Data Mining Approaches for Geo-Spatial Big Data: Uncertainty Issues

Frederick E. Petry, Geospatial Science and Technology Branch, Naval Research Laboratory, Stennis Space Center, MS, USA

ABSTRACT

The availability of a vast amount of heterogeneous information from a variety of sources ranging from satellite imagery to the Internet has been termed as the problem of Big Data. Currently there is a great emphasis on the huge amount of geophysical data that has a spatial basis or spatial aspects. To effectively utilize such volumes of data, data mining techniques are needed to manage discovery from such volumes of data. An important consideration for this sort of data mining is to extend techniques to manage the inherent uncertainty involved in such spatial data. In this paper the authors first provide overviews of uncertainty representations based on fuzzy, intuitionistic, and rough sets theory and data mining techniques. To illustrate the issues they focus on the application of the discovery of association rules in approaches for vague spatial data. The extensions of association rule extraction for uncertain data as represented by rough and fuzzy sets are described. Finally an example of rule extraction for both fuzzy and rough set types of uncertainty representations is given.

INTRODUCTION

Many large research efforts are currently focused on the problem known as Big Data (Boyd & Crawford, 2012, Michael & Miller, 2013). Issues for this involve effectively utilizing a vast amount of heterogeneous information from a variety of sources (Shekar, et al., 2012). Currently there is a great emphasis on the geophysical data that has a spatial basis or spatial aspects (Overpeck, et al., 2011). Advances in instrumentation and sensors have hugely increased the volume, velocity and variety of remote sensed data. For example the imagery data archived at the NASA EOSDIS (Earth Observing System Data and Information System) exceeds 3 PB (Petabytes) and is generating 5 TB (Terabytes) of data per day. To effectively utilize such volumes of data, data mining techniques are very critical (Vatsavi, et al., 2012). One factor that must be considered in particular is how to deal with the inherent uncertainty involved with the huge amount of such spatial data in databases.

Data mining or knowledge discovery (Witten, Frank & Hall 2011; Kantardzic, 2011)
generally refers to a variety of techniques that have developed in the fields of databases, machine learning (Alpaydın, 2004) and pattern recognition (Han & Kamber 2006). The intent is to uncover useful patterns and associations from large databases. For complex data such as that found in spatial databases (Shekar & Chawla 2003) the problem of data discovery is more involved (Lu et al., 1993, Miller & Han 2009).

Spatial data has traditionally been the domain of geography with various forms of maps as the standard representation. With the advent of computerization of maps, geographic information systems (GIS) have come to fore with spatial databases storing the underlying point, line and area structures needed to support GIS (Longley et al., 2010). A major difference between data mining in ordinary relational databases (Elmasri & Navathe 2010) and in spatial databases is that attributes of the neighbors of some object of interest may have an influence on the object and therefore have to be considered as well. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations), which are used by spatial data mining algorithms (Ester et al., 2000).

Additionally when wish to consider vague-ness or uncertainty in the spatial data mining process (Burrough & Frank 1996, Zhang & Goodchild 2002), an additional level of difficulty is added. In this chapter we describe one of the most common data mining approaches, discovery of association rules, for spatial data for which we consider uncertainty in the extraction rules as represented by both fuzzy set and rough set techniques.

BACKGROUND

Uncertainty Representations

Generally uncertainty is considered as arising from aleatory and epistemic sources (Parsons, 2001) where aleatory uncertainty is typically modeled by a probabilistic approach. Epistemic uncertainty arises from subjective consider-

ations and is our focus here. In this section we overview the uncertainty representations we will use for data discovery in spatial data, specifically fuzzy, intuitionistic, and rough sets theory (Saikia 2010).

Fuzzy Set Theory

Fuzzy set theory (Pedrycz & Gomide 1996, Zimmerman 2012) is an approach in which the elements of a set belong to the set to varying degrees known as membership degrees. Conventionally we can specify a set C by its characteristic function, Char C (x). If U is the universal set from which values of C are taken, then we can represent C as:

\[ C = \{ x | x \in U \text{ and } \text{Char}_C(x) = 1 \} \]

This is the representation for a crisp or non-fuzzy set. For an ordinary set C the range of Char C (x) are just the two values: \{ 0, 1 \}. However for a fuzzy set A we have a range of the entire interval [0,1].

That is, for a fuzzy set the characteristic function takes on all values between 0 and 1 and not just the discrete values of 0 or 1 representing the binary choice for membership in a conventional crisp set. For a fuzzy set the characteristic function is often called the membership function and denoted \( \mu_A(x) \).

One fuzzy set concept that we employ particularly in databases is the similarity relation, S(x, y), denoted also as xSy. For given domain D this is a mapping of every pair of values in the particular domain onto the unit interval [0, 1], which reflects the level of similarity between them. A similarity relation is reflexive and symmetric as a traditional identity relation. However, special forms of transitivity are used so a similarity relation has the following three properties, for \( x, y, z \in D \) (Zadeh 1970, Buckles & Petry 1982):

1. Reflexive: \( s_D(x, x) = 1 \);
2. Symmetric: \( s_D(x, y) = s_D(y, x) \);
3. Transitive: \( s_D(x, z) \geq \text{Max}_y \{ s_D(x, y), s_D(y, z) \} \): (T1).
The Computing of Digital Ecosystems
[www.igi-global.com/article/computing-digital-ecosystems/48209?camid=4v1a](www.igi-global.com/article/computing-digital-ecosystems/48209?camid=4v1a)

Population Diversity of Particle Swarm Optimizer Solving Single and Multi-Objective Problems
Shi Cheng, Yuhui Shi and Quande Qin (2012). *International Journal of Swarm Intelligence Research* (pp. 23-60).
[www.igi-global.com/article/population-diversity-particle-swarm-optimizer/75326?camid=4v1a](www.igi-global.com/article/population-diversity-particle-swarm-optimizer/75326?camid=4v1a)