Data Hierarchies for Generalization of Imprecise Data

D

Frederick Petry
Naval Research Laboratory, USA

Ronald R. Yager
Iona College, USA

INTRODUCTION

Issues related to managing imprecise data in areas as diverse as spatial and environmental data, forensic evidence and economics must be dealt with for effective decision making. In order to make use of such information, we have to settle on how the various pieces of data can be used to make a decision or take an action. This can involve some sort of summarization and generalization of the pieces of data as to what conclusions they can support (Yager, 1991; Kacpryzk, 1999; Dubois & Prade, 2000). A currently emerging issue is the management of uncertain information arising from multiple sources and of many forms that appear in the everyday activities and decisions of humans. This can range from data / information obtained by sensors to the subjective information from individuals or analysts. Today ever more massive amounts of multi-source heterogeneous data / information is prevalent such as in systems managing the problem of Big Data (Miller & Miller, 2013; Richards & Rowe, 1999) However while effective decision-making should be able to make use of all the available, relevant information about such combined uncertainty, assessment of the value of a generalization result is critical. One possible approach for such a generalization process can be found in the use of concept hierarchical generalization (Raschia & Mouaddib, 2002; Yager & Petry, 2006). In previous research the problem of evidence resolution was studied for crisp concept hierarchies (Petry & Yager, 2008).

As one example of where data generalization is needed for decision making is with data related to criminal forensics. Federal Bureau of Investigation (FBI) researchers have made use of GIS data for forensic evidence evaluation in criminal cases. Spatial distribution of soils (Pye, 2007), pollens (Brown, Smith & Elmhurst, 2002) and other trace evidence are represented by individual layers with uncertainty as to the exact spatial areas of such information. These are then overlaid, generalized and the areas aggregated to focus on possible sites of interest for further investigation of a crime.

For example, in the case of a suspicious death, depending on the environmental conditions, a medical examiner may provide a likely range of time of death, but allow a possible wider time interval. So, overlap between the above time of death and a temporal interval when a potential suspect might have been in the area of the murder could be crucial in an investigation. Also, forensic anthropology is concerned with evaluations of skeletal-age-at- death (Hoppa & Vaupel, 2002) and must deal with the uncertainties of missing remains and weathering effects to provide estimates of possible temporal age intervals.

To address these issues the use of fuzzy, interval valued and intuitionistic concept hierarchies for generalization can extend previous approaches to deal with the uncertainty of data. A number of approaches to characterizing such decompositions for the resolution of the evidence using these hierarchies is also needed. The characterization of hierarchies indicates that set decompositions are needed to represent the

DOI: 10.4018/978-1-7998-9220-5.ch117

uncertainty of the hierarchies. To characterize these decompositions granularity measures and overlap measures must be developed and examples of each discussed. Additionally, information measures can introduce to be used for these evaluations.

BACKGROUND

Uncertainty Representations

Here we review some of the most common representations of subjective uncertainty that have been developed in the computational intelligence area. We will specifically consider the first three described next, fuzzy set theory, interval valued sets, and intuitionistic fuzzy sets for approaches to generalization of such data to concept hierarchies. The other types could be used in a similar approach.

Fuzzy Set Theory

In ordinary sets a data values either belongs or does not belong to the set. However fuzzy set theory (Zadeh, 1965; Klir, St. Clair & Yuan, 1997) allows a gradual assessment of the membership of data values in a set described by of a membership function. Where elements can either belong or not belong to a regular set, with fuzzy sets elements can belong to the set to a certain degree with zero indicating not an element, one indicating complete membership, and values between zero and one indicating partial or uncertain membership in the set. For a domain S a fuzzy set is

$$Fs(S) = \{d_i / m(d_i); 0 \leq m(d_i) \leq 1\}, d_i \in S$$

Interval-Valued Sets

Uncertainty of data is commonly represented in many applications by the use of interval values. We will introduce here the formalisms for intervals and interval arithmetic (Moore, 1966; Deschrijver, 2007; Moore, Kearfott, & Cloud, 2009) as needed. We let D be the domain and intervals will be represented by the values of the lower bound, $z_{\dagger} = Lb(d_i)$ and an upper bound, $z^{\dagger} = Ub(d_i)$ of an interval $I(d_i)$, $d_i \in D$

$$I(d_i) = [z_{\dagger}, z^{\dagger}] = \{z \in D \mid z_{\dagger} \leq z \leq z^{\dagger}\}$$

For an interval $I(d_i)$, we define the size of the interval, Q, as the difference of the lower and upper bounds,

$$Q: I(d_i) \rightarrow R^{+}; Q(I(d_i)) = \mid z_{\dagger} - z^{\dagger} \mid$$

So a representation of uncertainty of a data value d_i by intervals using a lower bound, and an upper bound,

$$\{d_i \mid I(d_i) = [Lb(d_i), Ub(d_i)]\}$$