Computers, Environment and Urban Systems 43 (2014) 25–33

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Water quality analysis using a variable consistency dominance-based rough set approach

Jalal Karami [a,*], Abbas Alimohammadi [b], Tayebeh Seifouri [c,d]

[a] Department of GIS and Remote Sensing, University of Tarbiat Modares, Tehran, Iran
[b] Department of GIS Engineering, Faculty of Geodesy & Geometrics Engineering, K.N. Toosi University of Technology, Tehran, Iran
[c] Diseases Affairs, Jonoob Health Center, Tehran University of Medical Science, Tehran, Iran
[d] Faculty of Public Health and Health Promotion, Tehran University of Medical Science, Tehran, Iran

ARTICLE INFO

Article history:
Received 12 January 2011
Received in revised form 26 September 2013
Accepted 28 September 2013
Available online 2 November 2013

Keywords:
Moving average filter
Total dissolved solids
Variable consistency dominance-based rough set approach
Water quality
Latyan Dam watershed
Iran

ABSTRACT

Analysis and evaluation of water quality and its dynamics are of prime importance for water resources and environmental monitoring. Diverse methods such as multivariate statistics, time series analysis, and neural networks have been used for modeling and analysis of water quality indicators. Although these methods are useful to explore the main body of knowledge related to the water pollution problem, they are less effective for considering inherent uncertainties and vagueness in water pollution data. In this study, a variable consistency dominance-based rough set approach (VC-DRSA) was used to explore the underlying knowledge related to data for total dissolved solids (TDSs) in the Latyan Watershed, north of Tehran, Iran. Environmental parameters for the period of 2002–2007, including precipitation, river water temperature, runoff measured at 22 monitoring sites, and two products of the MODIS sensor (16-day NDVI and land surface temperature) were the explanatory variables. VC-DRSA was used in data mining analysis to explore the most effective and reliable rules for relating TDS data to the explanatory variables. Rule validation results show that the extracted rules were very effective and straightforward for examining the important relationships between the environmental parameters and TDS data. Application of the moving average filter in the TDS data led to decreased noise and a considerable reduction in the width of the boundary region between the lower and upper approximations.

1. Introduction

Water pollution is a critical environmental threat (Pimpunchat, Sweatman, Wake, Triampo, & Parshotam, 2009). Their close interaction with human activities such as agriculture, industry, transportation, and sewage discharges makes rivers as the main inland water bodies vulnerable to pollution. Because rivers play a significant ecological role, they are regarded as important indicators of the state of the environment. Population growth, industrialization, uncontrolled urbanization in developing countries, and other anthropogenic activities have led to ever-increasing river pollution (Su et al., 2010).

Total dissolved solids (TDSs), defined as any material in water that will pass through a filter with a pore size of 2 μm or smaller (Berdanier & Ziadat, 2006), is an important water pollution parameter. Most matter dissolved in fresh water consists of inorganic salts, small amounts of organic matter, and dissolved gases (Sawyer, McCarty, & Parkin, 2003). These usually include calcium, magnesium, sodium, potassium cations or carbonate, hydrogen carbonate, chloride, sulfate, and nitrate anions. Concentrations of TDS from natural sources vary from less than 30 mg/l to as much as 6000 mg/l, depending on the solubility of the minerals in different geological regions. Surface water with higher TDS levels are an important driver of the degradation of agricultural land. For example, irrigation with high TDS water is fatal to crop roots and can lead to soil salinization. Studies in Australia have shown that mortality from all categories of ischemic heart disease and acute myocardial infarction increased in a community with high levels of soluble solids, calcium, magnesium, sulfate, chloride, fluoride, alkalinity, total hardness, and pH in the water (Meyers, 1975).

Recent multivariate statistical methods such as cluster analysis, discriminant analysis, and factor and regression analysis have been used to explore the spatial and temporal variations of surface water quality and identify the sources of pollution (Huang, Wang, Lou, Zhou, & Wu, 2010; Lindenschmit, 2006; Maillard & Pinheiro Santos, 2008; Pekey, Karakas, & Bakoglu, 2004; Su et al., 2010; Zhou, Huang, Guo, Zhang, & Hao, 2007), especially

* Corresponding author. Address: Tarbiat Modares University, Jalal Ale Ahmad Highway, P.O. Box 14115-111, Tehran, Iran. Tel.: +98 21 82884698; fax: +98 21 82884180.

E-mail addresses: JL.karami@modares.ac.com (J. Karami), alimoh_abb@kntu.ac.ir (A. Alimohammadi), Taseifouri@yahoo.com (T. Seifouri).

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.09.005

of TDS (Brix, Gerdes, Curry, Kasper, & Grosell, 2010; Etemad-Shahidi, Afshar, Alikia, & Moshfeghi, 2009; Magazinovic, Nicholson, Mulcahy, & Davey, 2004). Although these methods explore the main body of knowledge about water pollution, but water pollution is a complex problem composed of issues such as decentralization, modularity, poor structure, and weak predictability (Sokolova & Fernandez-Caballero, 2009). These elements suggest that the problem of water quality inherently includes high rates of imprecision, vagueness, and uncertainty. It is necessary to employ sophisticated data analysis and mining algorithms suitable to solve problems with uncertain and complicated properties.

Examination of the capabilities of the dominance-based rough set (RS) to represent ambiguities and to explore complex relationships between the water quality and environmental parameters is the main objective of this paper. Therefore, the variable consistency dominance-based rough set approach (VC-DRSA) was used for rule extraction and classification of TDS data, with available environmental data (NDVI, LST, precipitation, runoff, RWT) used as explanatory variables.

RS theory is a powerful and flexible mathematical tool for imprecision, vagueness, and uncertainty, first proposed by Pawlak (1982). This algorithm extracts predictive and useful knowledge in the form of rules from imprecise data. The philosophy of RS theory is based on a classification where any union of elementary sets is called a crisp (or precise) set. Granularity of knowledge can be achieved approximately, rather than precisely, defining notions within the available knowledge. This type of set is referred to as the rough set (Triantaphyllou & Felici, 2006). In RS theory, it is possible to associate every set X with two crisp sets (lower and upper approximations of X), thus, each vague concept is replaced with a pair of precise concepts. The lower approximation of a concept consists of all objects that definitely belong to that concept. The upper approximation of a concept consists of all objects that may possibly belong to the concept. Hence, a boundary region can be assumed between the lower and upper approximations of a concept. The greater the boundary region, the vaguer the concept. If, for example, the boundary region of a concept is empty, that concept is considered precise.

RS theory overlaps, to some extent, other theories dealing with uncertainty and vagueness, especially the Dempster–Shafer (DS) theory of belief functions (Gorsevski, Jankowski, & Karami, 2008; Slowinski & Stefanowski, 1989) and fuzzy set theory (Dubois & Prade, 1990, 1992; Wygralak, 1989). The difference between the DS and RS theories is that RS theory uses sets of lower and upper approximations to represent knowledge in data collection, while the DS theory uses belief functions represented by lower and upper probability functions (Gorsevski et al., 2008). The approximations for a given data set derived by RS theory are based solely on the data, while the approximations derived by the DS theory involve calculations of belief values using both subjective judgments and data (Dempster, 1967). The differences between the DS and fuzzy sets are long and detailed (see (Yao, 1998)). Widely-used discipline-specific applications include remote sensing, image and signal processing (Czyzewski, 2003; Czyzewski & Krolikowski, 2001; Kostek, 1999, 2005; Peters, Han, & Ramanna, 2001; Tsumoto & Hirano, 2005; Wieczorkowska, Wroblewski, Synak, & Slezak, 2003), urban planning (Chen, Hipel, Kilgour, & Zhu, 2009; Wang, Hasbani, Wang, & Marceau, 2010), and GIScience (Duckham, Mason, Stell, & Worboys, 2001; Worboys, 1998).

In following sections of the paper, first a brief review of the VC-DRSA approach is provided followed by descriptions of the methodology and data, results, and conclusions of the study.

2. Methodology and data

2.1. Dominance-based RS approach and VC-DRSA

In a dominance-based RS approach (DRSA), as in equivalence RS, an information table is introduced by the 4-tuple $S = \langle U,Q,V,f\rangle$ where U is a finite set of objects or observations (i.e., the universe); $=c\cup D = \{q_1,q_2,\ldots,q_m\}$ is a finite set of attributes (including condition attribute set C and decision attribute set D); v_p is the domain of attribute q; and $V = \cup V_p$. Also, f: $U \times Q \rightarrow V$ is a total function such that for each $q \epsilon Q$, $x \epsilon U$ (information function) (Zhai, Khoo, & Zhong, 2009a).

In this table of information, the values of each condition and decision attributes are preference-ordered and inherently correlated. For example, an increase (or decrease) in a condition attribute value results in an upgrade (or downgrade) of the corresponding decision-class value. This is considered a key difference between the classical and dominance-based approaches. Usually, D has one member so that $D = \{d\}$ and it partitions U into a finite number of classes, such as $cl = \{cl_t, t \epsilon T\}$, where $T = \{1,\ldots,n\}$.

Classes in cl are ordered in an ascending sequence of class indices. For all r, $s \epsilon T$ and $r > s$, the objects included in cl_r are preferable to those contained in cl_s; therefore, for DRSA, the sets to be approximated are no longer single classes, but the upward and downward unions of the decision classes, respectively (Błaszczynski, Greco, & Slowinski, 2007).

The upward and downward unions of class cl_t are, respectively:

$$cl_t^{\geqslant} = \cup_{s \geqslant t} cl_s, \tag{1}$$

$$cl_t^{\leqslant} = \cup_{s \leqslant t} c_s, \tag{2}$$

where $t = 1,\ldots,n$.

The statement $x \in cl_t^{\geqslant}$ reads "x belongs to at least class cl_t", while $x \in cl_t^{\leqslant}$ reads "x belongs to at most class cl_t". Given a set of attributes $p \sqsubseteq c$ and $x \epsilon U$, the granules of knowledge used in DRSA for the approximation of the unions $cl_t^{\geqslant}$ and $cl_t^{\leqslant}$ are the open sets defined by the dominance cones with respect to x (Zhai, Khoo, & Zhong, 2009b). A set of objects dominating x and dominated by x are, respectively, called the P-dominating set and the P-dominated set:

$$D_p^+(x) = \{y \in U : yD_px\} \tag{3}$$

$$D_p^-(x) = \{y \in U : xD_py\} \tag{4}$$

Objects satisfying the dominance principle are called consistent and those that violate the dominance principle are called inconsistent. This inconsistency in the sense of the dominance principle is caused by the inclusion of object $x \epsilon U$ in the upward union of classes $cl_t^{\geqslant}$ for $t = 2,\ldots,n$, given that set of criteria $p \sqsubseteq c$ is true and provided one of the following conditions hold:

- x belongs to class cl_t or better, but is p-dominated by object y belonging to a class worse than cl_t; that is, $x \in cl_t^{\geqslant}$ but $D_p^+(x) \cap cl_{t-1}^{\leqslant} \neq \phi$.
- x belongs to a worse class than cl_t, but p-dominates object y belonging to class cl_t or better; that is, $x \notin cl_t^{\geqslant}$, but $D_p^-(x) \cap cl_t^{\geqslant} \neq \phi$ (Zhai et al., 2009b).

The p-lower and p-upper approximations of $cl_t^{\geqslant}, t \in \{1,\ldots,n\}$ (denotation $\underline{p}(cl_t^{\geqslant})$) and $\bar{p}(cl_t^{\geqslant})$), respectively, for $p \sqsubseteq c$ are:

$$\underline{p}(cl_t^{\geqslant}) = \left\{x \in U : yD_p^+x \subseteq (cl_t^{\geqslant}\right\} \tag{5}$$