

Extracting Fuzzy Measures from Sample Data

Xiaojing Wang and Martine Ceberio
University of Texas at El Paso

April 03, 2010

Abstract

Fuzzy measures and integrals have been successfully applied in different areas, such as multicriteria decision making (MCDM), pattern recognition, data mining, information fusion, game strategies evaluation, gene classification, and so on. Compared to the traditional additive measures which assume all attributes involved are independent, fuzzy measures replace the additive property by a weaker one, monotonicity, which is able to express the interaction of relevant attributes. Thus, fuzzy measures are more useful due to relationships among the attributes in most of the real applications. However, it is difficult to identify fuzzy measures in practical applications. The number of parameters to be identified for a fuzzy measure exponentially increases with the number of attributes.

Many algorithms have been proposed to solve this problem, including gradient descent algorithms (GD), neural networks, genetic algorithms (GA). Even though these approaches can identify fuzzy measures effectively, they have some common limitations including easily falling into locally optimal solutions and taking time to verify the monotonicity property.

In practice, it is possible to have a lot of sample data on which we can apply optimization algorithms to find the best fuzzy measure to fit them, that is, the minimum difference between the actual value and the measured value. The Choquet integral, a non-linear function with respect to fuzzy measures, is used to compute the measured value.

We consider extracting fuzzy measures from the learning data as a constraints satisfaction problem (CSP) and propose to use existing reliable global solvers to solve this problem. In this case, the constraints model the monotonicity property of the fuzzy measure, and the objective function is the least squares of the Choquet integral. Ideally, a fuzzy measure that exactly matches all the sample data would be found. However, it is almost impossible since the real data always contain noise. To address this issue, an interval-based approach is proposed to identify an approximately optimal solution.

We will then apply our approach to a problem of early diagnosis or prognosis of cancers.