Handling Fuzzy Similarity for Data Classification

Roy Gelbard
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.

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. 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

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 be used to classify nominal, discrete, or categorical attributes, since there is no meaning in placing such attribute values in a common Euclidean space. A similarity measure, which applicable to nominal attributes and used in our research is the Dice (Dice 1945).

Additional binary similarity measures were developed and presented (Illingworth, Glaser & Pyle, 1983) (Zhang & Srihari, 2003). Similarities measures between the different attribute values, as proposed in Zadeh (1971) model, are essential in the classification process.

In the current study we use similarities between entities and between entity's attribute values to get better classification. Following former reserches, (Gelbard & Spiegler, 2000) (Erlich, Gelbard & Spiegler, 2002), the current study also uses Dice measure.

Fuzzy Logic

The theory of Fuzzy Logic was first introduced by Lotfi Zadeh (Zadeh, 1965). In classical logic, the only possible *truth-values* are *true* and *false*. In Fuzzy Logic; however, more *truth-values* are possible beyond the simple true and false. Fuzzy logic, then, derived from fuzzy set theory, is designed for situations where information is inexact and traditional digital on/off decisions are not possible.

Fuzzy sets are an extension of classical set theory and are used in fuzzy logic. In classical set theory, membership of elements in relation to a set is assessed according to a clear condition; an element either belongs or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in relation to a set; this is described with the aid of a membership function $\mu \rightarrow [0, 1]$. An element mapped to the value 0 means that the member is not included in the given set, '1' describes a fully included member, and all values between 0 and 1 characterize the fuzzy members. For example, the continuous variable "Height" may have three membership functions; stand for "Short", "Medium" and "Tall" categories. An object may belong to few categories in different membership degree, e.g 180 cm. height may belong to the "Medium" and "Tall" categories, in different