International Journal of Innovative Computing, Information and Control
Volume **3**, Number **2**, April **2007**

ICIC International ©2007 ISSN 1349-4198
pp. **471–480**

UNUSUAL SUB-SEQUENCE IDENTIFICATIONS IN TIME SERIES WITH PERIODICITY

Rawshan Basha

Computer Department
University of Sharjah
P.O. Box 27272, Sharjah, United Arab Emirates
rawshan@sharjah.ac.ae

Jamal Ameen

Faculty of Advanced Technology
University of Glamorgan
Pontypridd, CF37 1DL, United Kingdom
jrmameen@glam.ac.uk

Received December 2005; revised June 2006

Abstract. *Fast and intelligent data mining has recently become an integral part of data analysis and a pre-requisite for modeling. This is largely due to the introduction of more sophisticated data collection tools and the possibility of observing large datasets at increased higher frequencies. This paper aims to investigate the current methodologies used for the detection of time series discord sub-sequences and especially those with periodicity. A strategy will be suggested to use classical data mining techniques and statistical decision making to take advantage of the special features of the time series to make the detection more efficient and more objective. An entropy-based measure will also be introduced as an alternative to the Euclidean distance measure for identifying discord sub-sequences.*
Keywords: Time series, Sub-sequence, Seasonality, Autocorrelation, Discord, Data mining

1. **Introduction.** Outlier detection in classical time series analysis has been studied by many authors [1,2,12]. However, investigating blocks of adjacent time series observations that are significantly distanced from the rest of the observed time series (unusual time series sub-sequences) is new. These may occur as a result of internal malfunctioning or the impact of some external sources on the process that has been observed for a short period of time. Their detection is therefore possible only when the concerned processes are observed at high frequency and using modern technological tools. For example, the use of a Laser Doppler in measuring the direction and speed of blood in blood flow monitoring with the ability of recording a minimum of 25 observations per second. The reliable detection of such phenomenon and its distinction from the impact that outliers may make, are of great help in studying their occurrence causes.

Current approaches in this area are mainly based on a combination of computer capabilities and human subjective assessment. In general, a sub-sequence length is subjectively stated for which, all pairs of sub-sequence distances (mainly Euclidean) are measured. The subsequence that is found to be 'far' from the rest of the sub-sequences is considered