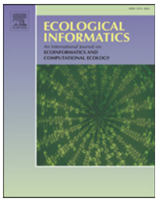
Ecological Informatics 39 (2017) 10–22

Contents lists available at ScienceDirect

Ecological Informatics

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



Mining lake time series using symbolic representation

Guangchen Ruan[a,*], Paul C. Hanson[b], Hilary A. Dugan[b], Beth Plale[a]

[a] *School of Informatics and Computing, Indiana University, 919 E. 10th Street, Bloomington, IN 47408, USA*
[b] *Center for Limnology, University of Wisconsin-Madison, 680 North Park Street, Madison, WI 53706, USA*

ARTICLE INFO

Article history:
Received 21 July 2016
Received in revised form 6 March 2017
Accepted 9 March 2017
Available online 14 March 2017

Keywords:
Lake time series
Symbolic representation
Mining

ABSTRACT

Sensor networks deployed in lakes and reservoirs, when combined with simulation models and expert knowledge from the global community, are creating deeper understanding of the ecological dynamics of lakes. However, the amount of data and the complex patterns in the data demand substantial compute resources and efficient data mining algorithms, both of which are beyond the realm of traditional limnological research. This paper uniquely adapts methods from computer science for application to data intensive ecological questions, in order to provide ecologists with approachable methodology to facilitate knowledge discovery in lake ecology. We apply a state-of-the-art time series mining technique based on symbolic representation (SAX) to high-frequency time series of phycocyanin (PHYCO) and chlorophyll (CHLORO) fluorescence, both of which are indicators of algal biomass in lakes, as well as model predictions of algal biomass (MODEL). We use data mining techniques to demonstrate that MODEL predicts PHYCO better than it predicts CHLORO. All time series have high redundancy, resulting in a relatively small subset of unique patterns. However, MODEL is much less complex than either PHYCO or CHLORO and fails to reproduce high biomass periods indicative of algal blooms. We develop a set of tools in R to enable motif discovery and anomaly detection within a single lake time series, and relationship study among multiple lake time series through distance metrics, clustering and classification. Furthermore, to improve computation times, we provision web services to launch R tools remotely on high performance computing (HPC) resources. Comprehensive experimental results on observational and simulated lake data demonstrate the effectiveness of our approach.

1. Introduction

Ecology has entered an era of “big data” in which the necessities for managing and analyzing data are inspiring scientists to develop novel approaches to use those data for answering contemporary ecological questions (Porter et al., 2012). Data from aquatic sensor networks are a good example of how observations for limnological variables, such as water temperature, dissolved oxygen, and phytoplankton biomass, have grown in volume. A few decades ago, one year of these data may have included hundreds of manually collected measurements; whereas, today, those same variables are sampled by autonomous sensors at frequencies up to one hertz, resulting in 0.5 M measurements per year per variable (Pellerin et al., 2016; Gries et al., 2016). When data from multiple depths, multiple years and multiple observatories are combined, the number of recorded measurements can reach into the billions. Data of these volumes stress the computational capacity of most desktop computers, and data analysis software commonly used by ecologists slows considerably when handling extremely large data sets. Even common data manipulations, such as searches, sorts, and sub-setting, become unacceptably slow.

Representation of time series data in ways that reduces the volume of data while retaining pattern relevant to ecological questions, would help ecologists overcome data volume as an obstacle to data analysis. In the field of computer science, symbolic representation of time series data was developed to help solve this problem. Specifically, *S*ymbolic *A*ggregate appro *X* imation, or SAX (Lin et al., 2003), was developed to transform time series data into symbols, as has been used in a variety of domains, such as acoustics, medicine, and image analysis (Kasten et al., 2007; Liu et al., 2006). Symbols not only require less space to store, but also retain essential statistical characteristics of the original time series and are amenable to machine learning algorithms designed for natural language processing (Lin et al., 2003). SAX has proven to perform as well as, or even better than, original representation of data for use in common data mining techniques, such as clustering, anomaly

[*] Corresponding author.
E-mail addresses: gruan@indiana.edu (G. Ruan), pchanson@wisc.edu (P.C. Hanson), hdugan@wisc.edu (H.A. Dugan), plale@indiana.edu (B. Plale).

http://dx.doi.org/10.1016/j.ecoinf.2017.03.001
1574-9541/© 2017 Elsevier B.V. All rights reserved.

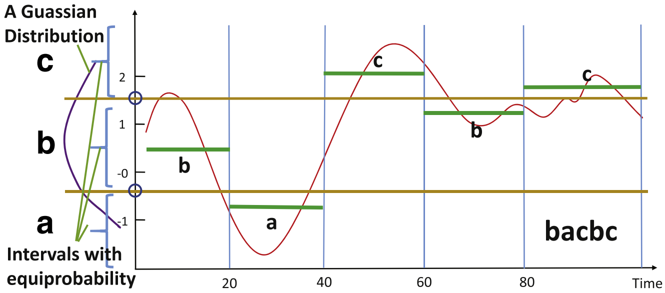


Fig. 1. Representing time series with *S*ymbolic *A*ggregate appro*X*imation (SAX). In this example, the x-axis of the time series (red line) is divided into five frames. The mean value of the data in each frame is represented by a symbol on the y-axis. The resulting letters form a word, in this case 'bacbc'.

detection, classification, and motif discover (Lin et al., 2007). Despite these advantages, we are unaware of its implementation in ecology.

Here, we introduce SAX as an approach to analyzing chlorophyll and phycocyanin (both light harvesting pigments) data from high frequency lake sensors and lake algal biomass predicted from a simulation model. We demonstrate how the SAX transformation greatly reduces data volume. In addition, we adapt pattern detection and classification algorithms from computer science as first steps in addressing the following questions: *What patterns exist in lake chlorophyll and phycocyanin data? Are the diversity and frequency of occurrence of patterns in fluorescence data matched by the patterns in predictions from lake simulations? What do the differences between observed and predicted variables tell us about lakes and the models we use to predict their time dynamics?*

With SAX, we developed a set of R tools to conduct motif discovery (Lin et al., 2002) and anomaly detection (Keogh et al., 2004; Kumar et al., 2005) within a single lake time series, and relationship study among multiple lake time series through distance metrics, clustering and classification. We conduct comprehensive experiments on observational and simulated data to demonstrate the effectiveness of our tools.

The remainder of this paper is organized as follows: Section 2 briefly describes the methodology of representing time series as symbolic characters, which serves as the foundation of mining algorithms to be developed. In Section 3, we present overall architecture and detail mining algorithms. Section 4 presents comprehensive experiments on observational and simulated lake time series data to validate the effectiveness of our approach. Related work is given in Section 5. Finally, we conclude the paper in Section 6.

2. Symbolic aggregation approximation

The challenges outlined in Section 1 all require a robust and efficient time series representation for which a rigorous distance metric can be defined. While there are many useful representations of time series data (Chan and Fu, 1999; Faloutsos et al., 1994; Huang and Yu, 1999; Keogh et al., 2001; Keogh and Pazzani, 1998; Yi and Faloutsos, 2000), *S*ymbolic *A*ggregate appro *X* imation or SAX (Lin et al., 2003), a symbolic representation of time series, solves many algorithmic problems through dimensionality and volume reduction, quantification of similarity between time series, and discovery and representation of otherwise difficult to handle data features, such as highly skewed probability densities. Essentially, SAX converts data to words and uses natural language processing algorithms to discover and classify patterns and collections of patterns in a manner akin to developing a dictionary (*i.e.*, the possible words) and a vocabulary (*i.e.*, the dictionary subset used in a time series). For example, a rare word in sensor data that might represent a surprising algal bloom can be catalogued and easily searched within the collection of words created by the SAX representation of the simulated data.

We use Fig. 1 to briefly illustrate how SAX works. The first phase of SAX is to apply *P*iecewise *A*ggregate *A*pproximation (PAA) dimensionality reduction (Keogh et al., 2001) where the original time series C of length n is reduced into a w-dimensional space as a vector $\overline{C} = \bar{c}_1, \ldots, \bar{c}_w$. The ith element of C is calculated by Eq. (1):

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j. \tag{1}$$

Simply stated, to reduce the time series from n dimensions to w dimensions, the data is divided into w equal sized "frames" (a frame shown as a segment within two consecutive blue vertical lines as in Fig. 1). The mean value of the data falling within a frame is calculated

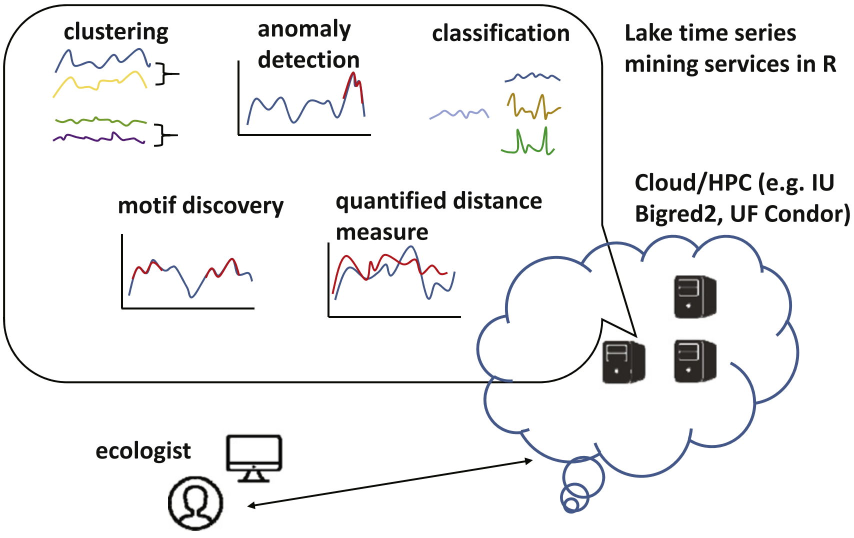

clustering
anomaly detection
classification
Lake time series mining services in R
motif discovery
quantified distance measure
Cloud/HPC (e.g. IU Bigred2, UF Condor)
ecologist

Fig. 2. Architecture sketch of lake time series mining system.