Chapter 11
Fuzzy Association Rules to Summarise Multiple Taxonomies in Large Databases

Trevor Martin
University of Bristol, UK; Intelligent Systems Research Centre, BT Innovate, UK

Yun Shen
University of Bristol, UK

ABSTRACT

When working with large datasets, a natural approach is to group similar items into categories (or sets) and summarise the data in terms of such categories. Fuzzy set theory allows us to represent and reason about sets of objects without providing crisp definitions for each group, an approach that often reflects the human interpretation of categories. Given two or more hierarchical sets of categories, our aim is to determine the correspondence between categories (e.g., approximate equivalence). Association rules are a useful tool in knowledge discovery from databases but are normally defined in terms of crisp rather than fuzzy categories. In this chapter, the authors describe a new method for calculating a fuzzy confidence value for association rules between fuzzy categories, using a novel approach based on mass assignment theory.

INTRODUCTION

A key feature of human intelligence is our ability to categorise and summarise large quantities of data, whether this data arises from sensory input or from other sources. The ability to group multiple entities together into an (approximately) uniform whole allows us to efficiently represent a whole group as a single concept, enabling us to reason, and to derive knowledge, about groups of entities. A simple form of derived knowledge is association - essentially, that the extensions of two concepts overlap significantly. One of the fundamental tenets underlying fuzzy set theory (Zadeh, 1965) is the idea that humans work with groups of entities (or conceptual categories) that are loosely defined, able to admit elements according to some scale of membership rather than according to an absolute yes/no test. This

DOI: 10.4018/978-1-60566-858-1.ch011
is particularly true where the knowledge and/or reasoning uses natural language - humans can communicate quickly and efficiently with an informal shared understanding of the vocabulary. Although different individuals may have slightly different interpretations of terms, meaning can still be conveyed sufficiently accurately in almost all cases.

A further step in the idea of grouping entities together leads us to the notion of a taxonomy, i.e. a hierarchical series of progressively more refined categories. This enables us to represent / reason about problems at the appropriate level of granularity, and the use of taxonomic hierarchies to organise information and sets of objects into manageable chunks (granules) is widespread. For example, taxonomies serve as the main organisational principle for the grouping of species, for systems of government (national - regional - local), for corporate and command structures, for libraries, for document repositories and very many other applications.

Granules were informally defined by (Zadeh, 1997) as a way of decomposing a whole into parts, generally in a hierarchical way using fuzzy representations. Although in principle a taxonomic hierarchy is crisply defined, in practice there is often a degree of arbitrariness in its definition. For example, we might divide the countries of the world by continent at the top level of a taxonomic hierarchy. However, continents do not have crisp definitions - Europe contains some definite members (e.g. France, Germany) but at the Eastern and South-Eastern border, the question of which countries belong / do not belong is less clear. Iceland is generally included in Europe despite being physically closer to Greenland (part of North America). Thus although the word “Europe” denotes a set of countries (i.e. it is a granule) and can be used as the basis for communication between humans, it does not have an unambiguous definition in terms of the elements that belong to the set. Different “authorities” adopt different definitions - the set of countries eligible to enter European football competitions differs from the set of countries eligible to enter the Eurovision song contest, for example.

Of course, mathematical and some legal taxonomic structures can be very precisely defined - in plane geometry, the class of polyhedra further subdivides into triangles, quadrilaterals, etc and triangles may be subdivided into equilateral, isosceles etc. Such definitions admit no uncertainty. Most information systems model the world in some way, and need to represent categories which correspond to the loosely defined classes used by humans in natural language. For example, a company may wish to divide adults into customers and non-customers, and then sub-divide these into high-value customers, dissatisfied customers, potential customers, etc. Such categories are not necessarily distinct (i.e. they may be a covering rather than a partition) but more importantly, membership in these categories is graded - customer $X$ may be highly dissatisfied and about to find a new supplier whilst customer $Y$ is only mildly dissatisfied. We argue that most hierarchical taxonomies involve graded or loosely defined categories, but the nature of computerised information systems means that a more-or-less arbitrary decision has to be made on borderline cases, giving the taxonomy the appearance of a crisp, well-defined hierarchy. This may not be a problem as long as a rigorous and consistent criterion for membership is used (e.g. a dissatisfied customer is defined as one who has made at least two calls complaining about service), but the lack of subjectivity in a definition is rare. The use of graded membership (fuzziness) in categories enhances their expressive power and usefulness.

There is rarely a unique way of splitting data into conceptual categories, and numerous methodologies exist to aid with design of databases and similar information systems. Our previous work “Smart Queries and Adaptive Data” (SQuAD) project is concerned with adding structure to data (i.e. moving up the metadata scale) and refining approximate knowledge (in the form of fuzzy association rules) from this data (see (Martin et al., 2008b) (Martin & Azvine, 2003, 2005; Martin et al., 2007b)). The aim is to assist in
Related Content

A Filter-and-Fan Metaheuristic for the 0-1 Multidimensional Knapsack Problem
www.irma-international.org/article/filter-fan-metaheuristic-multidimensional-knapsack/74738/

Home Blood Pressure Measurement: A Review
www.irma-international.org/article/home-blood-pressure-measurement/51670/

Discrete Artificial Bee Colony Optimization Algorithm for Financial Classification Problems
www.irma-international.org/article/discrete-artificial-bee-colony-optimization/52789/

Bi-Criteria Optimization for Finding the Optimal Replacement Interval for Maintaining the Performance of the Process Industries
www.irma-international.org/chapter/bi-criteria-optimization-for-finding-the-optimal-replacement-interval-for-maintaining-the-performance-of-the-process-industries/147533/

Marriage in Honeybee Optimization to Scheduling Problems
www.irma-international.org/chapter/marriage-honeybee-optimization-scheduling-problems/58521/