Fuzzy Methods in Data Mining

Eyke Hüllermeier
Philipps-Universität Marburg, Germany

INTRODUCTION

Tools and techniques that have been developed during the last 40 years in the field of fuzzy set theory (FST) have been applied quite successfully in a variety of application areas. A prominent example of the practical usefulness of corresponding techniques is *fuzzy control*, where the idea is to represent the input-output behaviour of a controller (of a technical system) in terms of fuzzy rules. A concrete control function is derived from such rules by means of suitable inference techniques.

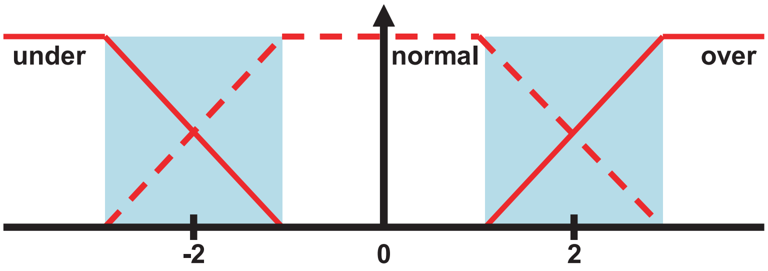
While aspects of knowledge representation and reasoning have dominated research in FST for a long time, problems of *automated learning and knowledge acquisition* have more and more come to the fore in recent years. There are several reasons for this development, notably the following: Firstly, there has been an internal shift within fuzzy systems research from "modelling" to "learning", which can be attributed to the awareness that the well-known "knowledge acquisition bottleneck" seems to remain one of the key problems in the design of intelligent and knowledge-based systems. Secondly, this trend has been further amplified by the great interest that the fields of *knowledge discovery in databases* (KDD) and its core methodical component, *data mining*, have attracted in recent years.

It is hence hardly surprising that data mining has received a great deal of attention in the FST community in recent years (Hüllermeier, 2005). The aim of this chapter is to give an idea of the usefulness of FST for data mining. To this end, we shall briefly highlight, in the next but one section, some potential advantages of fuzzy approaches. In preparation, the next section briefly recalls some basic ideas and concepts from FST. The style of presentation is purely non-technical throughout; for technical details we shall give pointers to the literature.

BACKGROUND ON FUZZY SETS

A fuzzy subset F of a reference set X is identified by a so-called *membership function* (often denoted $\mu_F(\bullet)$), which is a generalization of the characteristic function of an ordinary set $A \subseteq X$ (Zadeh, 1965). For each element $x \in X$, this function specifies the *degree of membership* of x in the fuzzy set. Usually, membership degrees $\mu_F(x)$ are taken from the unit interval [0,1], i.e., a membership function is an $X \rightarrow [0,1]$ mapping, even though more general membership scales (such as ordinal scales or complete lattices) are conceivable.

Figure 1. Fuzzy partition of the gene expression level with a "smooth" transition (grey regions) between underexpression, normal expression, and overexpression



Fuzzy sets formalize the idea of *graded membership* according to which an element can belong "more or less" to a set. Consequently, a fuzzy set can have "non-sharp" boundaries. Many sets or concepts associated with natural language terms have boundaries that are non-sharp in the sense of FST. Consider the concept of "forest" as an example. For many collections of trees and plants it will be quite difficult to decide in an unequivocal way whether or not one should call them a forest.

In a data mining context, the idea of "non-sharp" boundaries is especially useful for discretizing numerical attributes, a common preprocessing step in data analysis. For example, in gene expression analysis, one typically distinguishes between *normally expressed*, *underexpressed*, and *overexpressed* genes. This classification is made on the basis of the expression level of the gene (a normalized numerical value), as measured by so-called DNA-chips, by using corresponding thresholds. For example, a gene is often called overexpressed if its expression level is at least twofold increased. Needless to say, corresponding thresholds (such as 2) are more or less arbitrary. Figure 1 shows a *fuzzy partition* of the expression level with a "smooth" transition between under-, normal, and overexpression. For instance, according to this formalization, a gene with an expression level of at least 3 is definitely considered overexpressed, below 1 it is definitely not overexpressed, but in-between, it is considered overexpressed to a certain degree (Ortoloani et al., 2004).

Fuzzy sets or, more specifically, membership degrees can have different semantic interpretations. Particularly, a fuzzy set can express three types of cognitive concepts which are of major importance in artificial intelligence, namely *uncertainty*, *similarity*, and *preference* (Dubois, 1997). To operate with fuzzy sets in a formal way, FST offers generalized set-theoretical respectively logical connectives and operators (as in the classical case, there is a close correspondence between set-theory and logic) such as triangular norms (t-norms, generalized logical conjunctions), t-conorms (generalized disjunctions), and generalized implication operators. For example, a t-norm $\otimes$ is a $[0,1]\times[0,1]\rightarrow[0,1]$ mapping which is associative, commutative, monotone increasing (in both arguments) and which satisfies the boundary conditions $\alpha \otimes 0 = 0$ and $\alpha \otimes 1 = \alpha$ for all $0 \leq \alpha \leq 1$ (Klement et al., 2002). Well-known examples of t-norms include the minimum $(\alpha, \beta) \mapsto \min(\alpha, \beta)$ and the product $(\alpha, \beta) \mapsto \alpha\beta$.

BENEFITS OF FUZZY DATA MINING

This section briefly highlights some potential contributions that FST can make to data mining; see (Hüllermeier, 2005) for a more detailed (technical) exposition.

Graduality

The ability to represent gradual concepts and fuzzy properties in a thorough way is one of the key features of fuzzy sets. This aspect is also of primary importance in the context of data mining. In fact, patterns that are of interest in data mining are often inherently vague and do have boundaries that are non-sharp in the sense of FST. To illustrate, consider the concept of a "peak": It is usually not possible to decide in an unequivocal way whether a timely ordered sequence of measurements, such as the expression profile of a gene, has a "peak" (a particular kind of pattern) or not. Rather, there is a gradual transition between having a peak and not having a peak. Likewise, the spatial extension of patterns like a "cluster of points" or a "region of high density" in a data space will usually have soft rather than sharp boundaries.

Many data mining methods proceed from a representation of the entities under consideration in terms of *feature vectors*, i.e., a fixed number of features or attributes, each of which represents a certain property

Figure 2. Three exemplary time series that are more or less "decreasing at the beginning"