Chapter 7.12
Handling Fuzzy Similarity for Data Classification

Roy Gelberd
Bar-Ilan University, Israel

Avichai Meged
Bar-Ilan University, Israel

INTRODUCTION

Representing and consequently processing fuzzy data in standard and binary databases is problematic. The problem is further amplified in binary databases where continuous data is represented by means of discrete '1' and '0' bits. As regards classification, the problem becomes even more acute. In these cases, we may want to group objects based on some fuzzy attributes, but unfortunately, an appropriate fuzzy similarity measure is not always easy to find. The current paper proposes a novel model and measure for representing fuzzy data, which lends itself to both classification and data mining.

Classification algorithms and data mining attempt to set up hypotheses regarding the assigning of different objects to groups and classes on the basis of the similarity/distance between them (Estivill-Castro & Yang, 2004) (Lim, Loh & Shih, 2000) (Zhang & Srihari, 2004). Classification algorithms and data mining are widely used in numerous fields including: social sciences, where observations and questionnaires are used in learning mechanisms of social behavior; marketing, for segmentation and customer profiling; finance, for fraud detection; computer science, for image processing and expert systems applications; medicine, for diagnostics; and many other fields.

Classification algorithms and data mining methodologies are based on a procedure that calculates a similarity matrix based on similarity index between objects and on a grouping technique. Researches proved that a similarity measure based upon binary data representation yields better results than regular similarity indexes (Erlich, Gelbard & Spiegler, 2002) (Gelbard, Goldman & Spiegler, 2007). However, binary representation is currently limited to nominal discrete attributes suitable for attributes such as: gender, marital status, etc., (Zhang & Srihari, 2003). This makes the binary approach for data representation unattractive for widespread data types.

Copyright © 2009, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

Table 1. Standard binary representation table

	Regular Representation		Binary Representation							
Entity ID	Marital Status	Height	S	M	D	W	1.55	1.56	1.60	1.84
1	Married	1.60	0	1	0	0	0	0	1	0
2	Divorced	1.55	0	0	1	0	1	0	0	0
3	Single	1.84	1	0	0	0	0	0	0	1
4	Widowed	1.56	0	0	0	1	0	1	0	0
5	Single	1.60	1	0	0	0	0	0	1	0

The current research describes a novel approach to binary representation, referred to as Fuzzy Binary Representation. This new approach is suitable for all data types - nominal, ordinal and as continuous. We propose that there is meaning not only to the actual explicit attribute value, but also to its implicit similarity to other possible attribute values. These similarities can either be determined by a problem domain expert or automatically by analyzing fuzzy functions that represent the problem domain. The added new fuzzy similarity yields improved classification and data mining results. More generally, Fuzzy Binary Representation and related similarity measures exemplify that a refined and carefully designed handling of data, including eliciting of domain expertise regarding similarity, may add both value and knowledge to existing databases.

BACKGROUND

Binary Representation

Binary representation creates a storage scheme, wherein data appear in binary form rather than the common numeric and alphanumeric formats. The database is viewed as a two-dimensional matrix that relates entities according to their attribute values. Having the rows represent entities and the columns represent possible values, entries in the matrix are either '1' or '0', indicating that a given entity (e.g., record, object) has or lacks a given value, respectively (Spiegler & Maayan, 1985).

In this way, we can have a binary representation for discrete and continuous attributes.

Table 1 illustrates binary representation of a database consists of five entities with the following two attributes: Marital Status (nominal) and Height (continuous).

- Marital Status, with four values: **S** (single), **M** (married), **D** (divorced), **W** (widowed).
- Heights, with four values: **1.55**, **1.56**, **1.60** and **1.84**.

However, practically, binary representation is currently limited to nominal discrete attributes only. In the current study, we extend the binary model to include continuous data and fuzzy representation.

Similarity Measures

Similarity/distance measures are essential and at the heart of all classification algorithms. The most commonly-used method for calculating similarity is the Squared Euclidean measure. This measure calculates the distance between two samples as the square root of the sums of all squared distances between their properties (Jain & Dubes, 1988) (Jain, Murty & Flynn, 1999).

However, these likelihood-similarity measures are applicable only to ordinal attributes and cannot

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/handling-fuzzy-similarity-data-classification/8028

Related Content

Business Data Warehouse: The Case of Wal-Mart
Indranil Bose (2009). *Selected Readings on Database Technologies and Applications (pp. 112-133).*
www.irma-international.org/chapter/business-data-warehouse/28549

Introducing Fuzziness in Existing Orthogonal Persistence Interfaces and Systems
Miguel Ángel Sicilia, Elena Garcia-Barriocanaland José A. Gutierrez (2005). *Advances in Fuzzy Object-Oriented Databases: Modeling and Applications (pp. 241-268).*
www.irma-international.org/chapter/introducing-fuzziness-existing-orthogonal-persistence/4813

Adaptive Indexing in Very Large Databases
Andrew Johnsonand Farshad Fotouhi (1995). *Journal of Database Management (pp. 4-13).*
www.irma-international.org/article/adaptive-indexing-very-large-databases/51142

Challenges in Data Mining on Medical Databases
Fatemeh Hosseinkhah, Hassan Ashktorab, Ranjit Veenand M. Mehdi Owrang O. (2009). *Database Technologies: Concepts, Methodologies, Tools, and Applications (pp. 1393-1404).*
www.irma-international.org/chapter/challenges-data-mining-medical-databases/7980

INDUSTRY AND PRACTICE: Information Systems: Which Came First, The Information or the Systems?
Mark L. Gillenson (1997). *Journal of Database Management (pp. 37-38).*
www.irma-international.org/article/industry-practice-information-systems-came/51175