Computers and Mathematics with Applications 66 (2013) 190–200

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Towards new directions of data mining by evolutionary fuzzy rules and symbolic regression

P. Krömer[a,b,*], S. Owais[c], J. Platoš[a,b], V. Snášel[a,b]

[a] Department of Computer Science, VŠB – Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava – Poruba, Czech Republic
[b] IT4Innovations Center of Excellence, VŠB – Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava – Poruba, Czech Republic
[c] Department of Computer Science, IT College, ASU Applied Science University, Amman, Jordan

ARTICLE INFO

Keywords:
Fuzzy rules
Genetic programming
Fuzzy information retrieval
Data mining
Application

ABSTRACT

There are various techniques for data mining and data analysis. Among them, hybrid approaches combining two or more fundamental methods gain importance as the complexity and dimension of real world problems and data sets grows. Fuzzy sets and fuzzy logic can be used for efficient data classification by the means of fuzzy rules and classifiers. This study presents an application of genetic programming to the evolution of fuzzy rules based on the concept of extended Boolean queries. Fuzzy rules are used as symbolic classifiers learned from data and used to label data records and to predict the value of an output variable. An example of the application of such a hybrid evolutionary-fuzzy data mining approach to a real world problem is presented.

1. Introduction

The recent time has seen a rise in the demand for advanced data mining algorithms. Many real world application domains generate huge amounts of data collected in vast data sets. Information hidden in such a data ought to be extracted and used for optimization of processes, decisions, designs, and algorithms. The growing dimension and complexity of said data sets represents a challenge for traditional data mining and optimization methods while the increase of power of widely available computers encourages the deployment of soft computing methods such as populational meta-heuristic algorithms, artificial neural networks, fuzzy systems, and hybrid methods.

Fuzzy sets and fuzzy logic can be utilized for efficient soft classification of data [1–3]. In contrast to crisp classification, which produces hard decisions about investigated data records, fuzzy classification allows a more sensitive analysis of data [4]. Fuzzy decision trees and if–then rules are examples of efficient, transparent, and easily interpretable fuzzy classifiers [4,5].

Genetic programming is a powerful machine learning algorithm from the wide family of evolutionary computation methods [6]. In contrast to the traditional evolutionary algorithms, it can be used to evolve complex hierarchical tree structures and symbolic expressions. It has been used to evolve Lisp *S*-expressions, mathematical functions, various symbolic expressions including crisp and fuzzy decision trees, and recently to infer search queries from relevance ranked documents in a fuzzy information retrieval system [7,8].

The last approach can be used for general data mining as well. Extended Boolean queries, that is weighted Boolean search expressions, can be interpreted as symbolic fuzzy rules that describe a fuzzy subset of some data set by means of its features. Moreover, a fuzzy rule evolved over a training data set can be later used for efficient and fast classification of new data

[*] Corresponding author at: Department of Computer Science, VŠB – Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava – Poruba, Czech Republic. Tel.: +420 721364104.
E-mail addresses: pavel.kromer@vsb.cz (P. Krömer), dr_suhail@asu.edu.jo (S. Owais), jan.platos@vsb.cz (J. Platoš), vaclav.snasel@vsb.cz (V. Snášel).

0898-1221/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.camwa.2013.02.017

samples in order to e.g. predict quality of products, detect harmful actions in a computer network, and in general to assign labels to data samples.

Artificial evolution of fuzzy rules is a promising approach to data mining because genetic programming proved very good ability to find symbolic expressions in various problem domains [9–11]. The general process of classifier evolution can be used to evolve custom classifiers for different data classes and various data sets with different properties and with different inner structure. The resulting classifiers can be used as standalone data labeling tools or participate in the collective decision making of an ensemble of classification and prediction tools.

The remainder of this study is organized as follows: first, fuzzy information retrieval as the background of the investigated fuzzy classifier is described. Section 3 describes genetic algorithms and genetic programming as the tools to learn fuzzy rules from data. Next, evolutionary fuzzy rules, their construction, evaluation, and application to a selected real world problem are presented. In Section 6 are outlined other symbolic regression methods suitable for fuzzy rule evolution and Section 7 concludes this study.

2. Fuzzy information retrieval

The evolution of fuzzy rules for data mining was implemented as a generalization of a previous method for search query optimization designed for efficient information retrieval and based on symbolic regression by genetic programming [7,8,12,13]. For fuzzy data mining, data samples are mapped onto documents and data features are mapped to index terms. This section outlines the principles of fuzzy information retrieval which forms the background of evolutionary fuzzy rules.

The area of Information Retrieval (IR) is a branch of computer science dealing with storage, maintenance, and searching in large amounts of data [14]. It defines and studies IR systems and models. An IR model is a formal background defining the document representation, query language, and document-query matching mechanism of an IR system (i.e. a search engine).

2.1. Extended Boolean IR model

The proposed fuzzy classification algorithm builds on the extended Boolean IR model, which is based on the fuzzy set theory and fuzzy logic [14–16]. In the extended Boolean IR model, documents are interpreted as fuzzy sets of indexed terms.

In each document, every indexed term has a weight from the range [0, 1] expressing the degree of significance of the term for document representation. Many different weighting approaches can be used to assign weights to index terms, e.g. the $tf \cdot idf_t$ term weighting scheme [17].

A formal description of a document collection in the extended Boolean IR model is shown in (1) and (2), where d_i represents i-th document, t_{ij} is j-th term in i-th document, m is the number of terms, and n is the number of documents. The entire document collection can be represented by an index matrix $\boldsymbol{D}$.

$$\boldsymbol{d}_i = (t_{i1}, t_{i2}, \ldots, t_{im}), \quad \forall t_{ij} \in [0, 1] \tag{1}$$

$$\boldsymbol{D} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{pmatrix}. \tag{2}$$

The query language in the extended Boolean model of IR is improved with the possibility of weighting query terms in order to attribute different levels of importance to those in a search request and by weighting (parameterizing) aggregation operators (most often AND, OR, and NOT) to soften or blur their impact on query evaluation [14–16].

Consider Q to be the set of user queries over a collection; then the weight of term t in query q is denoted as $a(q, t)$ satisfying $a : Q \times T \rightarrow [0, 1]$. To evaluate the atomic query of one term representing single search criterion the function $g : [0, 1] \times [0, 1] \rightarrow [0, 1]$ will be used. The value of $g(F(d, t), a)$ is called the Retrieval Status Value (RSV). For RSV evaluation the interpretation of the query term weight a is crucial. The most commonly used interpretations see the query term weight as the importance weight, threshold or ideal document description [14–16].

The theorems for the evaluation of RSV in the case of importance weight interpretation and threshold interpretation are shown in (3) and (4) respectively [14–16], where $P(a)$ and $Q(a)$ are coefficients used for tuning the threshold curve. An example of $P(a)$ and $Q(a)$ could be as follows: $P(a) = \frac{1+a}{2}$ and $Q(a) = \frac{1-a^2}{4}$. The RSV formula from (4) is illustrated in (1a). Adopting the threshold interpretation, an atomic query containing term t of the weight a is a request to retrieve documents having $F(d, t)$ equal or greater to a. Documents satisfying this condition will be rated with high RSV and contrariwise documents having $F(d, t)$ smaller than a will be rated with a small RSV.

$$g(F(d,t), a) = \begin{cases} \min(a, F(d,t)) & \text{when } t \text{ is operated by OR} \\ \max(1-a, F(d,t)) & \text{when } t \text{ is operated by AND} \end{cases} \tag{3}$$

$$g(F(d,t), a) = \begin{cases} P(a)\dfrac{F(d,t)}{a} & \text{for } F(d,t) < a \\ P(a) + Q(a)\dfrac{F(d,t)-a}{1-a} & \text{for } F(d,t) \geq a. \end{cases} \tag{4}$$