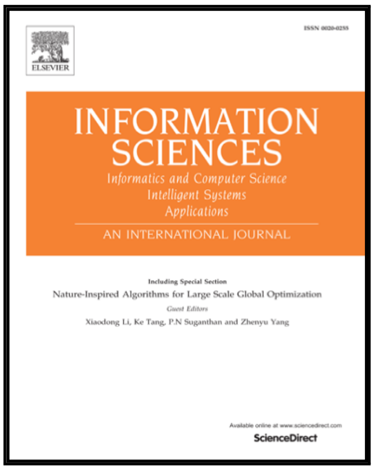
Accepted Manuscript



Dynamic probabilistic rough sets with incomplete data

Chuan Luo, Tianrui Li, Yiyu Yao

PII: S0020-0255(17)30810-1
DOI: 10.1016/j.ins.2017.06.040
Reference: INS 12962

To appear in: *Information Sciences*

Received date: 2 December 2015
Revised date: 12 June 2017
Accepted date: 26 June 2017

Please cite this article as: Chuan Luo, Tianrui Li, Yiyu Yao, Dynamic probabilistic rough sets with incomplete data, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.06.040

Dynamic probabilistic rough sets with incomplete data

Chuan Luo[a,b], Tianrui Li[a,*], Yiyu Yao[b]

[a]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b]Department of Computer Science, University of Regina, Regina, Saskatchewan, S4S 0A2, Canada

Abstract

Data in real-world applications are typically changing with time and are often incomplete. To address the challenge of processing such dynamic and incomplete data, we propose a model of dynamic probabilistic rough sets with incomplete data. We introduce incremental methods for estimating the conditional probability and present principles for updating probabilistic approximations when adding and removing objects, respectively. Based on the proposed updating strategies, algorithms are designed for dynamically updating probabilistic approximations with incomplete data. We report experimental evaluations of the efficiency and effectiveness of the proposed incremental algorithms for constructing probabilistic rough set approximations in terms of the size of data and updating ratio by comparing with a non-incremental algorithm. The results show that the new algorithms can effectively utilize the previously acquired knowledge, leading to significantly improved performance over a non-incremental algorithm.

Keywords: Incomplete data, incremental learning, rough sets, probabilistic approximations.

1. Introduction

Rough set theory provides effective methods for learning classification rules by using an information table with insufficient and incomplete information [16, 17, 21, 30]. It provides a ternary classification framework by approximating a target concept with three regions, namely, the positive, negative and boundary regions [41, 42]. There are two important classes of generalizations of rough set theory. One class is the probabilistic rough sets and the other class is rough sets with incomplete data.

In the Pawlak rough set model, the classification of data must be fully correct and certain, which is too restrictive to be practically useful in real-world applications [31]. By allowing certain acceptable level of classification errors, probabilistic approaches have been applied to rough set theory [40, 52]. According to whether or not there are missing data, information tables are classified into two categories, *i.e.*, complete and incomplete [7, 33]. Many proposals have been made in probabilistic generalizations of rough set theory with complete data. Wong et al. [36] proposed the notion of probabilistic approximate classification in rough set theory. Pawlak et al. [31] proposed 0.5-probabilistic rough set model by using 0.5 as a threshold to define probabilistic rough sets. Yao et al. [43] proposed a generalized probabilistic rough set model, called a decision-theoretic rough set model (DTRS), by using a pair of thresholds. Ziarko [53] introduced a variable precision rough set model based on a measure of misclassification. Greco et al. [9, 10] proposed a parameterized rough set model by using a pair of thresholds on a Bayesian confirmation measure, in addition to a pair thresholds on probability. Slezak et al. [34] presented a Bayesian rough set model, where the approximation regions are determined by using the prior probability. Herbert and Yao [2, 13] proposed a game-theoretic rough set model by integrating game theory and DTRS, in which an information theoretic approach is given for learning optimal thresholds. Yao and Zhou [44] proposed a naive Bayesian rough set model to estimate conditional probability with naive independence assumptions. Hu et al. [18] employed statistical techniques into rough sets, and proposed a new probabilistic rough set model with a probability distribution on the universe.

*Corresponding author.
Email addresses: `cluo@scu.edu.cn` (Chuan Luo), `trli@swjtu.edu.cn` (Tianrui Li), `yyao@cs.uregina.ca` (Yiyu Yao)