Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 258 (2015) 39–60

FUZZY sets and systems

www.elsevier.com/locate/fss

A fuzzy rough set approach for incremental feature selection on hybrid information systems

Anping Zeng [a,b], Tianrui Li [a,*], Dun Liu [c], Junbo Zhang [a], Hongmei Chen [a]

[a] *School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China*
[b] *School of Computer and Information Engineering, Yibin University, Yibin 644007, China*
[c] *School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China*

Received 3 September 2013; received in revised form 7 August 2014; accepted 22 August 2014

Available online 6 September 2014

Abstract

In real-applications, there may exist many kinds of data (*e.g.*, boolean, categorical, real-valued and set-valued data) and missing data in an information system which is called as a Hybrid Information System (HIS). A new Hybrid Distance (HD) in HIS is developed based on the value difference metric, and a novel fuzzy rough set is constructed by combining the HD distance and the Gaussian kernel. Considering the information systems often vary with time, the updating mechanisms for attribute reduction (feature selection) are analyzed with the variation of the attribute set. Fuzzy rough set approaches for incremental feature selection on HIS are presented. Then two corresponding incremental algorithms are proposed, respectively. Finally, extensive experiments on eight datasets from UCI and an artificial dataset show that the incremental approaches significantly outperform non-incremental approaches with feature selection in the computational time.

Keywords: Fuzzy rough sets; Incremental learning; Feature selection; Hybrid information systems; Big data

1. Introduction

Feature selection is an important technique in pattern recognition, machine leaning and data mining as there usually are many candidate attributes collected to represent recognition problems. In some real-world applications, data expand quickly, and tens, hundreds even thousands of attributes are stored in databases. These attributes may be of different types (*e.g.*, boolean, categorical, real-valued and set-valued data). The information system that includes different types of attributes is called as a Hybrid Information System (HIS). In HIS, some of attributes are irrelevant to the learning or recognition tasks. Experiments have shown irrelative attributes will deteriorate the performance of the learning algorithms for the cause of dimensionality, which increase training and testing times [1,2]. Feature selection

* Corresponding author.
E-mail addresses: zengap@126.com (A. Zeng), trli@swjtu.edu.cn (T. Li), newton83@163.com (D. Liu), junbozhang86@163.com (J. Zhang), hmchen@swjtu.edu.cn (H. Chen).

http://dx.doi.org/10.1016/j.fss.2014.08.014
0165-0114/

is an effective approach to remove the irrelative attributes for big data analysis. In recent years, feature selection in the HIS has received much attention. Many optimization algorithms of feature selection were presented to increase the classification accuracy in HIS [3–8].

The concept of feature selection (*aka.* attribute reduction), is viewed as one of the most important techniques in Rough Set Theory (RST) [9]. However, it is hard to use the traditional RST in handling hybrid attributes (*e.g.*, real-valued attributes). Discretization of the real-valued attributes is one way to solve this problem, but it may cause information loss. Another approach is the use of Fuzzy Rough Sets (FRS). FRS encapsulate the related but distinct concepts of fuzziness and indiscernibility, both occur as a result of uncertainty existing in knowledge [10–15].

Feature selection via FRS was first proposed in [16,17] which considers the problem of evaluating the hypoxic resistance of a patient on the basis of the values of his blood pressure during a barocamera examination. The measurements were evaluated by the FRS criterion. In [18], fuzzy-rough attribute reduction was proposed. The dependency function to measure the importance of attributes by FRS was defined and an algorithm to compute a reduct was designed. This algorithm was tested with some practical data sets from web categorization and was claimed to have better performance. In [19], Chen et al. introduced the concept of local reduction with FRS for decision systems. The local reduction can identify key conditional attributes and offer a minimal description for every single decision class. In [20], the FRS was used to compute both the relevance and significance of features. In [21,22], Hu et al. proposed Gaussian kernel based FRS and applied it into feature selection. In this model, Gaussian kernels were first introduced to acquire fuzzy relations between samples described by numeric attributes.

Nowadays we live in a world of big data [23–25]. Big data has the characteristics of 4Vs, *i.e.*, Volume, Variety, Velocity and Value [26]. Volume means the amount of data that needs to be handled is very large, and exabytes, zettabytes, and even higher amounts of data are described in big data applications. Variety means that the data is varied in nature, and structured, unstructured or semi-structured data needs to be properly combined to make the most of the analysis. Velocity means that a high rate of sampling is common in big data problems. Value means high yield will be achieved if the big data is used reasonable by analyzing correctly and accurately. Considering the velocity characteristic, algorithms which can improve efficiency of knowledge discovery are badly needed. Incremental learning is essential for the knowledge discovery which is an important manner of human intelligence. Especially in two cases, the first one is when the dataset cannot be collected at one time and the batch computing cannot be carried out, *e.g.* online applications [23] or interaction query [24]; the second case is when the data set is too big and it cannot be calculated at one turn due to the limitation of the computation capability and the memory size. The data set has to be cut to bulk and added in succession. Incremental learning is feasible when the data structure and information of the previous data is stored and the old data need not to be scanned again. Then the relationships between the new data added and the data structure stored are analyzed in order to require new knowledge. Therefore, in real-life applications, information systems may be very large [27–30] and vary with time. In FRS, the generation of fuzzy relations among samples is often very costly or even intractable. In fact, not only in FRS but also in other RST extensions, more and more computing problems arise due to the appearance of big data. In our study, the dynamic properties and the big size of the data set are both taken into consideration. Incremental learning method is studied which aims to improve the efficiency of feature selection under fuzzy rough set. Incremental learning which is fully used of the information get previously can improve the efficiency of knowledge discovery in big data.

The variation of objects has been widely considered in incremental learning. Shan et al. presented a discernibility-matrix based incremental methodology to find all maximally generalized rules [31]. Bang et al. proposed an incremental inductive learning algorithm to find a minimal set of rules for a decision table without recalculating all the set of instances when a new instance is added into the universe [32]. Liu et al. proposed an incremental approach for inducing interesting knowledge when the object set varies over time [33]. Zheng et al. presented an effective incremental approach for knowledge acquisition based on the rule tree [34]. Wang et al. constructed an incremental rule acquisition algorithm based on Variable Precision Rough Sets (VPRS) while inserting new objects into the information system [35]. Zhang et al. proposed an incremental rule acquisition algorithm based on neighborhood rough sets when the object set evolves over time [36]. Li et al. proposed a dynamic maintenance of approximations in Dominance-based Rough Sets Approach (DRSA) when adding or deleting objects in the universe [37]. In addition, Luo et al. proposed an incremental approaches for updating approximations in set-valued ordered information systems [38]. Zhang et al. proposed a composite rough set model for dynamic data mining [39]. In FRS, Zeng et al. proposed an incremental approach for updating approximations of Gaussian kernelized FRS under the variation of the object set [40]. In attribute reduction, several incremental reduction algorithms have been proposed to deal with