Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 281 (2015) 252–259

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy data mining and management of interpretable and subjective information

Christophe Marsala [a,b,*], Bernadette Bouchon-Meunier [a,b]

[a] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France*
[b] *CNRS, UMR 7606, LIP6, F-75005, Paris, France*

Received 5 January 2015; received in revised form 17 July 2015; accepted 24 August 2015

Available online 4 September 2015

Abstract

Fuzzy set theory offers an important contribution to data mining leading to fuzzy data mining. It enables the management of interpretable and subjective information in both input and output of the data mining process. In this paper, we discuss the notion of interpretability in fuzzy data mining and we present some references on the management of emotions as a particular kind of subjective information.

Keywords: Fuzzy data mining; Interpretable systems; Subjective information; Emotions

Introduction

Since its introduction by Lotfi Zadeh fifty years ago, fuzzy set theory has been fruitfully used in many domains. Data mining is one of the important fields of application, introduced in the 1990s as a particular step in knowledge discovery from database (KDD): "*The data-mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data-mining step of the KDD process*" [1]. Early after the appearance of data mining, several works have proposed the use of fuzzy set theory to this domain.

The contributions of fuzzy sets in data mining are various: increasing the interpretability, enhancing the robustness of the process and managing unclear information, in particular subjective and emotional information. Both are provided by the introduction of fuzzy set theory to build up *fuzzy data mining*, offering to that process the capacity to mine complex information difficult to treat in a classic environment, considering the particular case of emotions.

Interpretability is the focus of this paper and it is discussed hereafter. Robustness of the process enables it to produce similar results when facing only small changes in the data (for instance, in the presence of noise). Robustness of fuzzy systems has been extensively studied and is well-known even in non-machine learning domains.

* Corresponding author at: CNRS, UMR 7606, LIP6, F-75005, Paris, France. Tel: +33 144278806; fax: +33 144277000.
E-mail address: Christophe.Marsala@lip6.fr (C. Marsala).

http://dx.doi.org/10.1016/j.fss.2015.08.021
0165-0114/

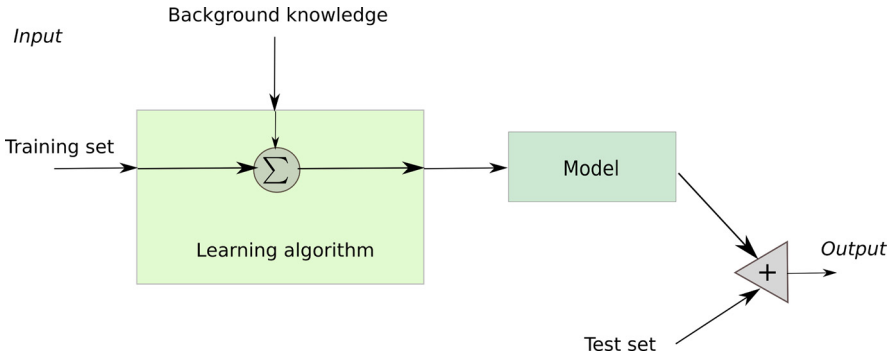

Fig. 1. Machine learning: from data to model.

Section 1 presents fuzzy data mining. In Section 2, interpretability in fuzzy data mining is discussed. In Section 3, the particular case of handling subjective information, such as emotions, is studied. Lastly, a conclusion and some future works are presented.

1. Fuzzy data mining

This section presents fuzzy data mining giving basic recalls on data mining and its links to machine learning. Then, various uses of fuzzy sets in data mining are presented.

Data mining refers to a global process that is composed of several steps (data pre-processing, learning, analyses, selection,...) [1]. Data mining and machine learning are two intertwined domains in the sense that data mining usually includes a machine learning algorithm as a step. In the literature, "data mining" can also refer to "the machine learning step of a data mining process". Here, data mining and machine learning differ mainly in the sense that each of them has a particular *aim*. In [2,3] the following difference is highlighted: "*Machine learning main aim is leaded by performances in predictive perspective, whereas data mining is related to understandability of discovered patterns*".

Machine learning aims to set up a *model* from a set of *data* providing some *background knowledge*. Data compose the training set available either to build up the model, or to tune it. If available, background knowledge is provided by an expert of the domain or is related to domain expertise. In this process, the model can be viewed as a new knowledge that is produced from the learning, it can be of various forms: e.g. mathematical function, neural network, rule base, patterns, association rules,...

Fig. 1 illustrates the links between these elements. It should be noted that the model can be included as a part of the learning algorithm if the algorithm is a way to tune it rather than to build it up completely (for instance, neural networks are tuned, and decision trees are constructed). In this figure, one possible use of the model is represented: the model can be exploited to provide an output for any forthcoming data (i.e. data that can be different from those present in the training set) composing the so-called test set. The test set, in a more general sense, can be associated with practical use cases of the model in practical situations.

Given that machine learning algorithms are highly reliable,[1] the choice of an algorithm to be used in data mining is made thanks to the interpretability of the model it builds.

Data mining is a process that is based on the use of a machine learning algorithm in a given application task. One main difference is that the constructed model is afterwards used not only to classify new cases, but also to provide an additional knowledge. Data mining is also concerned with accuracy of the obtained model in order to offer its validation or an evaluation of its performances [1]. In this case, models with a *good* accuracy rather than the *highest* one can be sufficient as the main aim is to obtain understandable knowledge on the data.

Fuzzy data mining is an extension of data mining where fuzzy set modelling is introduced. Various kinds of fuzzy data mining can be highlighted, depending on the application or the problem that has to be dealt with. The main

[1] Reliability in terms of *accuracy* refers to the efficiency of the model when applied to cases not belonging to the training set. The more accurate, the more usable the model for any forthcoming data and thus the more reliable it is.