A New Symbolization and Distance Measure Based Anomaly Mining Approach for Hydrological Time Series

Pengcheng Zhang, College of Computer and Information, Hohai University, Nanjing, China
Yan Xiao, College of Computer and Information, Hohai University, Nanjing, China
Yuelong Zhu, College of Computer and Information, Hohai University, Nanjing, China
Jun Feng, College of Computer and Information, Hohai University, Nanjing, China
Dingsheng Wan, College of Computer and Information, Hohai University, Nanjing, China
Wenrui Li, College of Computer and Information, Hohai University, Nanjing, China & School of Mathematics and Information Technology, Nanjing Xiaozhuang University, Nanjing, China
Hareton Leung, Department of Computing, Hong Kong Polytechnic University, Hong Kong, China

ABSTRACT

Most of the time series data mining tasks attempt to discover data patterns that appear frequently. Abnormal data is often ignored as noise. There are some data mining techniques based on time series to extract anomaly. However, most of these techniques cannot suit big unstable data existing in various fields. Their key problems are high fitting error after dimension reduction and low accuracy of mining results. This paper studies an approach of mining time series abnormal patterns in the hydrological field. The authors propose a new idea to solve the problem of hydrological anomaly mining based on time series. They propose Feature Points Symbolic Aggregate Approximation (FP_SAX) to improve the selection of feature points, and then measures the distance of strings by Symbol Distance based Dynamic Time Warping (SD_DTW). Finally, the distances generated are sorted. A set of dedicated experiments are performed to validate the authors’ approach. The results show that their approach has lower fitting error and higher accuracy compared to other approaches.

KEYWORDS
Data Mining, Distance Measure, Hydrological Time Series, Pattern Representation

1. INTRODUCTION

With the advance development of technology, the data which need to be dealt with is becoming various and complicated. Furthermore, the scale of the data is huge, the form of the data is more diverse and the speed of processing data is lacking. So how to gain valuable information and meaningful knowledge quickly from the numerous and complex big data has become a key challenge.

Data Mining (Lin & Chen, 2011) extracts potentially useful information which people is interested in and which is unknown in advance. There is a variety of knowledge representation, such as concepts, rules, regularities and patterns. Time series (Box, Jenkins et al. 2013), which reflect the characteristic of the attribute value related with time, have large data scale, high dimension and frequent update.

DOI: 10.4018/IJWSR.2016070102

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
As an important branch in the field of data mining research, time series data mining is the process of pattern discovery and knowledge extraction from the time sequence. The basic task includes similarity searching, periodical pattern mining, analysis and prediction, clustering and classification, visualization and abnormal data mining of time series. For a long time, the aim is finding the data pattern which appears frequently. People want to summarize some rules from the data pattern. Currently, almost all abnormal data are considered as noisy data and then ignored. In some cases, compared with the normal data, the frequency of abnormal data is not high, but it may hide some important information. And finding these abnormal data and the corresponding hidden information may provide valuable information and more enlightening knowledge.

Abnormal data are considered intuitively to be those which are associated with the data model or the data objects but do not conform to the general distribution. Nowadays there is not a definition generally accepted for abnormal data. It changes with the specific application. The research for abnormal time sequence mining is derived relatively late, but recently it is attracting more interest. For instance, this technology can monitor the crime hidden in the electronic commerce. And on the other hand, it can be used to detect possible intrusions by Hackers in the daily management of the Internet. In these applications, abnormal data often represent more useful meaning than other normal data. Although the probability of the abnormal data is very small, their importance cannot be ignored. Therefore, capturing these small probability events in the abnormal cases is important for some applications.

Hydrological data are discrete records for hydrological process, such as flow, rainfall, water level and so on (Ping, 2003). They are huge, noisy, unstable and have poor correlation. Nowadays, hydrological modernization is popular, which contains information collection, extraction, analysis and so on. During the hydrological process, it is indispensable to acquire valuable information and meaningful knowledge quickly from large amount of data. The rapid development of data mining provides a new approach for water resource management, hydrology and hydroinformatics research (Kozerski, 2010).

Hydrological time series data mining (Kozerski, 2010) can be used to extract the unknown process which contains important information from hydrological data, which is valuable for hydrological forecasting and hydrological data analysis. Abnormal hydrological time series are the data objects which are obviously inconsistent with the universal rule of the hydrological phenomenon. In the field of hydrological data mining, there are three branches: similarity search, sequence pattern mining and cycle analysis. The research of abnormal hydrological time series data mining is still at the starting stage.

There are many approaches for abnormal time series data mining. Most of them have clear problems. For example, the approach based on immunology (Chaovalit, Gangopadhyay et al. 2011) cannot apply to diverse data. Computational efficiency of Support Vector Machine (SVM) (Verdejo, García et al. 2011) is high, but its theory and modeling process are very complex which can only be adopted by experts. The accuracy of TSA-tree (Zhang, Meratnia et al. 2010) is low. To solve these problems, this paper puts forward a new approach which is based on Extended Symbolic Aggregate Approximation (ESAX) (Lkhagva, Suzuki et al. 2006) and Dynamic Time Warping (DTW) (Müller, 2007).

In summary, the contributions of this paper are as follows:

- Approaches of selecting feature points (extreme points, minimum or maximum) are added into ESAX in pattern representation to reduce dimension, which is called FP_SAX. FP_SAX looks for new feature points which are more representative to replace the maximum and minimum
Software-Defined Networking in Access Networks: Opportunities, Challenges, and Choices
www.igi-global.com/article/software-defined-networking-in-access-networks/125455?camid=4v1a

Exploring the Influence of Contexts for Mobile Recommendation
www.igi-global.com/article/exploring-the-influence-of-contexts-for-mobile-recommendation/188456?camid=4v1a