A Unified Fuzzy Data Model: Representation and Processing

Avichai Meged, Bar-Ilan University, Israel
Roy Gelbard, Bar-Ilan University, Israel

ABSTRACT

A novel fuzzy data representation model which enables data mining with standard tools is introduced. Many data elements in the world are fuzzy in nature. There is an obvious need to represent and process such data effectively and efficiently, using the same standard tools for crisp data that are popular with researchers and practitioners alike. Currently, however, standard tools cannot process or analyze data that are not adequately represented. The comprehensive data representation model put forward here extends principles of binary databases and provides a unified approach to all types of data: discrete and continuous, crisp and fuzzy. The model is illustrated on a baseline dataset and tested in clustering experiments matched against controlled groupings and a real dataset. The tests confirm that the implementation of the model not only enables the use of standard tools but also yields better results as regards segmentation and clustering of fuzzy datasets.

Keywords: Binary Databases, Clustering, Data Mining, Data Representation Models, Fuzzy Data, Fuzzy Databases

INTRODUCTION

Many types of data in our world are inherently fuzzy. Common vague descriptions such as “tall” or “short”, “skinny” or “fat” are often embedded in databases. Imprecise, uncertain or even incomplete data also reinforce fuzziness and make their representation an acute problem. The problem is acerbated when fuzzy data need to be processed by commercial tools which cannot handle fuzzy representations, thus making it impossible to apply standard techniques such as clustering, classification or association.

Typical fuzzy data representation models, such as the fuzzy relational models and fuzzy semantic models described in the background section, are unsuitable as input to standard clustering and mining tools which require one flat file matrix-like format such as a relational database table where each cell contains a single crisp value. The current paper proposes a method that enables the use of standard clustering and mining tools on fuzzy data. Since clustering has become increasingly popular as a data mining technique (Giannotti & Pedreschi, 2008; Manying, 2007) we concentrate on showing how our fuzzy data model applies to clustering.

Clustering is crucial in the social sciences, marketing, finance, computer science, biology, medicine and elsewhere. Standard data analysis and mining tools such as SPSS, SAS or Clementine implement widely available clustering algorithms methods, and therefore are preferable...
over proprietary tools (for a brief description of clustering issues see Estivill-Castro & Yang, 2004; Gan, Ma, & Wu, 2007; Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999; Lim, Loh, & Shih, 2000; Zhang & Srihari, 2004). There is an ongoing effort to devise a data model to resolve the discrepancy between the format used to store the database and the representation format demanded by clustering algorithms (Ryu & Eick, 2005). The few studies that have dealt with clustering of fuzzy data have been partially successful but the clustering methods were restricted to specific fuzzy data types and used dedicated and proprietary algorithms as described in the background section.

The current paper proposes a unified data representation model that can deal with both crisp and fuzzy data, thus enabling the use of standard clustering and mining tools. The model draws on principles of binary representation employed in commercial databases, motion databases and fuzzy databases (Gelbard & Meged, 2008; Gelbard & Spiegler, 2002; Spiegler & Maayan, 1985). In these binary database models, the data are represented in a matrix where the rows stand for the database entities and the columns stand for different attribute values. In the proposed model, matrix cells are numbers that indicate degrees of attribute value similarity to the “right value”. These similarity numbers are derived from fuzzy database models such as the possibility distribution model (Prade & Testemale, 1984) and the proximity-based fuzzy relational model (Shenoi & Melton, 1989). The model is described and illustrated on a simple baseline dataset. It is tested by a controlled clustering experiment using SPSS software.

**BACKGROUND**

In this section we examine the following four topics: fuzzy data, fuzzy data models, binary database models and mining/clustering of fuzzy data.

**Fuzzy Data**

Many types of data in our world are considered inherently imperfect. The following classification defines the basic characteristics exhibited by imperfect data (Ma, 2005; Motro, 1995):

- **Uncertainty**: It not possible to determine whether the information is true or false, e.g., John may be 38 years old.
- **Imprecision**: The information available is not specific enough, e.g., John may be between 37 and 43 years old, John is 34 or 43 years old, or even unknown.
- **Vagueness**: The information is represented in linguistic terms, e.g., John is in his early years, or John is young.
- **Inconsistency**: The statement contains two or more pieces of information which cannot be true at the same time, e.g., John is 35 and 37 years old.
- **Ambiguity**: Some elements in the data formulation lack complete semantics so that several interpretations are possible, e.g., it is not clear whether John is young or old.

Bahri, Bouaziz, Chakhar, and Naija (2005) used this type of classification to enumerate the different data types that support almost all kinds of data. These include:

- **Simple numbers**, e.g., his age is 30.
- **Matching degree**, e.g., quality = 0.7.
- **Symbolic**, e.g., the color is red.
- **Range**, e.g., his age is between 25 and 35.
- **Fuzzy Range**, e.g., his age is more or less between 20 and 30.
- **Approximate value**, e.g., his age is about 35.
- **More/Less than value**, e.g., his age is more/less than 35.
- **Linguistic label (Categorical)**, e.g., the person is very young.
- **Set of possible scalar assignments**, e.g., height = {tall, very tall}. 

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Related Content

Evaluation of Quantified Statements Using Gradual Numbers
www.igi-global.com/chapter/evaluation-quantified-statements-using-gradual/20356?camid=4v1a
Conflicts, Compromises, and Political Decisions: Methodological Challenges of Enterprise-Wide E-Business Architecture Creation
www.igi-global.com/chapter/conflicts-compromises-political-decisions/8010?camid=4v1a

Ara Irititja: Towards Culturally Appropriate IT Best Practice in Remote Indigenous Australia
www.igi-global.com/chapter/ara-irititja-towards-culturally-appropriate/7985?camid=4v1a

Extended Spatiotemporal UML: Motivations, Requirements and Constructs
www.igi-global.com/article/extended-spatiotemporal-uml/3255?camid=4v1a