Journal of Environmental Management 196 (2017) 365–375

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

A novel water quality data analysis framework based on time-series data mining

Weihui Deng [a, b], Guoyin Wang [a, *]

[a] Chongqing Key Laboratory of Big Data and Intelligent Computing, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form
23 February 2017
Accepted 8 March 2017
Available online 18 March 2017

Keywords:
Time-series data mining
Cloud model
Water quality analysis
Similarity measure
Anomaly detection
Pattern discovery

ABSTRACT

The rapid development of time-series data mining provides an emerging method for water resource management research. In this paper, based on the time-series data mining methodology, we propose a novel and general analysis framework for water quality time-series data. It consists of two parts: implementation components and common tasks of time-series data mining in water quality data. In the first part, we propose to granulate the time series into several two-dimensional normal clouds and calculate the similarities in the granulated level. On the basis of the similarity matrix, the similarity search, anomaly detection, and pattern discovery tasks in the water quality time-series instance dataset can be easily implemented in the second part. We present a case study of this analysis framework on weekly Dissolve Oxygen time-series data collected from five monitoring stations on the upper reaches of Yangtze River, China. It discovered the relationship of water quality in the mainstream and tributary as well as the main changing patterns of DO. The experimental results show that the proposed analysis framework is a feasible and efficient method to mine the hidden and valuable knowledge from water quality historical time-series data.

1. Introduction

Water resource is an indispensable material condition to human survival and society development as well as an essential part of the ecological environment. Apart from drinking and personal hygiene, it has a direct influence on ecological balance, agricultural production, industrial and manufacturing processes, etc. Over the past several decades, water quality problem has gradually become an important issue to the environment management department. Large amounts of water quality historical data have been collected and available by setting up a number of water quality monitoring stations. How to make the best of potentially useful information hiding behind these historical data emerges as a new scientific issue to the academic community. Theory-driven and data-driven are two main methods of analyzing water quality data. The former based on one's domain knowledge does some tasks of statistical analysis or establishes water quality models purposefully. The latter is to mining some interesting events or patterns from the water quality historical data without any priori knowledge. Usually, water quality data are recorded in time order, which takes the form of time series. That is, there exists strong stochastic dependencies among the observations. However, the common data-driven methods consider each observation rather than the whole time series as one independent sample, which cannot make use of these dependencies.

Time-series data mining (TSDM), as an important research field in computer science community, is to study discovering the meaningful knowledge and patterns hiding in time-series data. In recent years, new emerging applications of TSDM in water resources have contributed to environmental sciences. Solomatine (2002) gave an overview of successful applications of several data mining techniques in the problems of water resources control, including classification, clustering, and prediction. Hill and Minsker (2010) developed a real-time anomaly detection method to identify the abnormal patterns for environmental data streams. In Oueslati et al. (2015), time-series clustering and prediction techniques were

* Corresponding author. Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, No.266 Fangzheng Avenue, Shuitu Hi-tech Industrial Park, Shuitu Town, Beibei District, Chongqing, 400714, China.
E-mail addresses: dengweihui@cigit.ac.cn (W. Deng), wanggy@ieee.org (G. Wang).

http://dx.doi.org/10.1016/j.jenvman.2017.03.024
0301-4797/© 2017 Elsevier Ltd. All rights reserved.

used to classify the flow regimes of Mediterranean streams. Kim (2016) used the time series analysis techniques to model the soil moisture dynamics on a steep mountainous hillside. Asadi et al. (2016) applied a data-driven approach into the water treatment aeration process. They used the TSDM methods to select features (water quality variables), construct models, and predict the effluents. In Ouyang et al. (2010), authors analyzed hydrology daily discharge historical time-series data using similarity search and pattern discovery techniques. Other achievements in hydrological time-series data mining include (Toth, 2013; Bloomfield et al., 2015; Bichler et al., 2014).

For the water quality time-series data, the most common application of TSDM is prediction (Bowden et al., 2012; Wu et al., 2014; Chang et al., 2015). For example, Deng et al. (2015) proposed a novel water quality prediction model based on fuzzy time series and cloud model, and applied it into the dissolve oxygen (DO), chemical oxygen demand in manganese (COD_{Mn}), water temperature and electric conductivity time series prediction. Partalas et al. (2008) studied the greedy ensemble selection family of algorithms for ensembles of regression models, and applied it into water quality time series prediction. Burchard-Levine et al. (2014) examined the ability of Artificial Neural Network optimized by Genetic Algorithm (GA-ANN) for ammonia-nitrogen (NH_3-N), COD_{Mn} and total organic carbon (TOC) prediction. Han et al. (2016) presented a groundwater level (GWL) prediction model based on self-organizing map, stepwise cluster inference, and AR model. Besides prediction, other TSDM techniques also have been successfully applied in mining the water quality data (Cho, 2016; Camejo et al., 2013; Aubert et al., 2013), e.g., Aubert et al. (2013) clustered the water quality time-series data to extract the flood event patterns.

However, the aforementioned techniques are “tailed-made” for the special applications, and there is no general analysis framework and procedure based on time-series data mining in water quality data. Moreover, most existing studies cannot address the uncertainty inherent in certain water quality data directly and efficiently, such as incompleteness, fuzziness, and randomness. For example, the incomplete dataset of observations need to be preprocessed like data filling before performing the data analysis tasks, thus the analysis results may not reflect the objective fact because they are affected by the pretreatment methods; The randomness of observations may overwhelm the regularities or the patterns of water quality data. But with the increasing of utilizing the sensors, uncertainty in water quality time-series data is becoming more common.

The main aim of this paper is to develop a general analysis framework based on time-series data mining techniques to discover the latent patterns or information from the historical water quality data. Since almost every task of TSDM relies on the similarity measure between the candidate time-series, our analysis framework consists of two parts. In the first part, we propose using two-dimensional normal cloud representation (2D-NCR) to granulate and represent the observed water quality time series, and calculating their similarities in the granulated level. Having these implementation components, we discuss several common tasks of TSDM for water quality historical data in the second part. The proposed analysis framework hopes to provide a basic process and idea for analyzing the water quality historical data from the view of TSDM.

Dissolved oxygen (DO) is one of the important water quality variables for the survival of aquatic life. It is frequently used to assess the water quality of rivers. The river with low concentrations of DO is recognized as an unbalanced ecosystem with fish mortality, odours and other aesthetic nuisances. The overly low concentrations of DO will cause many problems like the reproductive problem or deformities. In general, DO is an indicator of the organic pollution level in an aquatic system. In the experiments of this paper, the novel analysis framework was applied to the weekly dissolved oxygen (DO) time series collected at five monitoring stations on the upper reaches of Yangtze River in China. Each DO time series includes 520 samples recorded weekly over a period of 10 years from 2005 to 2014. The experimental results demonstrate that the proposed analysis framework is a feasible and efficient method to mine the hidden and valuable knowledge from water quality historical time-series data.

The rest of this paper is organized as follows. Section 2 briefly reviews some TSDM techniques as well as some basic concepts of cloud model. In Section 3, the novel analysis framework based on TSDM for water quality time-series data is proposed. A case study to illustrate the analysis framework and procedure is presented in Section 4. And the last section summarizes the conclusions.

2. Preliminaries

2.1. Time-series data mining

The major tasks of TSDM include 7 aspects (Esling and Agon, 2012): query by content (or similarity search), anomaly detection, motif discovery, prediction, clustering, classification, and segmentation. In order to perform these tasks well, three major issues are involved: time series representation, similarity measurement, and indexing. In the recent decade, many achievements of TSDM have been proposed such as classification methods (Bagnall et al., 2015; Zhao and Itti, 2015), clustering methods (Gacek and Pedrycz, 2015; Ferreira and Zhao, 2016), and similarity measures (Wang et al., 2013). Also they have been successfully applied into a wide variety of domains (Chen and Chen, 2015; Deng et al., 2016b; Zhang et al., 2006). The tasks of TSDM involved in this study are similarity research, anomaly detection, clustering, and pattern discovery. For brevity, detailed review is not covered here. Readers can refer to the literature (Esling and Agon, 2012; Gupta et al., 2014; Wang et al., 2013).

Time series representation and similarity measure are two basic issues to implement the tasks of TSDM. Formally, a time series T is an ordered sequence of n real-value variables $T=(t_1,t_2,\ldots,t_n), t_i \in \mathscr{R}$, where t_i is an observed value in the data space. Esling and Agon (Bowden et al., 2012) formulated the representation as “Given a time series $T=(t_1,t_2,\ldots,t_n)$ of length n, a representation of T is a model $\overline{T}$ of reduced dimensionality $\overline{d}(\overline{d} \ll n)$ such that $\overline{T}$ closely approximates T”. There is a number of time series representations proposed for data mining, including Discrete Fourier Transform (DFT), Piecewise Aggregate Approximation (PAA), Piecewise Linear Approximation (PLA), Singular Value Decomposition (SVD), Symbolic Aggregate approXimation (SAX), PieceWise Cloud Approximation (PWCA), etc., these of which can be divided into two categories: data adaptive and non-data adaptive (Wang et al., 2013).

Given two time series T_1 and T_2, the similarity measure refers to a similarity function $Dist$ used to calculate the distance between the two time series, denoted as $Dist(T_1,T_2)$. Several common similarity measures include Minkowski Distance (L_p norms), Dynamic Time Warping (DTW), Longest Common SubSequence (LCSS), etc. For further reading, please refer to the literature (Wang et al., 2013).

2.2. Cloud model

The cloud model, as a new bidirectional cognition model, was proposed by Li and Du (2007) based on probability theory and fuzzy sets theory. In cloud model theory, it is possible to measure the