Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fuzzy Sets and Systems 150 (2005) 179–197
FUZZY sets and systems
www.elsevier.com/locate/fss

Interpretability issues in data-based learning of fuzzy systems

Ralf Mikut*, Jens Jäkel, Lutz Gröll

Forschungszentrum Karlsruhe GmbH, Institute for Applied Computer Science, P.O. Box 3640, D-76021 Karlsruhe, Germany

Received 1 October 2003; received in revised form 19 April 2004; accepted 18 June 2004
Available online 8 July 2004

Abstract

This paper presents a method for an automatic and complete design of fuzzy systems from data. The main objective is to build fuzzy systems with a user-controllable trade-off between accuracy and interpretability. Whereas criteria for accuracy mostly follow straightforwardly from the application, definition of interpretability and its criteria are subject to controversial discussion. For this reason, a set of interpretability criteria is given which guide the design process. Consequently, interpretability is maintained by structural choices regarding the type of membership functions, rules, and inference mechanism, on the one hand, and by including interpretability criteria in the rule/rule base evaluation, on the other hand. An application in Instrumented Gait Analysis, to characterize a certain group of patients in comparison to healthy subjects, illustrates the proposed algorithm.

Keywords: Classification; Data mining; Decision tree; Linguistic fuzzy system; Fuzzy rule; Inductive learning; Interpretability; Machine learning

1. Introduction

Interpretability is considered to be the main advantage of fuzzy systems over alternatives like statistical models or neural networks. Interpretability means that human beings are able to understand the fuzzy system's behavior by inspecting the rule base. It is crucial in the field of data mining and knowledge discovery where knowledge should be extracted from data bases and represented in a comprehensible form or for decision support systems where the reasoning process should be transparent to the user. Fuzzy systems constructed from expert knowledge—the traditional approach—usually are well understandable.

* Corresponding author. Tel.: +49-7247-825726; fax: +49-7247-825786.
E-mail addresses: ralf.mikut@iai.fzk.de (R. Mikut), jens.jaekel@iai.fzk.de (J. Jäkel), lutz.groell@iai.fzk.de (L. Gröll).

0165-0114/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2004.06.006

At present, a vast number of algorithms exist for automatic data-based fuzzy modeling, which can be classified as clustering [2,16,54], neuro-fuzzy [8,26,30], machine learning [17,20,24,47,51] or evolutionary approaches [4,13,31]. However, fuzzy systems generated by these algorithms are not necessarily comprehensible, especially when the algorithm aims at reaching a maximum accuracy. In recent years, research has therefore started to focus on the trade-off between interpretability and accuracy (for an overview see [10]).

Whereas the definition of accuracy in a certain application is straightforward, the definition of interpretability is rather problematic. Most researchers and practitioners would agree on interpretability involving the following aspects [3,9,14,28,49]:

- The number of rules is small enough to be comprehensible. Moreover, rules should not contain degrees of plausibility or rule weights.
- The rule base is formed of rules describing (locally) relevant relationships. The rules are consistent (similar premises lead to similar conclusions).
- Rule premises should be easy in structure and contain only a few features (input variables) only.
- The fuzzy system should preferably use features and combinations of these, which are familiar to the user.
- Linguistic terms should be intuitively comprehensible. The form and parameters of the membership functions should correspond to the understanding of the linguistic expressions.
- The inference mechanism should produce technically and intuitively correct results.

Generally, interpretability can be maintained or enhanced during the fuzzy system's generation or obtained by post-processing of the resulting data-driven fuzzy system.

Examples for the first approach include constraints on membership functions and their parameters [3,19,42], a special syntax of fuzzy rules [14,24] or a special structure of the fuzzy system, e.g. a hierarchical structure [15,22,52]. The second approach comprises simplification by merging similar fuzzy sets or rules [11,27,48,53] or using linguistic hedges [18,50].

This paper proposes a modular data-driven algorithm for fuzzy system learning according to the first alternative. Different elements to improve interpretability are explicitly (in form of evaluation measures) or implicitly (in form of efficient heuristics in all steps of the learning algorithm) integrated in this algorithm:

- Feature selection finds the most relevant features. The relevance measure can incorporate a priori information on preferred features (user or technical preference).
- Automatic generation of membership functions and labels takes their interpretability into account (form of membership functions, reasonable rounded parameters, adaptation to the distribution of feature values).
- Generation of rule hypotheses by decision tree induction and their pruning favor simple premises and lead to derived linguistic terms.
- Textual presentation of rules provides additional information in natural language and is better readable than a formal fuzzy rule.

An implementation as MATLAB toolbox KAFKA enables the user to solve complex real-world problems interactively controlling the trade-off between interpretability and accuracy.

The paper is organized as follows: Section 2 introduces the data and a priori information as the input to the learning algorithm as well as basic settings of the fuzzy system. Criteria for evaluating features and