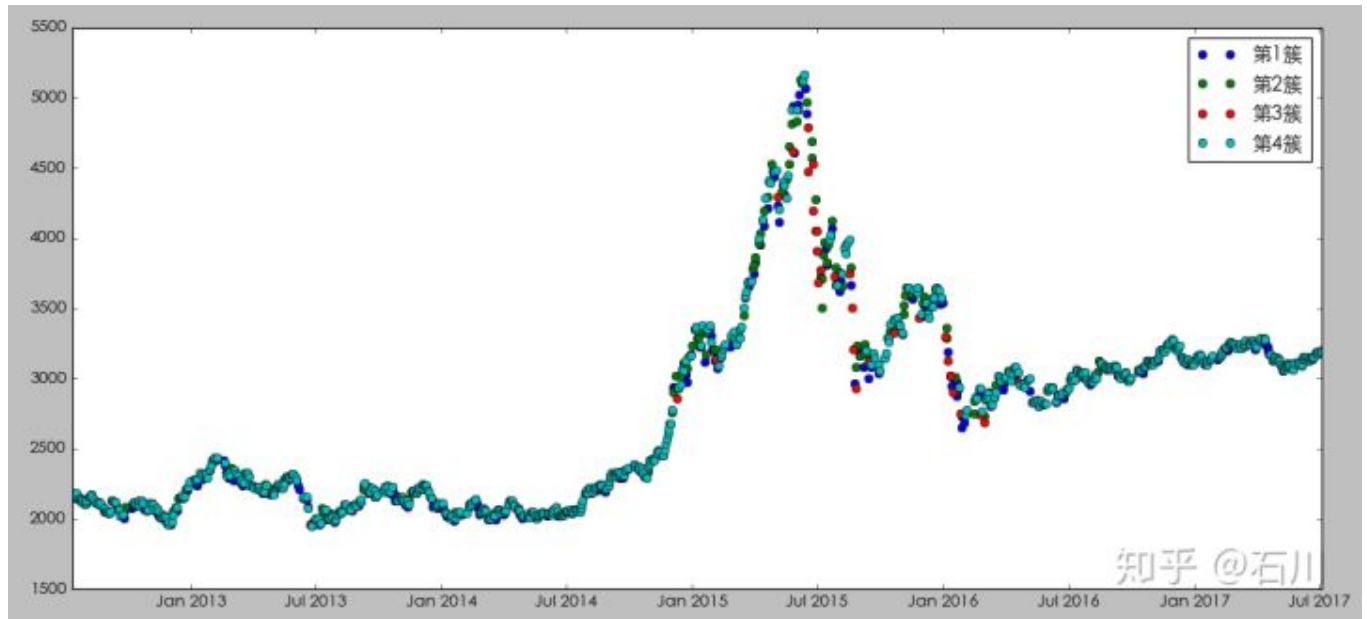
- 第一簇中的 K 线大部分都是**短的绿色线**，说明这一簇中以**小阴线**为主；
- 第二簇中的 K 线大部分都是**长的红色线**，说明这一簇中以**大阳线**为主；
- 第三簇中的 K 线大部分都是**长的绿色线**，说明这一簇中以**大阴线**为主；

知乎

首发于  
川流不息

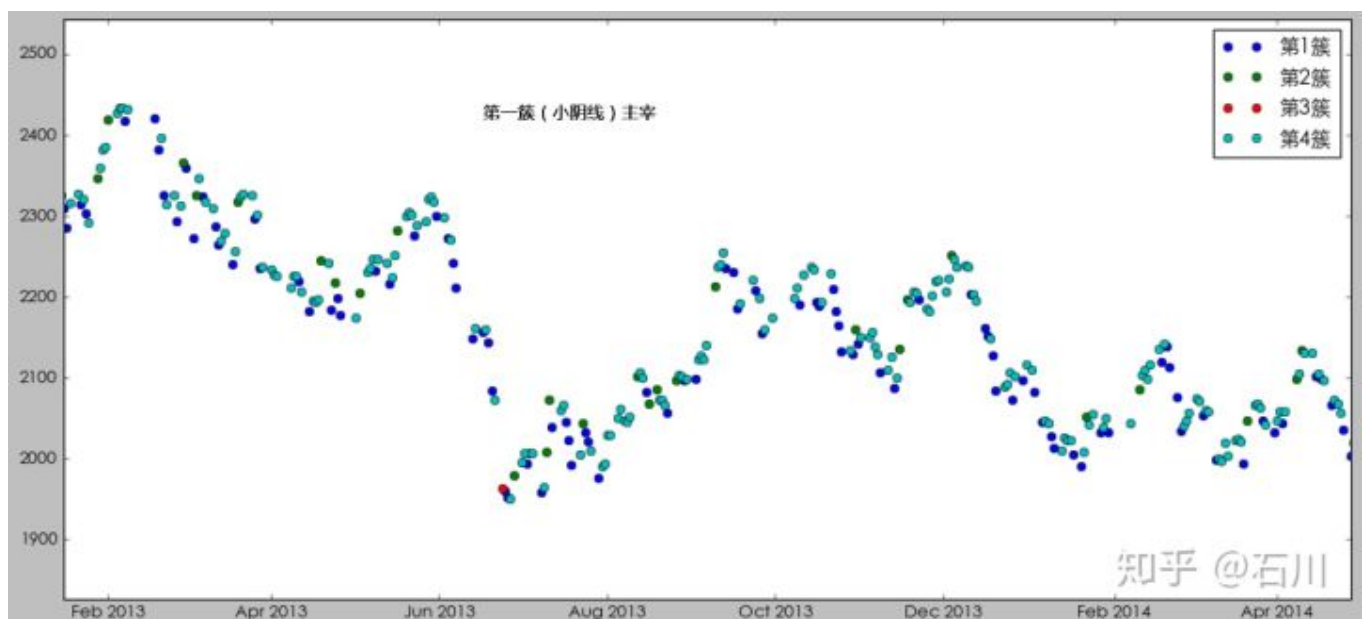
不过这个结果也清晰的说明，我们的样本是严重的不均衡的，第四簇小阳线内的观测点远超其他三簇。样本严重不均衡对所有的机器学习算法都是一个挑战。我们会在下面再谈到这个问题。

如果按照时间顺序把每个交易日的市场状态画出来，则得到下图。



我们分几个不同的时期来仔细看看。

在 2014 年底牛市启动之前，市场的状态受第一簇（小阴线）主宰，表现出来一个慢慢阴跌的态势。

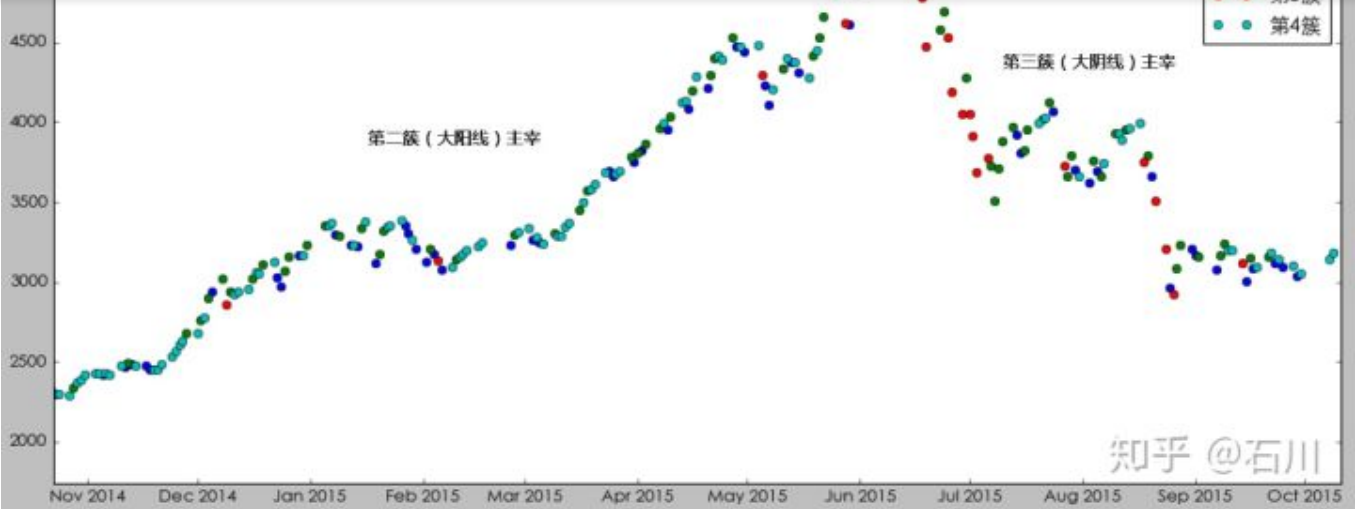


在 2014 年底到 2015 年底这个牛熊周期中，在牛市中市场状态由第二簇（大阳线）主宰，而在熊市中市场状态由第三簇（大阴线）主宰。

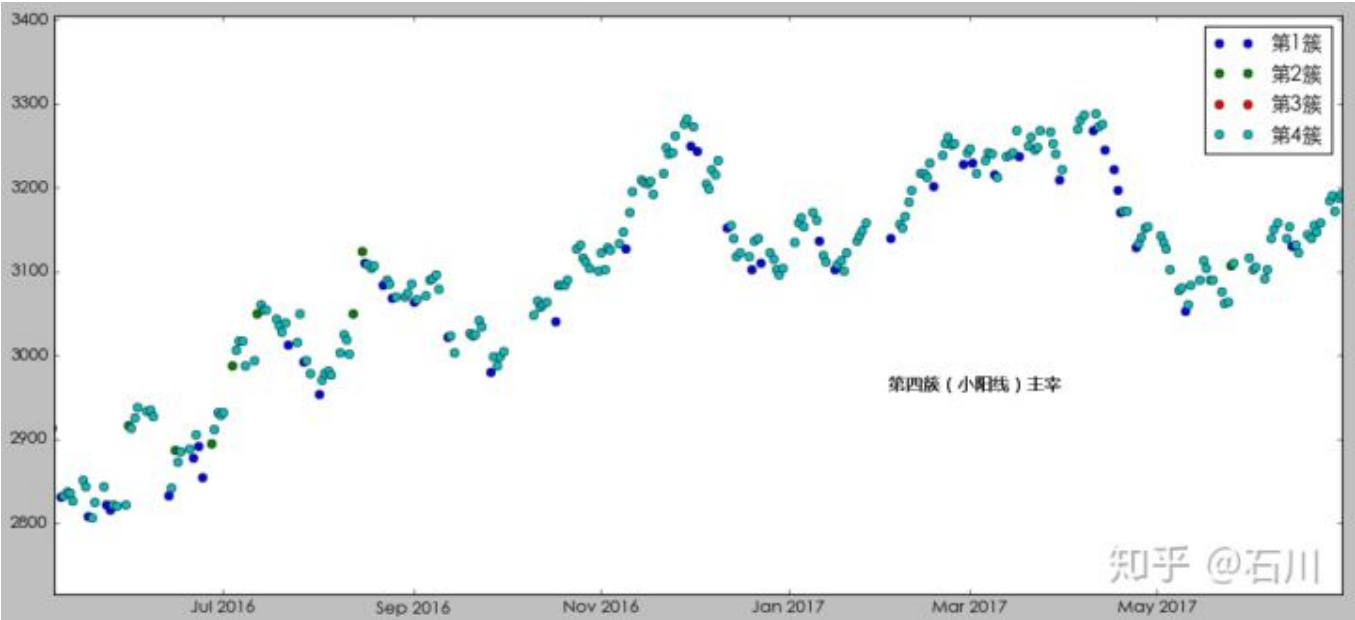


知乎

首发于  
川流不息



最后，从 2016 年二季度开始，市场状态由第四簇（小阳线）主宰，呈现出慢牛的走势。



在我们有了每个交易日的状态之后，便可以进行一系列的数据分析，得到进一步的推论。**这其中最有效的应该是求出市场状态的转移矩阵，它告诉我们在当前的状态下，下一个交易日市场将处于状态的条件概率。**这对策略择时和风控会有很大帮助。

基于上面的聚类结果，得到市场状态的转移矩阵如下。其中第 行第 列的数值表示在今天的市场状态为 的条件下，明天市场状态为 的条件概率。对于每一个，明天最有可能的状态 被用红色粗体表示出来。这个结果说明，除了大阴线外，在其他三种状态下，下一个交易日最有可能出现的都是小阳线，这和前面提到的样本严重不均衡密切相关。

知乎

首发于  
川流不息

第一簇 (小阴线)	0.196610	0.142373	0.027119	<b>0.633898</b>
第二簇 (大阳线)	0.195804	0.202797	0.041958	<b>0.559441</b>
第三簇 (大阴线)	0.333333	<b>0.366667</b>	0.233333	0.066667
第四簇 (小阳线)	0.269283	0.082544	0.012179	<b>0.635995</b>

知乎 @石川

本文的标题提出一个问题“大阳线之后更危险？”。这个问题可以通过这个状态转移矩阵回答。如果今天是大阳线，则下一个交易日是大阴线的**条件概率**为 4.2%（第二行、第三列的数值）。让我们再来看看大阴线出现的**非条件概率**。在回测的 1207 个交易日中，有 30 个交易日属于第三簇，因此大阴线的非条件概率仅为 2.5%，小于前面这个 4.2% 的条件概率。基于这个结果，我们得出“大阳线之后更危险”的推论。**这个结论事实上是符合人的认知的。这是因为无论大涨还是大跌，都意味着波动率的上升；而波动率的上升意味着风险的加大；风险加大意味着大跌的可能性增大。**

假如上述聚类分析的结果是有效的，那么使用这个转移矩阵可以回答很多类似的问题、得到很多有益的推论。

## 6 结语

**样本不足和样本不均衡是金融数据的两大特色。**这些对于 K-means 聚类算法在量化投资中的应用提出了严峻的挑战。对于待分析的数据，“如何有效的选取特征？”、“适合的簇数 K 是多少？”，这些都属于算法本身之外的问题，但它们又对算法的分析结果至关重要。比如在上面的例子中，使用 OHLC 数据描述市场状态是否恰当？K = 4 是否有足够的依据？要回答这些问题，自然需要更多的研究。

任何机器学习算法都仅仅是工具。在金融领域，核心的问题不是工具的使用，而是从对市场的理解。唯有理解了市场，才能选择正确的工具。掌握一门算法并不需要很长的时间；但要想深刻理解市场则需要时间的积淀。

(全文完)

**免责声明：**文章内容不可视为投资意见。市场有风险，入市需谨慎。




编辑于 2019-07-03

聚类   机器学习   量化交易

▲ 赞同 29   ▼   ● 添加评论   ➤ 分享   ★ 收藏   ...

文章被以下专栏收录



川流不息

北京量信投资管理有限公司是一家在中国基金业协会备案登记的专业私募基金管理人...

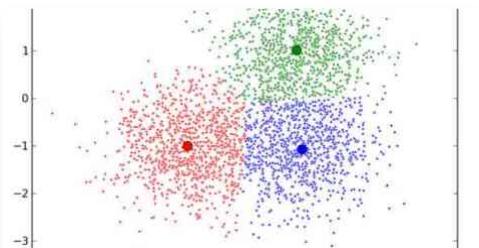
关注专栏

推荐阅读



吴恩达机器学习第八周一聚类降维Kmeans算法

Lyon   发表于DeepA...



K-means算法的改进：K-means++

weapo...   发表于高阶Pyt...



【点宽量化神经网络及其】

DigQuant点

还没有评论

因作者设置，评论已关闭