高维数据的可视化和快速聚类算法

杨 莉

（西南科技大学理学院 绵阳 621002）

摘 要 本文通过介绍一种用于高维数据的可视化方法，引入了可用于快速聚类的一种距离算法，该方法不仅具有鲁棒性而且有较低的计算复杂性 $O(n^1)$，最后我们将该方法用于金融数据立方体的聚类算法，主要用于挖掘庄家行为模式并作为是否存在操纵行为的依据。

关键词 数据可视化，聚类算法，数据挖掘

Visualization of High-dimension Data and the Study of Quick-Clustering Algorithms

YANG Li

(School of Science'SWUST, Mianyang 621002)

Abstract This paper gives an approach to a distance algorithms of which can be used into quick-clustering through a visual method of high-dimension data. The new method has robustness and lower computering complexity. At last, it is used into clustering algorithms of financial data cube, which is used to mine banker deed mode and whether is pursuant from manipulating deed or not.

Keywords Visualization of data, Clustering algorithms, Data mining

1 引言

数据挖掘[1]是一种新兴的面向决策支持的数据处理手段，金融数据挖掘是其中内容最丰富的研究方向之一，相关的文献[2,3]这里不一一例举。数据挖掘关注规则的发现和算法的简洁快速，面对海量、高维的金融数据，很多传统的数据挖掘和聚类方法在可视化、快速适时、鲁棒性等方面都存在一些问题，迫切需要建立更好的可视化算法和聚类算法。

金融证券数据的挖掘主要面临两个方面的困难，一是数据海量、维度高，如何有效地可视化显得很重要，比如中国沪深证券市场在全球看来仅仅是一个微型的市场，但每天需要处理的数据也非常可观，这还不包括各类文本信息；二是适时要求，这需要在算法的快速和鲁棒性上做文章，因此需要设计具有较低计算复杂性的挖掘算法，以适应在瞬息万变的市场中进行高速挖掘和快速决策。本文介绍了一种数据可视化方法—三角多项式图，提出了一种数据的鲁棒距离，并可在此基础上进行金融数据的快速聚类并发现庄家行为和股票操纵。

设 X_{kij} 表示数据立方体，其中 $k=1,2,\cdots,p$，p 表示该类品种成员数，比如研究国债、股票等，则 p 表示全部品种的个数，$t=1,2,\cdots,\tau$，τ 为一个确定的时间点，$i=1,2,\cdots,m$，m 表示单个品种的重要时间计量个数，比如某只股票的收盘价、开盘价、最高价、最低价等等。$j=1,2,\cdots,n$，n 表示一个单元内的序列值个数，比如每周的交易天数，每天的交易小时数等等。

注：当我们仅仅考虑单个品种时，$p=1$，数据可简化为 X_{kij}，在此基础上仅仅考虑某个计量值，可令 $m=1$，数据可记为 X_{ij}，这时如果再固定的考虑以天作为分析单元时，数据立方体就退化为一个数列 $\{X_t\}$。

一般地，我们记矩阵 $X_{kt}=(X_{kij})_{m\times n}$，通过对它右乘一个变量 η_n 得到的向量 $X_{kt}\eta_n$ 称为加权均线，比如取 $\eta_n=(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n})'$，则 $X_{kt}\eta_n$ 为 n 日均线。如果左乘一个向量 ξ'_m，相应的结果 $\xi'_m X_{kt}$ 称为中价。设

$$y_{kt}=\xi'_m X_{kt}\eta_n \tag{1}$$

从而得到 k 个序列 $\{y_{kt}\}$。

与传统的时间序列分析[4]所不同的是，对上述序列的数据挖掘需要同时自动快速地处理大量的序列集[5]，上面的金融序列 $\{y_{kt}\}$ 存在于海量的动态数据库内，我们需要挖掘存在于众多序列中对我们有用的信息，以便为决策和快速反映提供支持，这时，单个序列的运行规律的研究就不那么重要了，从而可以在方法上进行大量的简化，以适应算法上简洁快速的要求，满足金融投资的适时性要求。

2 数据的可视化和快速聚类算法

数据仓库和 OLAP 工具基于多维数据模型，该模型需要处理以数据立方体形式出现的数据。关于立方体数据的可视化本文建议采用三角多项式图[6]。取 $[-\pi,\pi]$ 上的正交函数系 $\{\sin t,\cos t,\sin 2t,\cos 2t,\cdots\}$，建立如下映射

$$V=(v_1,v_2,\cdots,v_m)\rightarrow f_v(t)=v_1\sin t+v_2\cos t+v_3\sin 2t+v_4\cos 2t+\cdots \quad -\pi\leqslant t\leqslant\pi \tag{2}$$

该映射具有如下优良性质：

(1) 保线性关系；

(2) 保欧氏距离；

(3) 为 R^m 到 L^m 上的一一对应的映射；

(4) 如果 V 的各个分量独立同方差 σ^2，则当 m 为偶数时，

$$var(f_v(t))=\frac{m\sigma^2}{2},\ -\pi\leqslant t\leqslant\pi \tag{3}$$

杨 莉 副教授，主要从事应用非线性分析、多维空间数据处理等。

当 m 为奇数时，

$$\frac{(m-1)\sigma^2}{2}\leqslant var(f_v(t))\leqslant\frac{m\sigma^2}{2},\ -\pi\leqslant t\leqslant\pi \tag{4}$$

(5)若 $V\sim N_p(\mu,\sigma^2 I_p)$，则以 $1-\alpha$ 的概率有

$$|f_v(t)-f_\mu(t)|^2\leqslant\frac{\sigma^2(m+1)}{2}\chi^2_m(\alpha),\ -\pi\leqslant t\leqslant\pi \tag{5}$$

其中 $\chi^2_m(\alpha)$ 为卡方分布的 α 分位点，上述性质的详细证明可参见文[6]。

此外，三角多项式图可以选择的投影方向为无穷多个，这也是它与众不同的地方。

设 $v,w\in R^m$，那么在 v,w 之间的欧氏距离为

$$d^2_{v,w}=(v-w)'(v-w) \tag{6}$$

由于 $f_v(t)$ 和 $f_w(t)$ 均为$[-\pi,\pi]$上的平方可积函数，它们之间的欧氏距离可以定义为

$$d^2_{f_vf_w}=\int_{-\pi}^{\pi}|f_v(t)-f_w(t)|^2\mathrm{d}t \tag{7}$$

文[6]已经证明了

$$d^2_{f_vf_w}=\int\pi\mathrm{d}^2_{vw} \tag{8}$$

这个严格的等式解决了这种图示法距离计算的复杂性问题，即 $d^2_{f_vf_w}$ 的复杂性至多为 $O(n^2)$。但我们知道用两条曲线间的面积作为它们之间的距离度量是非常恰当的，因此考虑下面的距离定义

$$\tilde{d}_{f_vf_w}=\int_{-\pi}^{\pi}|f_v(t)-f_w(t)|\mathrm{d}t \tag{9}$$

显然它不仅仅具有计算上的优势，而且比欧氏距离更具有鲁棒性，在维数比较多的情况下，不会夸大某局部范围的差异。本文采用它作为聚类的距离定义，可以达到快速聚类的目的。

实际计算中，我们可以采用两种近似算法，它们的算法复杂性均为 $O(n^1)$！

a. $\tilde{d}_{f_vf_w}\approx\sqrt{\pi}\sum_{i=1}^{m}|v_i-w_i|$

如果对精度没有要求可以用这个公式，目前数据挖掘中聚类算法广泛采用的距离如最小距离、最大距离、平均距离等(见文[1] pp. 237)和它类似。

b. $\tilde{d}_{f_vf_w}\approx\frac{2\pi}{s}\sum_{i=1}^{s}|\sum_{j=1}^{m}c_{ij}(v_j-w_j)|$，

其中 c_{ij} 均为常数，对每一个给定的 i

$$(c_{i1},c_{i2},\cdots,c_{im})=(\sin(-\pi+2\pi i/s),\cos(-\pi+2\pi i/s),\sin2(-\pi+2\pi i/s),\cos2(-\pi+2\pi i/s),\cdots) \tag{10}$$

这个公式可以对精度加以控制，由于 $\tilde{d}_{f_vf_w}$ 表示两条曲线 f_v 和 f_w 之间的面积，与 f_v 和 f_w 之间的欧式距离相比，不仅具有鲁棒性，且具有更明显的几何意义，采用常用的近似算法，先将区间$[-\pi,\pi]$等分成 s 个小区间，从而

$$\tilde{d}^*_{f_vf_w}=\frac{2\pi}{s}\sum_{i=1}^{s}\left|f_v(-\pi+\frac{2\pi}{s}i)-f_w(-\pi+\frac{2\pi}{s}i)\right|\xrightarrow[n\to\infty]{}\tilde{d}_{f_vf_w} \tag{11}$$

假定 v 和 w 是 $2p$ 维空间的点，则

$$f_v(-\pi+\frac{2\pi}{n}i)=(v_1,v_2,\cdots,v_{2p})(\sin(-\pi+\frac{2\pi}{n}i),\cos(-\pi+\frac{2\pi}{n}i),\cdots,\sin p(-\pi+\frac{2\pi}{n}i),\cos p(-\pi+\frac{2\pi}{n}i))^T\quad i=1,2,\cdots,n$$

$$f_w(-\pi+\frac{2\pi}{n}i)=(w_1,w_2,\cdots,w_{2p})(\sin(-\pi+\frac{2\pi}{n}i),\cos(-\pi\frac{2\pi}{n}i),\cdots,\sin p(-\pi+\frac{2\pi}{n}i),\cos p(-\pi+\frac{2\pi}{n}i))^T$$

$$i=1,2,\cdots,n$$

$$f_x(-\pi+\frac{2\pi}{n}i)-f_y(-\pi+\frac{2\pi}{n}i)=(x_1-y_1,x_2-y_2,\cdots,x_{2p}-y_{2p})\times(\sin(-\pi+\frac{2\pi}{n}i),\cos(-\pi+\frac{2\pi}{n}i),\cdots,\sin p(-\pi(\frac{2\pi}{n}i),\cos p(-\pi+\frac{2\pi}{n}i))^T\quad i=1,2,\cdots,n$$

由此可以得到

$$\tilde{d}_{f_vf_w}=\frac{2\pi}{s}\sum_{i=1}^{s}|\sum_{j=1}^{m}c_{ij}(v_j-w_j)| \tag{12}$$

通过选择不同的 s，可以适应对计算精度的不同要求。下面用这种距离算法进行金融证券数据的聚类分析。

3 应用算例

聚类是数据挖掘中知识发现的重要手段，可以从海量的数据集合中抽取隐含的、先前未知的、对决策有潜在价值的信息。当挖掘任务面临缺少领域知识或领域知识不完整的数据集时，采用聚类分析技术可以将无标识数据对象自动划分为不同的类，并且可以不受人的先验知识的约束和干扰，得到原本存在于数据集中的信息。我们可以采用可视化距离对证券交易数据作聚类，以发现可能隐含其中的操盘行为和潜在的模式变化。

对于证券数据聚类而言，我们并不关心绝对数值的聚类，而是变化规律的聚类，希望从中发现直观上不容易发现的信息，或者印证已有的猜测。为此我们需要先对数据 X_{tk} 进行标准化处理，以消除量纲的影响。我们采用如下的标准化数据

$$\tilde{X}^T_{tk}=\frac{X_{tk1}-X_{(t-1)k1}}{X_{(t-1)k1}},\frac{X_{tk2}-X_{(t-1)k2}}{X_{(t-1)k2}},\cdots,\frac{X_{tkm}-X_{(t-1)km}}{X_{(t-1)km}} \tag{13}$$

下面我们采用中国沪深 A 股 2001～2002 将近两年的日线数据进行分析，取 $s=10$，$m=5$，$data=(v_1,v_2,v_3,v_4,v_5)$=(close, open, high, low, vol) 对应证券日线的收盘价、开盘价、最高价、最低价、成交量，并假定数据已经标准化，则由(10)式可算得系数矩阵$(c_{ij})_{10\times5}$为

$$\begin{pmatrix}-0.98&-0.20&0.39&-0.92&0.83\\0.5&0.86&0.87&0.49&1.00\\0.30&-0.95&-0.58&0.81&0.80\\-0.91&0.41&-0.75&-0.67&0.30\\0.91&0.41&0.75&-0.67&-0.30\\-0.30&-0.95&0.58&0.81&-0.80\\-0.51&-0.86&-0.87&0.49&-1.00\\0.98&-0.20&-0.39&-0.92&-0.83\\-0.80&-0.60&0.96&-0.28&-0.35\end{pmatrix} \tag{14}$$

我们将相邻的两个数据先按式(13)标准化后再按公式(12)算出 $\tilde{d}^*_{f_vf_w}$，根据实际情况确定聚类阈值 u，当两个数据之间的距离 $\tilde{d}^*_{f_vf_w}<u$ 时，将它们聚为一类。具体算法如下：

```
d1:=(c-ref(c,1))/ref(c,1);
d2:=(open-ref(open,1))/ref(open,1);
d3:=(high-ref(high,1))/ref(high,1);
d4:=(low-ref(low,1))/ref(low,1);
d5:=(vol-ref(vol,1))/ref(vol,1);
a1:=abs(-0.98*(d1-ref(d1,1))-0.2*(d2-ref(d2,1))+0.39*(d3-ref(d3,1))-0.92*(d4-ref(d4,1))
```

(下转第 138 页)

种结合了图像帧全局信息和局部信息的局部高层语义特征提取方法。基于全局的方法能够从场景宏观角度分析图像帧与特定语义的关系，而基于局部的方法能够从物体微观角度分析图像帧与特定语义的关系。实验表明，此方法能够比单独基于全局或者局部的方法取得更好的结果。

参 考 文 献

1 Naphade M R, Smith J R. On the Detection of Semantic Concepts at TRECVID. In: Proceedings of the 12th annual ACM international conference on Multimedia, New York, 2004
2 Naphade M R, Smith J R, et al. IBM Research TRECVID-2004 Video Retrieval System. TRECVID, 2004
3 Naphade M R, Smith J R, et al. IBM Research TRECVID-2005 Video Retrieval System. TRECVID, 2005
4 Chang S F, Hsu W, et al. Columbia University TRECVID-2005 Video Search and High-level Feature Extraction. TRECVID, 2005
5 Hauptmann A, Baron R V, Chen M Y, et al. Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video. TRECVID, 2003
6 Chang C C, Lin C J. LIBSVM: a Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, 2002
7 Manjunath B S, Ma W Y. Texture Feature for Browsing and Retrieval of Image Data. IEEE Trans on Pattern Analysis and Machine Intelligence, 1996
8 Multimedia Description Schemes Group. Text of 15938-5 FCD Information Technology - Multimedia content description interface - Part 5 Multimedia description schemes. ISO/IEC JTC 1/SC29/WG11/N3966, 2001
9 Swets D L, Weng J. Using Discriminant Eigenfeatures for Image Retrieval. IEEE Trans on Pattern Analysis and Machine Intelligence, 1996

(上接第 133 页)

+0.83*(d4−ref(d4,1)));

a2:=abs(0.51*(d1−ref(d1,1))+0.86*(d2−ref(d2,1))+0.87*(d3−ref(d3,1))+0.49*(d4−ref(d4,1))+1*(d4−ref(d4,1)));

a3:=abs(0.3*(d1−ref(d1,1))−0.95*(d2−ref(d2,1))−0.58*(d3−ref(d3,1))+0.81*(d4−ref(d4,1))+0.8*(d4−ref(d4,1)));

a4:=abs(−0.91*(d1−ref(d1,1))+0.41*(d2−ref(d2,1))−0.75*(d3−ref(d3,1))−0.67*(d4−ref(d4,1))+0.3*(d4−ref(d4,1)));

a5:=abs(0.91*(d1−ref(d1,1))+0.41*(d2−ref(d2,1))+0.75*(d3−ref(d3,1))−0.67*(d4−ref(d4,1))−0.3*(d4−ref(d4,1)));

a6:=abs(−0.3*(d1−ref(d1,1))−0.95*(d2−ref(d2,1))+0.58*(d3−ref(d3,1))+0.81*(d4−ref(d4,1))−0.8*(d4−ref(d4,1)));

a7:=abs(−0.51*(d1−ref(d1,1))+0.86*(d2−ref(d2,1))−0.87*(d3−ref(d3,1))+0.49*(d4−ref(d4,1))−1*(d4−ref(d4,1)));

a8:=abs(0.98*(d1−ref(d1,1))−0.2*(d2−ref(d2,1))−0.39*(d3−ref(d3,1))−0.92*(d4−ref(d4,1))−0.83*(d4−ref(d4,1)));

a9:=abs(−0.8*(d1−ref(d1,1))−0.6*(d2−ref(d2,1))+0.96*(d3−ref(d3,1))−0.28*(d4−ref(d4,1))−0.35*(d4−ref(d4,1)));

a1+a2+a3+a4+a5+a6+a7+a8+a9<0.15;

经对中国沪深 1267 只 A 股进行分析，对每一只股票，我们将相邻的两个数据依据距离是否小于 0.03π 来聚类，存在不同的类显示了不同的投资环境或不同的操作行为。比如说，缓慢下降的股票数据相互之间很容易表现这种粘连(相互间距离很小)，这和上涨时完全不一样(见图 1)。

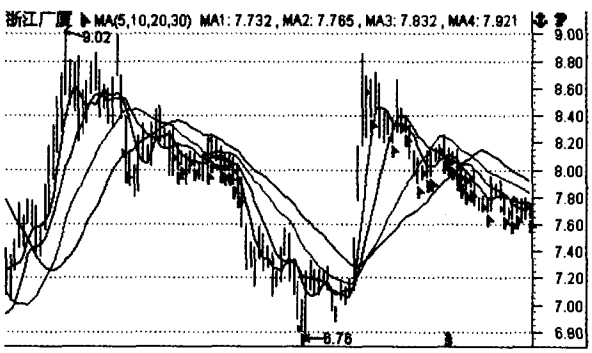

图 1 股票下跌初期往往表现出很好的聚类特征

这和很多控盘者采取这种方式出货有关，必须控制波动以凝聚人气。因此我们可以通过这一现象分析股票的操纵行为，我们判定连续 20 个交易日(一个月)可以聚入同一类的股票具有操纵行为，取 uπ/5 为距离阈值，当 u=0.2 时，在 2001 和 2002 年将近两年的时间里，具有这种操纵行为的股票有 289 种；当 u=0.15 时，我们称为中度操纵，挖掘出的股票 49 种；当 u=0.1 时，称为严重操纵行为，共有 11 只，它们是东风科技、啤酒花、新疆屯河、长春长铃、徐工科技、世纪中天、湘火炬 A、蜀都 A、合金投资、思达高科、太钢不锈。这其中很多是有名的庄股，如德隆系的新疆屯河、湘火炬 A、合金投资等。

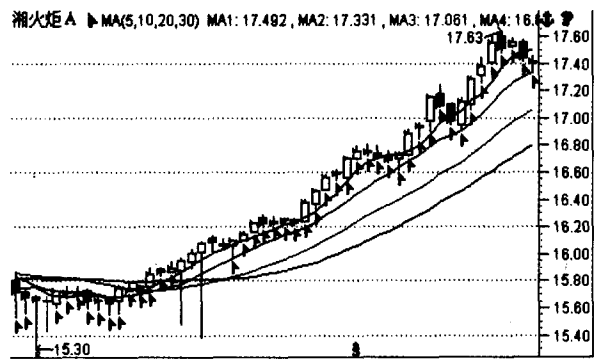

图 2 存在操纵行为的个股

结论 本文通过介绍一种数据可视化方法(三角多项式图)，在实现立方体数据的可视化的同时，引入了一种立方体数据的快速聚类算法，在此基础上进行了金融数据的快速聚类以发现庄家行为，尤其是股票操纵行为。结论表明，庄家行为会在其操盘上留下痕迹，我们可以通过聚类进行挖掘和分析，尤其具有严重操纵行为的股票，是可以被挖掘出来的。本文的结论对于实施有效的金融监管和加强市场执法有一定的参照意义。

参 考 文 献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc., 2001
2 John G H, Miller P. Building long/short portfolios using rule induction. In: Computational Intelligence for Financial Engineering, Piscataway NJ: IEEE Press, 1996
3 John G H. Stock selection using rule Induction. IEEE Expert, 1996. 52～58
4 Box G E P, Jenkins G M. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco, 1970
5 Liu L, Bhattacharyya S, Scove S L, Chen R, Lattyak W J. Data mining on time series: an illustration using fast-food restaurant franchise data. Computational Statistics & Data Analysis, 2001, 37: 455～476
6 Andrews D F. Plots of high-dimensional data. Biometrics, 1972, 28: 125～136