第25卷第3期
2005年3月

计算机应用
Computer Applications

Vol. 25 No. 3
Mar. 2005

文章编号:1001-9081(2005)03-0652-02

基于离散小波变换的时间序列数据挖掘

余璟明,何希琼,程冬爱
(中国科学院 成都计算机应用研究所,四川 成都 610041)
(yujingming@sohu.com)

摘 要:提出了一种利用离散小波变换进行时间序列分析预测的新方法。该方法的特点主要是在小波系数的选取依据上与以往方法不同,以往方法大多是选取前k个位置的系数或者是选取数值最大的k个位置的系数,其依据是能量保持;本文方法的选取依据是各系数在训练集数据上的分类能力大小,即通过对已知类别的训练集的学习过程,找出使得类内距离最小、类间距离最大的若干系数作为特征系数。对于未知类别的时间序列,根据特征系数计算出该序列属于各个类别的隶属度,隶属度最高的类别即为预测结果。实验结果表明,本方法用于时间序列分析预测,显示出了较高的效率和准确性。

关键词:时间序列;离散小波变换;特征提取;趋势预测
中图分类号:TP311.13 **文献标识码**:A

Time series data mining using discrete wavelet transform

YU Jing-ming, HE Xi-qiong, CHENG Dong-ai
(*Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610041, *China*)

Abstract: A new method of time series data mining using discrete wavelet transform was proposed. The main property of this method lied in its feature extraction strategy, which was different from other methods before. Instead of using only the first k coefficients or the largest k coefficient, which emphasized on energy preservation, this new method decided extracted coefficients according to their classification ability on real time sequences in the training set. Generally speaking, this method tried to select coefficients from all wavelet coefficients to form a feature coefficient set, which best enlarged the distance between different classes and reduced the distance within same class. With this feature coefficient set, we can make prediction on test sequences set by calculating pertaining degree of each sequence to each class. The prediction result is the pertaining relation with largest pertaining degree. For real time series data used in our research, the efficiency and accuracy of this method is satisfying.

Key words: time series; discrete wavelet transform; feature extraction; trend prediction

时间序列数据是一组有序的、随时间变化的数值序列,其中相邻数值间的时间间隔一般是相等的。世界上的许多事物、现象的发展变化都离不开时间,所以时间序列数据库的分布相当广泛。数据挖掘的一个重要分支就是挖掘基于时间序列的数据。大部分时间序列数据有着很多的共性,像周期性、季节性、随机性等,因而时间序列数据挖掘领域的任何进展都可能带来广泛的社会效益。例如在证券行业,通过分析股票市场历史走势的变化特点,我们可以对未来走势进行预测;商品销售中,通过对销售数据的分析,可以预测未来的市场需求状况,从而指导生产计划的制定;电力部门可以通过对用电量分析,指导电力分配;在医学领域,医生通过对药物疗效的数据分析,掌握药物的特性等。

时间序列挖掘发展过程中曾经面临的一个困难就是它的高维性。我们分析时间序列时,不可能只关注某个点,往往需要同时关注连续的若干个点的特征,这些点构成了一个窗口,或者称作向量,随着窗口的扩大,向量的维数不断增长,以致产生所谓的"维灾",同时也使依据窗口序列距离进行分类变得毫无意义[1]。然而大部分连续的时间序列中的点不是相互独立的,而是彼此相关的,因此一定存在信息冗余。于是人们想到可以对时间序列进行压缩,通过特征提取只保留重要的信息,去除噪声和相关性[2]。通过特征提取,我们不仅可以大大加快挖掘算法的速度,而且能够产生比直接应用原始数据更好的结果。目前较流行的特征提取方法就是离散小波变换[3],利用它可以提取时间序列的频谱特性,它具有多分辨分析的特点,在时频两域都具有表征信号局部特征的能力。通过保留小波变换后的前k个位置的系数,我们就可以得到原时间序列的一个粗略逼近,这是因为前面的系数对应着时间序列的低频特征。然而这对于频率较高的时间序列并不是个好办法,它会丢掉所有的高频信息。为了尽量保留时间序列的能量信息,我们可以保留小波系数中最大的k个系数作为特征系数,这种方法能够最大限度地保留原时间序列的能量信息[4]。然而这种最优的方法有个极大的缺点就是效率太低,因为各个时间序列的最大系数位置各不相同,使得必须保留每个时间序列的小波系数位置信息,增大了算法的时间复杂度和空间复杂度。一种改进的方法是对上面两个方法进行折中,即通过对已知时间序列的学习,找出使得对多数时间序列能量保持最优的系数位置作为统一的特征系数,这种方法提高了算法的效率,然而却丢掉了许多可能对下一步数据挖掘很有价值的分类信息。

这些方法特征提取过程的共同依据是能量保持。而本文的特征提取方法依据在于其对后续数据挖掘过程的价值。选取依据是各系数在训练集数据上的分类能力大小,即通过对已知类别的训练集的学习过程,选出使得类内距离最小、类间距离最大的若干系数作为特征系数。对于未知类别的时间序

收稿日期:2004-08-16;修订日期:2004-10-13

作者简介:余璟明(1978-),男,江苏无锡人,硕士研究生,主要研究方向:数据挖掘; 何希琼(1951-),女,四川成都人,研究员,主要研究方向:数据库、数据仓库; 程冬爱(1980-),女,湖北省武汉人,硕士研究生,主要研究方向:数据挖掘.

列,根据特征系数计算出该序列属于各个类别的隶属度,隶属度最高的类别即为预测结果。实验结果表明,本方法用于时间序列分析预测,显示出了较高的效率和准确性。

1　方法流程

本文的数据挖掘方法主要包含 4 个步骤。第一步是数据采集,对时间序列历史数据进行挑选和整理,按照时间序列的后续走势情况对数据序列进行分类,建立训练集与测试集;第二步是数据转换,利用离散小波变换对原时间序列进行转换,得到转换后所有位置的小波系数值;第三步是特征提取,根据训练集的时间序列对各个小波系数进行评价,计算出各系数对于缩小类内距离和增大类间距离的贡献度,选取贡献度最高的若干系数作为该类别的特征系数。第四步是预测与评价,对于一系列未给出类别信息的时间序列测试集,利用特征系数计算出其属于各类别的隶属度,隶属度最高的类别即为预测结果,统计预测结果的准确率,提取有意义的模式作为知识。

2　数据采集

表 1　迭代二分法

第一次迭代	第二次迭代	第三次迭代	…
$-10.00 \leqslant X \leqslant -0.20$	$-10.00 \leqslant X \leqslant -1.37$	$-10.00 \leqslant X \leqslant -2.33$	
		$-2.33 < X \leqslant -1.37$	
	$-1.37 < X \leqslant -0.20$	$-1.37 < X \leqslant -0.74$	
		$-0.74 < X \leqslant -0.20$	
$-0.20 < X \leqslant 10.00$	$-0.20 < X \leqslant 1.20$	$-0.20 < X \leqslant 0.48$	…
		$0.48 < X \leqslant 1.20$	
	$1.20 < X \leqslant 10.00$	$1.20 < X \leqslant 2.56$	
		$2.56 < X \leqslant 10.00$	

注:分类方式 X 为后一交易日收盘价对于前一日的涨跌幅(%)

数据采集过程的主要目标是建立已知类别的训练集和测试集,其中的关键是对时间序列确定窗口宽度以及定义时间序列类别。本次试验选取了 1998 年到 2003 年上海股票交际市场 50 只股票的收盘价作为试验对象,其中 1998 ~2002 年的数据组成训练集,2003 年的数据作为测试集。根据股票数据的特点,这里将时间序列窗口宽度定为 32。对时间序列进行分类的依据主要是其后一交易日的收盘价。为便于后续的特征提取,这里的分类方式采取了迭代二分法(如表 1 所示),即每次迭代将后一交易日的走势情况分成两类,逐步求精。表 1 中的第一次迭代选取 -0.20 作为分类点的原因是大于它与小于它的时间序列数目相等,一般称之为中位点,后续过程同样如此。至于取正负 10 为边界是因为股票政策限制股票日涨跌幅不能超出此范围。这种迭代的主要目的是使得后续过程中的特征提取过程每次只需对两个类进行操作,也使得预测过程是逐步求精的。通过数据采集,便可得到具有已知类别的两类训练集,接下来就是要对两类训练集进行数据转换和特征提取。

3　数据转换

数据转换过程利用 Haar 小波变换将宽度为 32 的训练集/测试集的时间子序列变换到频谱空间,得到 32 位小波系数。小波变换相对于之前的傅立叶变换的优点是其在时频两域都具有表征信号局部特征的能力。而 Haar 小波变换又是小波变换家族中相当经典的一种,它的最大特点其完美的正交特性,而且形式简单、易于实现。Haar 小波的父波形式为:

$$\phi(t) = \begin{cases} 1, & t \in [0, 1] \\ 0, & t \notin [0, 1] \end{cases}$$

它主要提取时间序列的平均值信息,也称低频信息;母波形式为:

$$\psi(t) = \begin{cases} 1, & t \in [0, 0.5] \\ -1, & t \in [0.5, 1] \\ 0, & t \notin [0, 1] \end{cases}$$

它提取的是时间序列的变化信息,也称高频信息。通过对父波与母波形式进行偏移和缩放可以得到两个函数族,分别为 $\phi_i^j(t) = \phi(2^j t - i)$ 和 $\psi_i^j(t) = \psi(2^j t - i)$,其中 $j = 0, 1, \cdots$; $i = 0, 1, \cdots, 2^j - 1$。这两个函数族构成了变换基函数向量空间:$V^j = sp\{\phi_i^j\}_{i=0,1,\cdots,2^j-1}$, $W^j = sp\{\psi_i^j\}_{i=0,1,\cdots,2^j-1}$。它具有以下性质:$V^j \subseteq V^{j+1}$, $W^j \subseteq W^{j+1}$, $V^{j+1} = V^j \oplus W^j$,由此公式可以递推求出任意宽度时间序列变换基函数向量空间,例如 $V^3 = V^0 \oplus W^0 \oplus W^1 \oplus W^2$。对时间序列 $f_k = (f(0), f(1), \cdots, f(N-1))$, $N = 2^k$ 作转换就是将其转换成带有系数的基函数之和的形式。例如 $f_{k=3} = < f, \phi_0^0 > \phi_0^0 + < f, \psi_0^0 > \psi_0^0 + < f, \psi_0^1 > \psi_0^1 + < f, \psi_1^1 > \psi_1^1 + < f, \psi_0^2 > \psi_0^2 + < f, \psi_1^2 > \psi_1^2 + < f, \psi_2^2 > \psi_2^2 + < f, \psi_3^2 > \psi_3^2$(其中的尖括号表示内积)。变换后得到的系数个数与时间序列宽度相同,然而它与原序列的最大差别是体现了不同频率、不同位置的数值变化信息,因此比原数据更有意义。然而这些信息未必都对后续的数据挖掘过程有用,这就需要对这些系数进行评价,找出有价值的系数作为特征系数,即下一步特征提取所要做的工作。

4　特征提取

该步骤的主要工作是对训练集成员的小波系数进行评价,找出最有助于分类的系数位置及其数值作为该类别的特征系数。因为每次迭代只有两个类别,所以只需在两个类别训练集上进行。这里将两类别分别标记为 A 和 B,将 A 类成员 32 个系数记为 $a_1, a_2, \cdots, a_{32}$,其均值为 $\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_{32}$,标准差为 $v(a_1), v(a_2), \cdots, v(a_{32})$;同理 B 类成员 32 个系数记为 $b_1, b_2, \cdots, b_{32}$,均值为 $\bar{b}_1, \bar{b}_2, \cdots, \bar{b}_{32}$,标准差为 $v(b_1), v(b_2), \cdots, v(b_{32})$;特征系数集合记为 $S = (p_1, p_2, \cdots, p_k)$,其中的 p_i 代表系数位置。初始时,S 集合为空,顺序考查各个系数位置,算出待考查系数位置加入以后训练集中成员的类间距离之和与类内距离之和的比值,作为加入后的分类能力值,将之与加入前的能力值进行比较,若增强则加入,若减弱则考查下一位置,直到考查完毕。

这是一种最优方法,可以得到最优解,然而它的缺点是效率太低,主要是由于计算类内距离与类间距离的时间复杂度成组合式增长。这里使用的是一种次优方法,即分别计算每个位置的分类能力,第 i 个系数位置的分类能力计算公式为 $\dfrac{|\bar{a}_i - \bar{b}_i|}{v(a_i) + v(b_i)}$,然后对各个位置的分类能力进行排序,选择能力最高的若干系数位置加入 S。选择方式有两种,一种是规定选择的个数,另一种是给分类能力设定一个域值 ε,大于 ε 的系数位置加入 S。实验发现设定域值方法更为有效。通过本步骤,最后得到的特征系数以有序链表形式存储,即 $S = ((p_1, \bar{a}_{p_1}, v(a_{p_1}), \bar{b}_{p_1}, v(b_{p_1})), (p_2, \bar{a}_{p_2}, v(a_{p_2}), \bar{b}_{p_2}, v(b_{p_2})), \cdots, (p_k, \bar{a}_{p_k}, v(a_{p_k}), \bar{b}_{p_k}, v(b_{p_k})))$,注意 k 值不是常数,而是由小波系数的实际分类能力决定。当没有满足域值要求的系数位置($k = 0$)存在时,迭代过程结束。通过 n 次迭代,将得到 $2^0 + 2^1 + \cdots + 2^{n-1} = 2^n$ 个这样的特征系数。

(下转第 663 页)

由于本系统选用的词典等方面的问题,导致试验结果中某些类别的分类准确率较低,但这并不影响对本文提出的互信息比值法作用的验证。

实验数据如表 1、表 2 所示。

表 2　$N=3000$ 时的测试结果

序号	类别	测试样本数		封闭测试集				开放测试集			
		封闭测试集	开放测试集	改进的互信息		互信息比值法		改进的互信息		互信息比值法	
				正确样本	准确率(%)	正确样本	准确率(%)	正确样本	准确率(%)	正确样本	准确率(%)
1	政治	1000	264	85	8.50	135	13.50	13	4.92	42	15.91
2	经济	1109	274	521	46.98	640	57.71	120	43.80	99	36.13
3	军事	1000	255	741	74.10	755	75.50	176	69.02	173	67.84
4	体育	1000	255	366	36.60	562	56.20	118	46.27	155	60.78
5	旅游	1001	255	156	15.58	274	27.37	28	10.98	25	9.80
6	教育	999	255	919	91.99	945	94.59	245	96.08	248	97.25
7	健康	241	255	76	31.54	141	58.51	22	8.63	46	18.04
8	通信	1000	257	892	89.20	944	94.40	207	80.54	243	94.55
总计		7350	2070	3856	52.46	4396	59.81	929	44.88	1031	49.81

从以上结果我们不难看出,使用互信息比值法进行特征选择有效提高了文本分类的准确率,在特征子集维数较低时的效果尤其明显。

本文还把互信息比值公式中的次大值换成第三大、第四大的值依次进行实验,但分类结果基本没有变化。

4　结语

前面已经提到,采用本论文提出的互信息比值法很有可能把生僻词纳入到特征子集中来,应当采取适当措施首先剔除生僻词,然后再使用互信息比值法。

词典是文本分类系统中对文本进行分词、进而构造其特征向量的关键所在,信息充分的词典能够极大地影响文本分类的性能。词典的构造已经成为影响文本分类领域的一个瓶颈问题,由于缺乏有效的自动或半自动构造方法,目前大多数工作是由大量的人工完成的,而且目前尚没有形成一个标准的字典。因此今后对于词典的构造是亟待解决的一大问题。

参考文献:

[1] 史忠植. 知识发现[M]. 北京:清华大学出版社, 2002.

[2] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004,18(1).

[3] 王玉玲,王娟. 文本分类中的特征选取算法[J]. 孝感学院学报, 2003,23(6).

(上接第 653 页)

5　预测与评价

该阶段的工作是对测试集数据进行预测,并对预测结果进行评价。预测过程同样是个迭代求精的过程,需要用到上一步骤得出的一系列特征系数,每一次迭代都需计算出预测对象属于两个类的隶属度,隶属度大的类就是预测结果。将测试集小波系数序列记为 c_1, c_2, …, c_{32},则对 A 类的隶属度计算公式为 $\dfrac{\sum_{i=1}^{k}(c_{p_i}-\bar{a}_{p_1})^2}{\sum_{i=1}^{k}v^2(a_{p_i})}$ 的倒数,B 类公式与此类同。通过在测试集上实验,发现此方法显示出了较高的分类能力。我们对不同的迭代分支分别计算其准确率,得到的准确率曲线如图 1 所示。横轴代表迭代的层次,纵轴代表准确度,由于每次迭代(第一次除外)有多个分支,我们将准确度记为每次迭代多个分支的均值。我们发现随着迭代层次增加,准确率呈下降趋势,这个结果也很容易理解,即随着预测精度的增加,特征系数的分类能力不断减弱。这里的准确率的测量是针对单次独立测量,实际的准确率应该是每次迭代准确率的乘积,所以实际准确率比图示更低,特别是当迭代次数较大时,因此预测仅在前几次迭代中有实际意义。这种迭代方法还有一个很大的缺点:虽然最初几次迭代有较高的正确率,但对于具体的测试集成员来说,一旦某次迭代判断错误,则后续迭代将没有意义。为了尽量不作出错误决策,这里计算正确率还有一个作用,就是用一个域值对迭代树进行剪枝,去掉准确率太低的分支,使得在实际预测时保持一定的准确率。另外在实际预测中,遍历所有准确率较大的分支,给出多个结果和相应的准确率,以供参考。通过评价我们可以提取准确率较高的预测模式作为知识,这才是数据挖掘的终极目标。

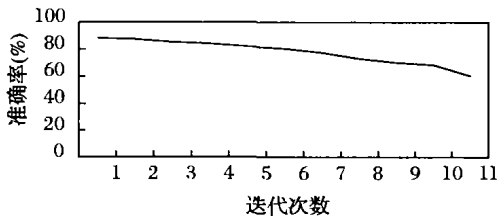

图 1　准确率变化曲线

参考文献:

[1] AGGARWAL CC, HINNEBURG A, KEIM DA. On the Surprising Behavior of Distance Metrics in High Dimensional Space[A]. Proceedings of the 8th International Conference on Database Theory [C]. London, UK: Springer-Verlag, 2001.420 - 434.

[2] GEURTS P. Pattern extraction for time series classification [A]. Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery[C]. Freiburg, Germany: Springer-Verlag, 2001.115 - 127.

[3] VLACHOS M, LIN J, KEOGH E, *et al*. A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series[A]. the 3 SIAM International Conference on Data Mining[C]. San Francisco, CA, 2003.

[4] WU Y-L, AGRAWAL D, ABBADI AE. A comparison of DFT and DWT based similarity search in time-series databases[A]. Proceedings of the 9th International Conference on Information and Knowledge Management[C]. McLean, VA: ACM Press, 2000. 488 - 495.

[5] POPIVANOV I, MILLER RJ. Similarity Search Over Time - series Data Using Wavelets[A]. Proceedings of the 18th International Conference on Data Engineering[C]. Washington, DC: IEEE Computer Society, 2002.212 - 221.