计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(8):1306~1311, 2005

主曲线异常检测及其在股票市场中的应用

齐红威[1,2] 张军平[1] 王 珏[1]
[1](中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100080)
[2](中国科学院计算技术研究所数字化技术研究室 北京 100080)
(hongwei.qi@mail.ia.ac.cn)

A Principal Curve-Based Outlier Detection Model and Its Application in Stock Market

Qi Hongwei[1,2], Zhang Junping[1], and Wang Jue[1]
[1](*Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)
[2](*Digital Technology Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)

Abstract To solve the outlier detection problems where outliers highly intermix with normal data, a general variance-based outlier detection model (VODM) is presented, in which the information of data is decomposed into normal and abnormal components according to their variances. With minimal loss of normal information in the model, outliers are viewed as the top k samples holding maximal abnormal information in a dataset. The VODM is a theoretical framework, and then, the principal curve is introduced as an algorithm of it. Experiments carried out on abnormal returns detection in stock market show that the VODM is feasible.

Key words outlier detection; principal curve; stock market

摘 要 复杂领域中,异常检测的困难是异常信息和正常信息高度混杂,针对此问题,提出了基于方差的异常检测模型(variance-based outlier detection model,VODM). 此模型把数据集的信息分解为正常信息和异常信息两部分,使得在正常信息损失最小的目标下,异常点集合就是前 k 个包含最多异常信息的样本. VODM 只是一种检测异常的理论框架,为此,采用主曲线作为其实现算法. 股票市场中异常收益检测的实验表明,VODM 及其算法是有效的.

关键词 异常检测;主曲线;股票市场
中图法分类号 TP18;TP391.41

1 引 言

异常检测(outlier detection)的目的是揭示数据集中隐藏的非平凡知识,即异常信息[1]. 相对于正常数据而言,异常数据可能包含更令人激动的信息,因此异常检测越来越受到研究者与应用者的关注.

然而,对于"异常",至今仍无统一定义,但对其特征,许多研究者却给出了相似的"描述"[1,2],Han 等人认为:异常是指那些与数据集中大多数样本极度不一致的样本[1]. 基于上述描述,在异常检测中最常用的思想是:设计一个描述数据集信息的模型(如高斯模型),认为不符合此模型的样本为"异常"[1,2].

收稿日期:2004-10-27;修回日期:2004-11-01
基金项目:国家"九七三"重点基础研究发展规划基金项目(G1998030500)

但是,在复杂领域问题中,异常信息和正常信息高度混杂,异常信息中可能包含部分正常信息,反之亦然. 例如,在股票市场中,日收盘价由随机波动和长期趋势共同组成[3]. 因此,在这些复杂数据集中,正常数据和异常数据的区别并不如上述描述所假设的那样明显,传统的方法在检测此类异常数据时存在一定困难,需要一种更有效的描述异常的方法及相应的检测模型. 为此,本文提出了一种基于方差的异常检测模型(variance-based outlier detection model,VODM). 此模型首先把数据集的总方差分解为估计方差和残差两部分,然后检测前 k 个包含最大残差的样本为异常. 与传统的异常检测模型相比,VODM 的最大优点是模型中同时描述了正常和异常信息,这对提高异常检测的精度是有益的. 但 VODM 只是一种理论框架,为此,本文随后讨论了为其设计算法的两种指导原则——基于分布和直接基于数据,并给出了基于数据的主曲线算法[4]. 上交所综合指数(INDEXSH)异常收益检测的实验结果表明,VODM 及其主曲线算法是有效的.

2 基于方差的异常检测模型(VODM)

令 $D\subseteq R^d$ 为样本空间,W 是一参数集合.

定义 1. 设 $T\subseteq D(\mathrm{card}(T)=n)$ 是与 D 独立同分布的一组样本子集,$X=(x_1,x_2,\cdots,x_d)\in T$. 如果存在 T 上的一个函数 f,使

$$X = f(w) + \varepsilon, \tag{1}$$

其中,$f(w)=(f(x_1,w),f(x_2,w),\cdots,f(x_d,w))$,$w\in W$,且 $f(w)$ 是 X 的正常部分,ε 是 X 的残余部分,则称 X 是可分解的.

异常检测的具体任务就是把 X 分解为 $f(w)$ 和 ε. 其中 X 的正常部分 $f(w)$ 是稳定和确定的,但其异常部分 ε 受各种因素影响是不确定的. 因此可以把原任务转换为从已知数据集 T 中求解 $f(w)$,目标是最小化 $\| T-f(w)\|^2$.

定义 2. 设 $f(x)=E(X\mid w)$,其中,

$$f(x_j,w) \overset{\mathrm{def}}{=} E(x_j \mid w),\ j=1,2,\cdots,d, \tag{2}$$

则 $f(w)$ 称为 X 在条件 w 下的估计值.

定义 3. 平方距离. 距离

$$\varepsilon(X,w) \overset{\mathrm{def}}{=} \sum_{j=1}^{d}(x_j - E(x_j \mid w))^2 \tag{3}$$

称为 $f(w)$ 和 X 之间的欧几里德平方距离.

定义 4. 期望平方距离. 距离

$$\Delta(f) \overset{\mathrm{def}}{=} E[\| T - E(T \mid w)\|^2] \tag{4}$$

称为 $f(w)$ 的期望平方距离.

定理 1. VODM. 设 $X=(x_1,x_2,\cdots,x_d)\in T$,$w\in W$. 在条件 w 下,X 的总方差可被分解为估计方差和残差两部分,即

$$\sum_{j=1}^{d} Var(x_j) = \sum_{j=1}^{d} Var(E(x_j \mid w)) + \sum_{j=1}^{d} E(x_j - E(x_j \mid w))^2. \tag{5}$$

定理 1 的证明见文献[5].

显而易见,估计方差表达的是数据信息的正常部分,而残差反映的是其异常部分.

为了最大可能估计出数据集信息的正常部分,在定理 1 的条件下,应使 $\Delta(f)$ 最小. 这样,可以自然地定义异常如下.

定义 5. 异常. 根据定理 1,在最小化期望平方距离的条件下,

$$Outlier \overset{\mathrm{def}}{=} \underset{X_i}{\mathrm{argmax}} \| X_i - E(X_i \mid w)\|^2,\quad i=1,2,\cdots,n, \tag{6}$$

称为在样本集 $T=\{X_1,X_2,\cdots,X_n\}$ 中的异常,其中 n 代表样本的个数.

定义 5 的含义是异常为具有最大残差的样本. 实际问题中,可以取前 $k(k<n)$ 个具有最大残差的样本构成异常集合.

3 VODM 的算法实现

为 VODM 设计算法时,一般可遵循两个原则. 如果事先知道数据的分布形式,那么根据分布函数构造算法是最好的选择. 否则,依照 Vapnik 原则:"在解决一个给定的问题时,要设法避免把解决一个更为一般的问题作为其中间步骤[6]". 因此在这种情况下,最好设计一个基于数据的算法而不要试图估计数据的分布. 本文的目的是解决复杂领域问题的异常检测,因此我们为 VODM 采用了基于数据的主曲线算法.

3.1 主曲线

主曲线是通过数据集"中间"的光滑无参曲线[4]. 为便于理解,本节概述主曲线基本原理.

定义 6[4]. 曲线 $f:\Lambda\mapsto R^d$ 是 d 维欧几里德空间中连续函数,其中 $f=(f_1,f_2,\cdots,f_d)$,$\Lambda=[a,b]\subset R^1$.

根据定义 6,曲线 f 可以看作是单变量 $\lambda\in[a,$

$b]\subset R^1$ 的 d 维函数向量,即 $f(\lambda)=(f_1(\lambda),f_2(\lambda),\cdots,f_d(\lambda))$,其中 $f_1(\lambda),f_2(\lambda),\cdots,f_d(\lambda)$ 称为坐标函数.

定义 7[4]. 设 $T\subseteq D$. 对于任意 $X\in T$,

$$\lambda_f(X)\overset{\text{def}}{=}\sup\{\lambda:\|X-f(\lambda)\|=\inf_{\tau}\|X-f(\tau)\|\},\tag{7}$$

称为其相应的投影指标. 其中 $\tau\in[a,b]\subset R^1$,$f(\lambda)$是由参数 $\lambda\in R^1$ 控制的 T 上的曲线.

投影指标 $\lambda_f(X)$是指 X 投影到曲线 f 上最短正交距离处 $f(\lambda)$的参数 λ 值,如果存在多个最小值,则取其中最大的一个. 相应地,X 到 f 上的投影点是 $f(\lambda_f(X))$.

定义 8[4]. X 到其在曲线 f 上投影点的平方距离

$$\varepsilon(X,f)\overset{\text{def}}{=}\|X-f(\lambda_f(X))\|^2\tag{8}$$

称为曲线 f 和 X 之间的欧几里德平方距离.

定义 9[4]. 曲线距离. 设 $T\subseteq D$,距离

$$\Delta(f)\overset{\text{def}}{=}E[\|T-f(\lambda_f(T))\|^2]\tag{9}$$

称为 T 和曲线 f 之间的期望平方距离.

定义 10[4]. 主曲线. 设 $T\subseteq D$. 如果定义 6 中的光滑曲线 $f(\lambda)$满足如下条件:①f 不自相交;②f 在 T 的任何有界子集中长度有限;③f 是自相合的,即

$$f(\lambda)=E(T\mid\lambda_f(T)=\lambda),\ \forall\lambda\in\Lambda\subset R^1,\tag{10}$$

则称其为主曲线.

自相合是指曲线上的每一点是数据集中投影到该点的样本点的期望平均. 因此,主曲线是指满足自相合特性的光滑无参曲线,它通过数据集的"中间"且提供对数据集信息的非线性的描述[4](如图 1 所示).

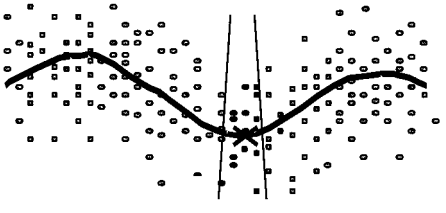

Fig. 1 Self-consistency of the principal curve.

图 1 主曲线的自相合特性

3.2 算法实现

定理 2. 设 $T\subseteq D$,$X=(x_1,x_2,\cdots,x_d)\in T$,$w\in W$. 如果 $f(\lambda)$为一主曲线,且 $w=\lambda_f(X)$,那么,

$$\sum_{j=1}^{d}Var(x_j)=E\|X-f(\lambda_f(X))\|^2+\sum_{j=1}^{d}Var(f_j(\lambda_f(X))).\tag{11}$$

定理 2 的证明见文献[5].

式(11)的含义为 X 的总方差可被分解为由主曲线表示的估计方差和由 X 到它在曲线上投影的平方距离表示的残差两部分. 因此,定理 2 表明主曲线可以作为 VODM 的一种实现算法.

主曲线是曲线距离式(9)取得的最小值时的曲线[4],因此根据定义 5,在主曲线意义下,数据集 $T=\{X_1,X_2,\cdots,X_n\}$中的异常可定义如下:

$$Outlier=\arg\max_{X_i}\|X_i-f(\lambda_f(X_i))\|^2,\quad i=1,2,\cdots,n.\tag{12}$$

实际问题中,可以先把所有样本按 $\|X_i-f(\lambda_f(X_i))\|^2(i=1,2,\cdots,n)$降序排序,然后取前 $k(k<n)$个样本构成异常集合.

根据定义 10、定理 2 和式(12),可得到如下基于主曲线的异常检测算法:

Step1. 令初始曲线 $f^{(0)}(\lambda)$为 T 的第 1 主成分线. 设 $j=0$.

Step2. (投影步)对所有 $X\in T$,求 $\lambda_f^{(j)}(X)=\max\{\lambda:\|X-f(\lambda)\|=\min_{\tau}\|X-f(\tau)\|\}$.

Step3. (期望步)求 $f^{(j+1)}(\lambda)=E[T\mid\lambda_f^{(j)}(T)=\lambda]$.

Step4. 如果$(1-\Delta(f^{(j+1)})/\Delta(f^{(j)}))<\delta$,则转到 Step5,否则令 $j=j+1$,转到 Step2,其中 δ 为预先给定的阈值.

Step5. 将所有样本按值 $\|X_i-f(\lambda_f(X_i))\|^2(i=1,2,\cdots,n)$降序排序.

Step6. 取前 $k(k<n)$个样本为异常点集合.

上述算法是针对连续分布数据的,然而在实际中,我们只能得到有限数据集,在这种情况下,根据主曲线的自相合特性,最多只有一个点投影到曲线上的给定点,这会导致上述算法在 Step3 后遍历所有数据点. 为解决此问题,可采用 cubic 样条函数来近似自相合条件并对主曲线做光滑处理,即在限定弧长为单位长度($\lambda\in[0,1]$)的条件下,算法通过最小化式(13)来实现期望步(Step3):

$$\Delta_n(f)=\frac{1}{n}\sum_{i=1}^{n}\|X_i-f(\lambda_i)\|^2+B\int_0^1\|f''(\tau)\|^2\mathrm{d}\tau,\tag{13}$$

其中,$i=1,2,\cdots,n$,B 为固定光滑因子,$f''(\tau)$为

曲线在投影指标处的二阶导数.具体的优化方法可参见文献[7].

4 实 验

本文提出的 VODM 及其主曲线算法可以应用在许多领域,其中之一就是股票市场中异常收益的检测.检测异常收益的传统方法是假设收益率服从正态分布[8],并认为落在 $\mu \pm 3\sigma$(μ,σ 为正态分布的均值和方差)之外的样本为异常.但众多研究表明,股票市场的收益率并不总是服从正态分布的,而是具有高峰度和高偏度的特征[9].另一种检测异常收益的常用方法是 GARCH 模型[10].此模型在市场环境相对稳定的条件下比较有效,但却不适合市场波动比较大的情况[11].

本节将通过上交所综合指数(INDEXSH)异常收益检测的实验验证 VODM 及其主曲线算法是有效的.实验中采用 INDEXSH 在 1998 年 1 月 1 日至 2001 年 12 月 31 日期间内日收益率作为实验数据(如图 2 所示).实验数据来自 http://finance.yahoo.com/?u.

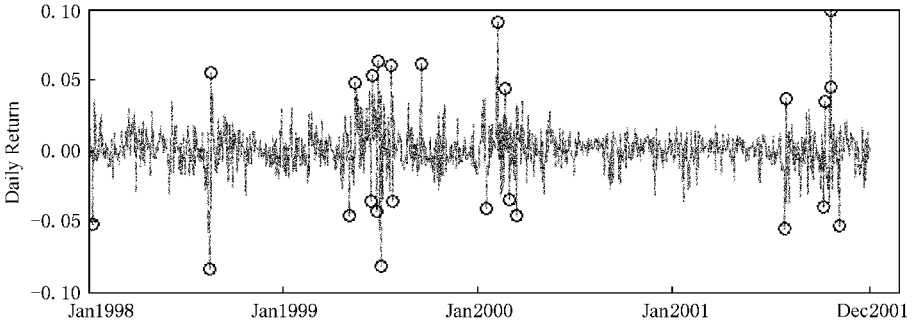

Fig. 2 INDEXSH: The daily time series and the top 25 outliers identified by stock analyst.

图 2 上交所综合指数每日收益率序列及分析师指定的 25 个异常点

4.1 数据前处理

(1)构造 5 日、20 日及 60 日累计收益率

一般而言,由不同因素引起的异常收益将持续不同的时间,因此有必要区别检测短期和长期异常收益.我们通过对单日收益率和 5 日(1 周)累计收益率来检测短期异常收益,而通过 20 日(1 个月)和 60 日(3 个月)的累积收益率来检测长期的异常收益.5 日、20 日及 60 日累计收益率数据集可采用类似移动平均的方法从日收益率集合 $R=\{r_1,r_2,\cdots,r_n\}$ 得到.具体方法如下:

$$r_i^M=\sum_{j=i}^{i+M-1} r_j,\ M=5,20,60, \tag{14}$$

其中,$i=1,\cdots,n-M+1$,r_i^M 是 M 日累计收益率,n 是日收益率样本个数.因此可得到 $R^M=\{r_1^M,r_2^M,\cdots,r_{n-M+1}^M\}$,$M=5,20,60$.

(2) 为收益率增加序列信息

对于给定的收益率集合 $R=\{r_1,r_2,\cdots,r_n\}$(或 $R^M,M=5,20,60$),为在主曲线算法下保持收益率样本的序列性,本文采用了 Reinhard 等人[12]提出的 TC-PCA (time constraint principal component analysis)方法.此方法把原收益率数据通过式(15)添加一维:

$$r_{i\eta}=(i\times\eta,r_i),i=1,2,\cdots,n. \tag{15}$$

因此得到 $R_\eta=\{(\eta\times1,r_1),(\eta\times2,r_2),\cdots,(\eta\times n,r_n)\}$.同理也可由 R^M 得到 R_η^M($M=5,20,60$).在 R_η 和 R_η^M($M=5,20,60$)中,添加的一维表示对收益率样本增加变尺度的时间序列约束,其中可调因子 η 表示时间序列约束的权重.TC-PCA 最终把 R_η(或 $R_\eta^M,M=5,20,60$)变换到正交坐标系中,其目标是调节 η 使第 1 主成分线尽量和时间轴平行.

4.2 标准异常

为使主曲线算法的异常检测结果有一个可衡量的标准,本文请股票分析师从实验数据 R 和 R^M($M=5,20,60$)中分别指定 25,20,15 和 10 个异常点,构成集合 O_{25},O_{20},O_{15}和 O_{10},其中 O_{25}的结果示意在图 2 中.上述 25,20,15 和 10 是在假设异常点只占总样本很小比例的前提下,由分析师根据其领域知识确定的.另外,由于 R^5,R^{20},R^{60}是累计的单日收益率,所以 O_{20},O_{15}和 O_{10}中的异常点是某个时间段,而且时间上可能会有部分重叠.

4.3 主曲线算法实验结果

为验证主曲线算法异常检测的正确率,我们从

R_η 和 R_η^M($M=5,20,60$)中同样检测出25,20,15和10个异常点. 在此,两个参数:式(15)中的 η 和式(13)中的 B,需要进行进一步的实验分析.

在实验中我们发现,当 η 的取值为收益率的两个数量级以上时,则对检测结果影响很小,不失一般性,实验中令 $\eta=2.0$. 同样,在实验中我们采用不同的 B 来验证其对检测结果的影响,结果表明光滑因子 B 在一定的取值范围内对检测结果没有重大影响. 在此,我们令 $B=0.001$,由主曲线算法从集合 R_η 和 R_η^M($M=5,20,60$)中分别得到异常点集合 O_{25}^B,O_{20}^B,O_{15}^B和 O_{10}^B,其中图3是 O_{25}^B的结果. 然后对 O_{25}^B,O_{20}^B,O_{15}^B和 O_{10}^B中的异常点进行合并,最后再与当时的宏观政策与事件进行比较,查出造成异常收益的原因,结果如表1所示.

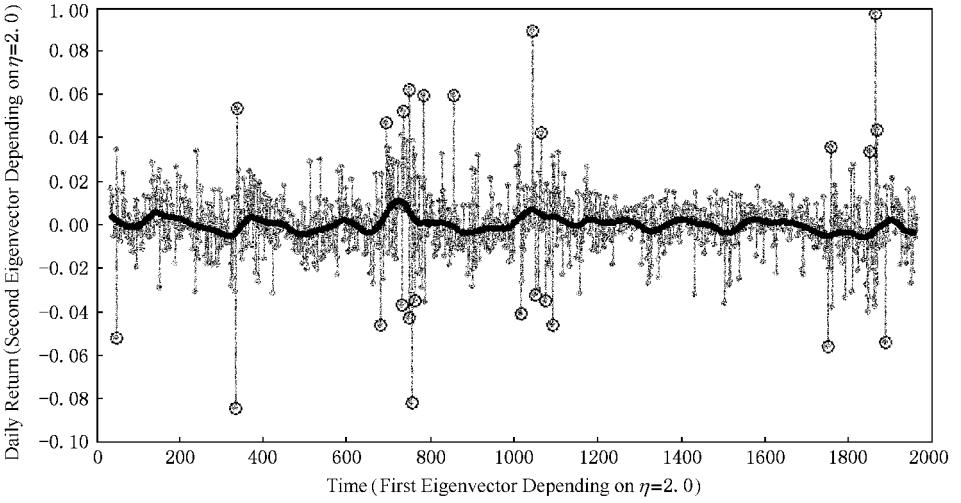

Fig. 3 The outliers detection result by the principal curve to R_η.

图3 主曲线对 R_η 的异常检测结果

Table 1 The External Events of INDEXSH

表1 上海大盘的异常事件检测结果

No.	Time	Duration				Events
		1	5	20	60	
1	19980113	Y				The Currency of Southeast Asian countries was impacted.
2	19980612-19980915	Y	Y	Y	Y	① The impact of Asia Financial Crisis on Southeast Asia; ② An unprecedented big flood occurred in China on Aug.
3	19990510	Y				US-led NATO attacked Chinese Embassy in Yugoslavia On May 9.
4	19990511-19990629	Y	Y	Y	Y	The six policies of "invigorating the market" caused the 5.19 market.
5	19990701-19990810	Y	Y	Y	Y	"Securities Act" formally took effect from 7.1.
6	20000104-20000111	Y	Y			General Office of the State Council announced the notice about several suggestions of setting up the risk investment mechanism.
7	20000125-20000316	Y	Y			The secondary market rationed new shares and stocks were pledged.
8	20001124	Y				China Securities Regulatory Commission issued notice on strengthening market supervision and beating market manipulation behavior; meanwhile nine inspection bureaus were to be established and 236 illegal cases in violation of rules and regulations were to be investigated, etc.
9	20010724-20010801	Y	Y		Y	The People's Bank of China investigated the funds in violation of rules and regulations.
10	20010918-20011022	Y	Y	Y	Y	① The reduction of state-owned stocks; ② Yinchuan Guangxia fake incident; ③ The 911 incident occurred in U.S.A.
11	20011023-20011107	Y	Y			The reduction of state-owned stocks was suspended.

Note: "Y" means there is an abnormal return in one of the daily, 5, 20 and 60 aggregate days' return sets. The announcements come from http://www.csrc.gov.cn/, the website of Chinese Securities Regulatory Commission.

从表 1 可知,在 11 次异常收益中,只有 1 次是由政治事件引起的,3 次与经济事件有关,而 7 次则是由股票市场的宏观调控政策引起的. 这说明,虽然宏观政治因素和经济因素也能引起股市的异常收益,但政府对股市的宏观调控政策是导致上海股票市场大盘发生异常的主要原因.

对于本实验,作者声明:本实验表 1 的分析结果只用于算法测试,除此之外没有任何其他意图. 表 1 的分析结果表明,本文提出的方法在解决异常收益检测的问题中是可行的.

5 总 结

本文研究了如何从异常信息和正常信息高度混杂在一起的复杂领域问题中检测异常,并提出了基于方差的异常检测模型(VODM)及其相应的主曲线算法. 股票异常收益检测实验表明,VODM 及其算法是有效的. 但这并不能说明 VODM 是完美的,因为此模型在理论上仍然存在很多缺陷,可以探讨的问题还很多,该模型在其他不同应用领域的有效性需要进行进一步的研究和实验验证.

参 考 文 献

1 Jiawei Han, M. Kamber. Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann, 2001

2 V. Barnett, T. Lewis. Outliers in Statistical Data. New York: John Wiley & Sons Inc., 1994

3 M. Last, Y. Klein, Abraham Kandel. Knowledge discovery in time series databases. IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics, 2001, 31(1): 1~9

4 T. Hastie. Principal curves and surfaces. Laboratory for Computational Statistics, Department of Statistics Stanford University, Tech. Rep.: 11, 1984

5 Hongwei Qi, Jue Wang. A model for mining outliers from complex data sets. The 19th Annual ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004

6 V. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995

7 B. W. Silverman. Some aspects of spline smoothing approaches to non-parametric regression curve fitting. Journal of the Royal Statistical Society, Ser. B, 1985, 47(1): 1~52

8 J. Smith. The probability distribution of market returns: A logistic hypothesis: [Ph. D. dissertation]. Utah: University of Utah, 1981

9 C. Corrado, Tie Su. Implied volatility skews and stock return skewness and kurtosis implied by stock option prices. The European Journal of Finance, 1997, 3(1): 73~85

10 P. Franses, D. Dijk. Outlier detection in GARCH models. Econometric Institute, Tech. Rep.: EI-9926/RV, 2000

11 C. Gourieroux. ARCH Models and Financial Applications. Berlin: Springer-Verlag, 1997

12 K. Reinhard, M. Niranjan. Parametric subspace modeling of speech transitions. Speech Communication, 1999, 27(1): 19~42

Qi Hongwei, born in 1975. Ph. D. in patten recognition and intelligence system. He currently focuses on text/Web mining, and his general interests include machine learning and data mining.

齐红威,1975 年生,博士,主要研究方向为机器学习和数据挖掘.

Zhang Junping, born in 1970. Ph. D. in patten recognition and intelligence system. He currently focuses on text/Web mining, and his general interests include manifold learning and data mining.

张军平,1970 年生,博士,主要研究方向为流形学习和数据挖掘.

Wang Jue, born in 1948. Professor in the Institute of Automation, the Chinese Academy of Sciences. His main research interests are machine learning and artificial intelligence.

王珏,1948 年生,研究员,博士生导师,主要研究方向为计算机科学与人工智能的研究.

Research Background

Being an integral part of data mining and having attracted much attention recently, outlier mining aims at revealing the nontrivial knowledge, i.e., abnormal information, hidden in data. In this paper, to solve the outlier detection problems where outliers highly intermix with normal data, a general variance-based outlier detection model (VODM) is presented, in which the information of data is decomposed into normal and abnormal components according to their variances. With minimal loss of normal information in the model, outliers are viewed as the top k samples holding maximal abnormal information in a dataset. The VODM is a theoretical framework, and then, the principal curve is introduced as an algorithm of it. Experiments carried out on abnormal returns detection in stock market show that the VODM is feasible. This research is supported by the National Key Project for Basic Research in China (G1998030508).