

基于小波变换的例外挖掘

Wavlet Transform-based Outlier Mining

周丽华 王丽珍

(云南大学信息学院计算机科学与工程系 昆明650091)

Abstract In real project, the useful signal usually shows as low-frequency signal, but noise (outlier) shows as high-frequency signal. Wavlet transform can maps 1-D signal into a 2-D time-scale plane, and the localization of signal is described in different scales, so it is suit to detect instantaneous strange embedded in normal signal. The notion of TS (ω) outlier and CL (δ) outlier, a kind of wavlet transform-based outlier mining approach are introduced. Experiments results show its efficiency and effectiveness.

Keywords Wavlet transform, Outlier mining, Time series outlier, Cluster-based outlier

1 引言

例外常常是指数据集中远远偏离其它对象的那些小比例对象,在大多数研究中作为噪声被遗弃。但是在一些应用中,例外的检测能为我们提供比较重要的信息,使我们发现一些真实而又出乎预料的知识,具有很高的实用价值,如文[1]挖掘了时序数据中的例外,从而在存储容量相同的情况下,可以获得对原始序列更精确的表示。

目前例外挖掘方法可以广义地分为3类^[2,3]:基于分布的方法、基于深度的方法及基于距离的方法。基于深度的方法避免了基于分布的方法中的分布适应性问题,扩展了例外挖掘的应用,但已有的基于深度的方法只有在数据维数 $k \leq 2$ 时,效果才被人们接受。基于距离的方法不受数据维数的限制,超越了数据空间,它仅仅依赖于米制距离函数的距离值计算。

在实际工程中,有用信号通常表现为低频信号或是一些比较平稳的信号,而噪声(例外)则通常表现为高频信号。文[4]将数据集中每个元组看成是 n 维特征空间(由元组的 n 个属性构成)中的一个点,将所有元组的集合看成是一个 n 维信号,因此用信号处理的方法来处理数据。受其启发,本文用小波变换来挖掘例外。小波变换是一种信号的时间-尺度分析方法,具有多分辨率分析的特点,而且在时频两域都具有表征信号局部特征的能力:在低频部分具有较高的频率分辨率和较低的时间分辨率,在高频部分具有较高的时间分辨率和较低的频率分辨率,被誉为分析信号的显微镜,很适合于探测正常信号中夹带的瞬态反常现象,并且小波变换具有快速、高效的特点,因此用小波变换来挖掘例外是比较有效的。

2 小波变换的多分辨率分析^[5,6]

设 $h = \{h_k\} \in l^p(R)$, $-M \leq k \leq M$ 满足下述条件:(1) $\sum_k h_k = 2^{-1/2}$; (2) 对所有的 $l \neq 0$, $\sum_k h(k)h(k+2l) = 0$; (3) 存在 $q > 0$, $\sum_k |h(k)|^q < \infty$, 则有 $g = \{g_k\}$, $g_k = (-1)^k h(-k+1)$ 。其中 h, g 称为长度为 $R=2M$ 的镜像滤波器,和 h, g 相联系,有尺度函数 φ 和小波函数 Ψ ,它们具有以

下性质:

$$\varphi(x) = 2^{\frac{1}{2}} \sum_k h(k) \varphi(2x-k), \hat{\varphi}(0) = 1 \quad (1)$$

$$\Psi(x) = 2^{\frac{1}{2}} \sum_k g(k) \varphi(2x-k) \quad (2)$$

由 Ψ 伸缩和平移形成 $\Psi_{jk}(x)$, $j, k \in Z$ 构成一个 $l^p(R)$ ($p \geq 1$) 的规范正交基。相应于镜像滤波器 h, g , 定义算子 H, G 为:

$$Hf(i) = \sum_k h(k-2i)f(k), Gf(i) = \sum_k g(k-2i)f(k) \quad (3)$$

由此得到它们的对偶算子 H^*, G^* 为

$$H^*f(k) = \sum_i h(k-2i)f(i), \quad (4)$$

$$G^*f(k) = \sum_i g(k-2i)f(i)$$

上述算子具有以下性质:(1) $H^*H + G^*G = I$; (2) $HG^* = GH^* = 0$; (3) $l^p = H^*l^p \oplus G^*l^p$ 。

以上即为小波分析的分解与重构公式,原信号经 H 作用得到平滑版本(Averages),经 G 作用得到细节版本(Details),重复上述过程,对其平滑版本再进行小波变换,得到它的更平滑版本和一系列细节版本,即

$$l^p = \bigoplus_j (H^*)^j G^* l^p \oplus (H^*)^j l^p. \quad (5)$$

小波分解如图1所示。其中, $[H], [G]$ 表示信号与滤波器的卷积; $[\downarrow 2]$ 表示偶抽样。

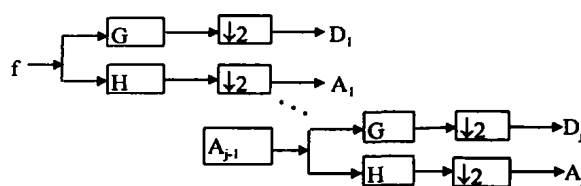


图1 小波分解示意图

$$\text{由(5)式可得 } f = A_j + \sum_{i=1}^j D_i \quad (6)$$

上式表明,任何函数 $f \in l^p(R)$ 都可以根据分辨率为 2^{-j} 时 f 的低频部分 A_j (平滑版本) 和分辨率为 2^{-j} 以下的 f 的高频部分 $D_i, i=1, 2, \dots, j$ (细节版本) 完全重构。

小波变换具有能量守恒的性质,它不仅是全局能量守恒

*)基金项目:云南省自然科学基金资助项目(1999F0015M),周丽华领域为数据库与面向对象技术。

讲师,主要研究领域为模式识别和图像处理。王丽珍 副教授,主要研究

的,如果考虑所有伸缩小波构成的空间支持,则小波变换也是局部能量守恒的,总能量可被分解为不同尺度 j 的能量。小波变换允许局部重构,即可以只重构信号的一部分或仅对给定的尺度范围来重构信号,反映其对该范围的局部贡献。

3 时序数据的例外挖掘

时序中的例外一般是指那些与邻近点相比,其幅度变化较大,而持续时间较短的点,将它们从序列中移走后,序列变得很平滑,可以获得较简洁的表示。这些例外不一定是序列中的极值点,其变化和持续时间是针对某一时段而言,具有局部性。利用小波变换的空间局部化性质,可以有效挖掘时序数据中的例外。

由于例外表现为高频信号,用小波变换对信号进行分解后,关于例外的信息应包含在一系列细节版本中。当然,对于时序信号,除了例外点而外,其余数据点的值也不完全相同,这些值是上下波动的,只是相对于例外点而言,这些波动较小。在小波分解过程中,这些波动随着分解层次的加深也被分解到细节版本中。但由于例外点的模值大于波动点的模值,例外点的能量是高频能量的主成分,故可以使用阈值将例外与波动区分开。

定义1(时序例外 $TS(\omega)$) 设时间序列为 $X = x_1 x_2 \cdots x_N$,用小波变换对 X 进行分解,分辨率为 2^{-j} 时 X 的低频部分记为 A_j , 2^{-j} 以下的高频部分记为 $D_i = d_{i1} d_{i2} \cdots d_{iN}$, $i = 1, \cdots, j$ 。所有高频部分的和 $D = \sum_{i=1}^j D_i = d_{11} d_{12} \cdots d_{jN}$ 。给定阈值 ω (例外数据在高频信号中的能量比),根据 d_i 的绝对值大小将 $d_{11} d_{12} \cdots d_{jN}$ 降幂排列成 $|d_1| \geq |d_2| \geq \cdots \geq |d_N|$, 若 $\sum_{i=1}^N |d_i|^2 / \sum_{i=1}^N |d_i|^2 > \omega$, 则 $th = |d_k|$ 。如果 $|d_i| \geq th$, 则序列 X 中的点 x_i 就称为序列 X 的一个时序例外 $TS(\omega)$ 。

挖掘 $TS(\omega)$ 例外的算法 FindTS 如下:

输入:时间序列 X , 阈值 ω
输出:时序 X 中的 $TS(\omega)$ 例外
1 对时间序列 X 进行尺度为 j 的小波分解;
2 计算各尺度高频信号之和 $D = \sum_{i=1}^j D_i = d_{11} d_{12} \cdots d_{jN}$;
3 计算区分例外与波动的阈值 th
根据 d_i 的绝对值大小将 $d_{11} d_{12} \cdots d_{jN}$ 降幂排列成 $|d_1| \geq |d_2| \geq \cdots \geq |d_N|$,
$$\frac{\sum_{i=1}^N |d_i|^2}{\sum_{i=1}^N |d_i|^2} > \omega \quad \text{则 } th = |d_k|$$

4 区分例外与波动
如果 $|d_i| \geq th$, 则报道点 x_i 是一个例外。

语句1的时间复杂度为 $O(lN) = O(N)$, l 是一个小的常数,表示小波变换中所使用的滤波器的长度, N 是时间序列长度;语句2,4的时间复杂度为 $O(N)$;算法的花费主要在语句3的排序上,其时间复杂度为 $O(N \log_2 N)$, 故整个算法的时间复杂度为 $O(N \log_2 N)$ 。

文[1]对时序例外的挖掘是在限定存储容量 B 的条件下进行的。若将一组点从序列中移走后,能使序列的直方图表示误差减到最小,则被移走的这组点称为序列的最优例外集,为便于描述,我们将其称为 XBD 例外集。为了找到最优例外集,文[1]使用了一种动态程序算法,避免了点与点的庞大组合。该算法时间复杂度为 $O(N^2(B-k)k^2)$, k 是例外点数目。实际上,只有将序列中变化幅度较大的点移走,才有可能使序列的直方图表示误差减小。本文挖掘的 $TS(\omega)$ 例外正好是原序列

中变化幅度较大的点,由所有 $TS(\omega)$ 例外构成的集合是 XBD 例外集的超集,因此 XBD 例外集的挖掘可在 $TS(\omega)$ 例外集的基础上进行:从序列中移走点时只移走 $TS(\omega)$ 例外集中的点。由于 $TS(\omega)$ 例外集中点的数目远远小于序列长度 N , 因此只移走 $TS(\omega)$ 例外集中的点可避免对许多非例外点的检测,从而大大降低算法的时间复杂度。例如文[1]中的序列 $X = (38, 53, 33, 390, 210, 371, 46, 72, 174, 47, 373, 21, 30, 107, 46)$, 在限定 $B = 4$ 的条件下, XBD 例外集只包含 373 一个点。用本文算法 FindTS 在 X 上挖掘到的 $TS(\omega)$ 例外集为 $(390, 371, 373)$, 在 $TS(\omega)$ 例外集上只需进行 7 次移动即可找到 XBD 例外集。

4 多维数据的例外挖掘

对于多维数据的例外,不能简单地使用小波分解的高频部分进行挖掘,因为此时的高频部分除了包含例外信息,还包含边缘信息,若直接用高频信号来挖掘例外,则边缘也被当作例外。文[7]将例外的挖掘问题看成是一种特殊的两类别(例外类和非例外类)的聚类问题,由此本文提出了基于聚类的例外 $CL(\delta)$ 。

定义2(基于聚类的例外 $CL(\delta)$) 设数据集 $T = \{t_1, t_2, \cdots, t_N\}$, t_i 称为 T 的第 i 个元组,将 T 中的元组聚类成若干聚类族 $C = \{c_1, c_2, \cdots, c_r\}$, 属于聚类族 c_i 的元组数目称为该聚类族的族码 $|c_i|$ 。给定阈值 δ , 若聚类族 c_i 的族码 $|c_i| < \delta$, $1 \leq i \leq r$, 则该聚类族称为 T 的一个例外集,属于该聚类族的所有元组称为 T 的聚类例外 $CL(\delta)$ 。

由定义可知,要想挖掘到正确而稳定的 $CL(\delta)$ 例外,则必须要求聚类结果正确、稳定,因此要求所使用的聚类算法具有:能识别任意形状的聚类族;去噪声能力强;对数据的输入顺序不敏感等优良性质。文[4]提出了一种新的聚类算法——基于小波变换的波聚类算法,该算法满足上述要求。算法的主要思想为:一个具有 n 个属性的元组,它的 n 个属性值构成一个 n 维的特征矢量,该矢量可看成是 n 维特征空间中的一个点,数据集中所有元组的集合构成一个 n 维信号,信号的高频成分对应于元组急剧变化的区域,即聚类族的边界;具有较大模值的低频成分对应于元组集中分布的区域,即聚类族。因此用小波变换将信号变换到频域后,通过寻找平滑版本的密集区域来标识聚类。在波聚类算法中采用了格的技术:将 n 维特征空间的每一维 i ($1 \leq i \leq n$) 分成 m 个间隔,整个特征空间共有 $K = m^n$ 个格。根据各元组的属性值大小,将各元组分配到相应的格中,小波变换针对格进行。数据输入后,处理时间仅是 K 的函数,独立于元组数 N , 由于 $N \gg K$, 所以波聚类算法对元组数较多,而维数、间隔数较少的数据库效率特别高。波聚类算法除具有上述所有优良特性外,还具有多解特性。本文用波聚类算法对数据集聚类,用波聚类的结果挖掘例外。由于要计算各聚类族的族码,因此需要将波聚类算法的第6步(给各元组分配类别标识)修改为:

```
6.1  $\forall c_i \in C, |c_i| = 0$  //各聚类族族码初始化
6.2  $\forall t_j, l_{ij} = -1, j = 1, 2, \cdots, N$  //将所有元组类别标识初始化为例外
6.3  $\forall c_i \in C, \forall M_j, \forall u_k \in M_j, l_{ik} = c_i, |c_i| = |c_i| + 1$  //给各元组类别标识赋值并计算各聚类族的族码
```

挖掘 $CL(\delta)$ 例外的算法 FindCL 如下:

输入:多维数据对象的特征矢量, 阈值 δ
输出: $CL(\delta)$ 例外
1 wavecluster(); //调用波聚类算法标识聚类元组,同时计算各聚类族的族码
2 for $i = 1$ to N

if $(l_i = -1 \text{ or } l_i = c_k \text{ but } |c_k| \leq \delta)$
 then 报道元组 t_i 是一个例外。

由于波聚类算法具有很强的去噪声能力,能识别任意形状的聚类族,因此数据集中的小比例元组要么被当成噪声去掉,其标识 l_i 的初值不会被改变($l_i = -1$);要么其所属聚类族 c_k 的族码 $|c_k|$ 小于阈值 δ 。而大比例元组所属聚类族的族码大于阈值 δ 。由于波聚类算法的时间复杂度为 $O(N)$,遍历数据库的次数为1次(读入元组并将其分配到相应格),因此整个算法的时间复杂度为 $O(N)$,遍历数据库的次数为1次。文[2]中挖掘 DB(P,D) 例外的 NL 算法的时间复杂度为 $O(kN^2)$ (k 是特征维数),遍历数据库的次数 $\geq n-2$ ($n = \lceil 200/B \rceil$, B 是缓冲区的百分比);CS 算法的时间复杂度与 N 成线性关系,遍历数据库的次数最多为3次。由此可见,本文算法 FindCL 优于 NL 算法及 CS 算法。

5 实验

图2是对一个实际时间序列应用算法 FindTS 的实验结果,图中的时间序列 X 为一股票的日成交量。实验中采用了双正交样条小波滤波器,例外数据在高频信号中的能量比 $\omega = 80\%$ 。结果表明,原序列 X 中的局部偏离被准确定位。又由于算法只需进行一次小波分解就可将所有 $TS(\omega)$ 例外同时挖掘出来,故算法具有高效、正确的特点。

随着分解层数的加深,低频部分的频率分辨率增大,使得高频部分中的能量加大,可以根据需要在不同的分解层上挖掘例外(图2给出了在尺度3和尺度5上挖掘到的例外),因此算法具有多解的特性。

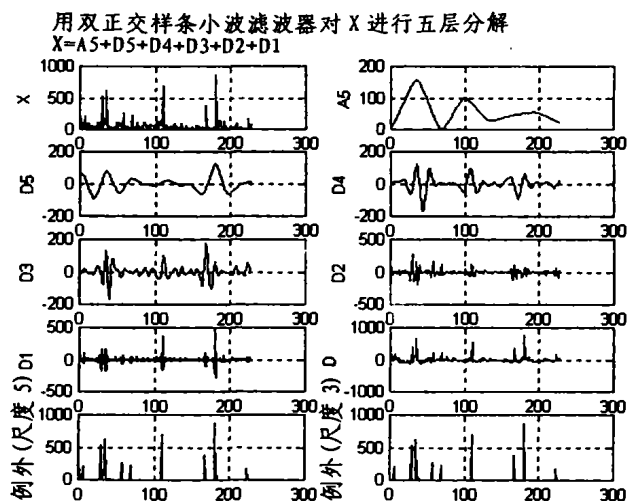


图2 用小波变换对时序进行分解的结果及挖掘到的例外

为便于可视化,图3利用算法 FindCL 对2维图像进行例外挖掘,图像上的每一个像素对应于波聚类算法中的一个格。图中显示了波聚类的结果及算法 FindCL 挖掘到的例外。结果表明,波聚类算法具有很好的聚类结果,基于波聚类算法挖掘到的例外与原图中的例外非常吻合。

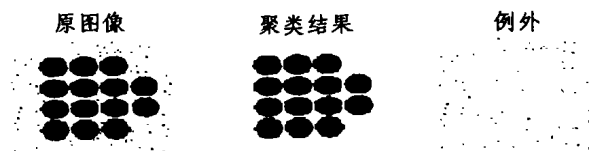


图3 波聚类的结果及基于波聚类挖掘到的例外

结束语 例外挖掘是一项重要的、有意义的工作。本文提出了 $TS(\omega)$ 例外和 $CL(\delta)$ 例外的概念,并用小波变换的多分辨率分析性质对它们进行挖掘。实验表明,无论对时序数据还是高维数据,基于小波变换的例外挖掘方法都具有较高的效率和较好的结果。

例外数据的挖掘包括例外数据的发现以及例外数据的分析^[8],本文仅仅完成了例外数据的发现,因此下一步的主要工作是进行例外数据的分析。

参考文献

- 1 Jagadish H V, et al. Mining Deviants in a Time Series Database. In: Proc. of the 25th VLDB Conf. Edinburgh, Scotland, 1999. 102~113
- 2 Knorr E, Ng R T. Algorithms for Mining Distance Based Outliers in Large Databases. In: Proc. of the 24th VLDB Conf. New York: USA, 1998. 392~403
- 3 Knorr E M, Ng R T. Finding Intentional Knowledge of Distance-Based Outliers. In: Proc. of the 25th VLDB Conf. Edinburgh: Scotland, 1999. 211~222
- 4 Sheikholeslami G, et al. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In: Proc. of the 24th VLDB Conf. New York: USA, 1998. 428~439
- 5 胡昌华, 张军波, 等. 基于 MATLAB 的系统分析与设计——小波分析. 西安: 西安电子科技大学出版社, 2000
- 6 徐佩霞, 孙功亮. 小波分析与应用实例. 合肥: 中国科学技术大学出版社, 1996
- 7 Arning A, et al. A Linear Method for Deviation Detection in Large Databases. KDD, 1995
- 8 史东辉, 蔡庆生, 等. 基于规则的分类数据离群挖掘方法研究. 计算机研究与发展, 2000, 37(9): 1094~1100

(上接第139页)

通过自适应搜索自动确定聚类个数,而不再是直接考察统计误差, $\sum_{i=1}^K \sum_{j=1}^n d(X_i, Q_j)$ 从而可以实现对数据的智能分类。

2) 数据对象的处理过程在时间上无先后顺序之分,这一特性使得算法在处理分布式数据时效率很高。

3) 算法实现简单,时间复杂度为 $o(n^2)$,使之能有效地完成大规模数据的聚类分析工作。

CLEC 方法的提出为从其它角度重新审视传统的聚类分析方法提供了新的思路。

参考文献

- 1 Bobrowski L, Bezdek J C. c-Means Clustering with the l_1 and l_∞ Norms. IEEE Trans on Systems, Man and Cybernetics, 1991, 21(3): 545~554
- 2 Milligan G W, Cooper M C. An Examination of Procedure for Detecting the Number of Clusters in a Data Set. Psychometrika, 1985, 50: 159~179
- 3 Leouski A V, Croft W B. An Evaluation of Techniques for Clustering Search Results: [Technical Report IR-76]. Department of Computer Science, University of Massachusetts. 1996
- 4 Zhu Ming, Wang Jun Pu. A New Approach of Clustering. Pattern Recognition and Artificial Intelligence. 262~265