Knowledge-Based Systems 73 (2015) 81–96

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Incremental updating approximations in probabilistic rough sets under the variation of attributes

Dun Liu [a,*], Tianrui Li [b], Junbo Zhang [b]

[a] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China
[b] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 29 May 2014
Received in revised form 5 August 2014
Accepted 15 September 2014
Available online 2 October 2014

Keywords:
Rough sets theory
Probabilistic rough sets
Incremental learning
Updating approximations
Knowledge discovery

ABSTRACT

The attribute set in an information system evolves in time when new information arrives. Both lower and upper approximations of a concept will change dynamically when attributes vary. Inspired by the former incremental algorithm in Pawlak rough sets, this paper focuses on new strategies of dynamically updating approximations in probabilistic rough sets and investigates four propositions of updating approximations under probabilistic rough sets. Two incremental algorithms based on adding attributes and deleting attributes under probabilistic rough sets are proposed, respectively. The experiments on five data sets from UCI and a genome data with thousand attributes validate the feasibility of the proposed incremental approaches.

1. Introduction

With today's databases increase at an unprecedented rate, information systems evolve over time. Various approaches for updating knowledge incrementally are getting more and more popular. The algorithms or strategies for updating knowledge from information systems are vital in inductive machine learning and data mining.

An information system is composed by the objects, the attributes and the domain of attributes' values [30]. The literature for updating knowledge from information systems are mainly on three aspects, namely, variation of objects (instances) [2,7,19,22,23,29,31,43,46,57,58,60,62], variation of attributes (features) [4,9,20,21,26,32,59,64] and variation of attributes' values [5,6,8,24,25,27]. All these studies help decision makers to update knowledge with different viewpoints from different kinds of information systems.

By considering many databases with thousands of attributes (features), we face lots of massive challenges in real technical applications. The updating strategies and incremental algorithms under the variation of attributes are significantly affecting the knowledge updating, both in quantitative and qualitative aspects [21]. In this paper, we mainly focus on discussing scenarios of the variation of attributes.

In view of granular computing, the variation of attributes affects the granularity of the knowledge space. Yao pointed out that the two simple and commonly operators named "Zooming-in" and "Zooming-out" can describe the dynamic character while systems vary [51]. The Zooming-in operator refines the granules of the universe, like decomposing a granule into many granules; The Zooming-out operator coarsens the granules of the universe by omitting some details of the problem, such as combining many granules to form a new granule [51]. The granularity becomes refining when adding the attributes. On the contrary, the granularity becomes coarsening when deleting the attributes. Qian et al. proposed a multi-granulation rough set model and discussed several important measures in both complete information systems [40] and incomplete information systems [41]. Furthermore, Hu et al. unitized the granularity to the neighborhood rough set for heterogeneous feature subset selection. They further provided useful properties and prefect experimental results on variation of feature numbers [14,15].

Rough set theory (RST) is one special case of granular computing [34]. In RST, the variation of attributes, includes adding and deleting of attributes, affects reducts and approximations of a concept in information systems. For incremental attribute reduction approaches, Qian et al. investigated an accelerator for attribute reduction in positive approximation [42]. Hu et al. proposed an incremental attribute reduction based on elementary sets [13]. For incremental approaches of updating approximations in RST, Chan discussed an incremental approach for updating approximations of a concept when adding or deleting an attribute in a

* Corresponding author.
E-mail addresses: newton83@163.com (D. Liu), trli@swjtu.cn (T. Li), JunboZhang86@163.com (J. Zhang).

http://dx.doi.org/10.1016/j.knosys.2014.09.008
0950-7051/

complete information system by using the lower and upper boundaries [3,4]. Li et al. presented a method for updating approximations of a concept in an incomplete information system under the characteristic relation when the attribute set varies over time [21]. Cheng proposed an incremental model for fast computing the rough fuzzy approximations [9]. Li et al. investigated an approach for incremental updating approximations in dominance-based rough sets under the variation of the attribute sets [19]. Luo et al. discussed an approach for incremental updating approximations in set-valued ordered decision systems under the attribute generalization [32]. Zhang et al. proposed a rough sets based matrix approaches to compute the incremental updating approximations [58], and further considered a parallel method for computing rough set approximations for massive data analysis [60,61,63]. Davide summarized the classification of dynamics in rough sets with synchronic cases and diachronic cases. He further discussed dynamic in information systems, approximation spaces and coverings, which provided a basic framework on dynamic studies of rough sets [10,11].

As a generalized model of RST, probabilistic rough sets (PRS) allow a flexible approximation boundary region by using the threshold parameters with a better tolerance ability for inconsistent data [65,67]. The model of PRS generalizes the restrictive definition of the lower and upper approximations by allowing certain acceptable levels of errors. A pair of threshold parameters define the lower and upper approximations [52,53]. By considering the PRS leads to a new direction of research and applications on one hand and some confusions and inconsistencies on the other [53], this paper focuses on development of the incremental approach for updating approximations in PRS under the variation of attributes.

The rest of the paper is organized as follows: Section 2 provides basic concepts of Pawlak rough sets and PRS. The related propositions, strategies and algorithms for incremental learning knowledge in PRS are presented when attributes vary in Section 3. Section 4 illustrates the proposed approach with experiments under five data sets from UCI and a genome data with thousand attributes. The paper ends with conclusions and further research topics in Section 5.

2. Preliminaries

Basic concepts, notations and results of rough sets as well as their extensions are outlined in this section [1,16,18,33,34,36,38,39,45,47–50,52–55,65,66].

2.1. Pawlak rough sets

An approximation space $apr = (U, R)$ is defined by a universe U and a binary relation R. Let U be a finite and non-empty set and R be an binary relation on U. The pair $apr = (U, R)$ is defined as Pawlak approximation space when R is an equivalence relation. In Pawlak rough sets, the equivalence relation R induces a partition of U, denoted by $[x]_R$ or $[x]$. For a subset $X \subseteq U$, its lower and upper approximations are defined respectively by:

$$\begin{aligned} &\underline{R}(X) = \{x \in U | [x] \subseteq X\}; \\ &\overline{R}(X) = \{x \in U | [x] \cap X \neq \emptyset\}. \end{aligned} \tag{1}$$

Intuitively, these two approximations divide the universe U into three disjoint regions: the positive region $\mathrm{POS}_R(X)$, the negative region $\mathrm{NEG}_R(X)$ and the boundary region $\mathrm{BND}_R(X)$.

$$\begin{aligned} &\mathrm{POS}_R(X) = \underline{R}(X); \\ &\mathrm{BND}_R(X) = \overline{R}(X) - \underline{R}(X); \\ &\mathrm{NEG}_R(X) = U - \overline{R}(X). \end{aligned} \tag{2}$$

The positive region $\mathrm{POS}(X)$ is defined by the lower approximation, the negative region $\mathrm{NEG}(X)$ is defined by the complement of the upper approximation, and the boundary region $\mathrm{BND}(X)$ is defined by the difference between the upper and lower approximations. If $\underline{R}(X) = X = \overline{R}(X)$ holds, a subset X of U is definable in (U, R). Otherwise, X is indefinable in (U, R).

2.2. Probabilistic rough sets

To introduce the PRS, Pawlak and Skowron suggested using a rough membership function to redefine the two approximations [37]. The rough membership function μ_A is defined by:

$$\mu_A(x) = Pr(X|[x]) = \frac{|[x] \cap X|}{|[x]|}, \tag{3}$$

where $|\cdot|$ stands for the cardinal number of objects in sets. $Pr(X|[x])$ denotes the conditional probability of the classification. The degree of the overlap between an equivalence class $[x]$ and a set X is calculated as the conditional probability $Pr(X|[x])$ of the set given the equivalence class $[x]$.

From a rough membership function, the Pawlak approximations in Eq. (1) can be equivalently defined respectively as follows:

$$\begin{aligned} &\underline{R}(X) = \{x \in U | Pr(X|[x]) = 1\}; \\ &\overline{R}(X) = \{x \in U | Pr(X|[x]) > 0\}. \end{aligned} \tag{4}$$

In Eq. (4), two extreme values, 1 and 0, are utilized to define the two approximations by rough membership function. The classifications in the lower approximation must be absolutely consistent. However, the definition in Eq. (4) is too strict because the magnitude of the value $Pr(X|[x])$ is not taken into account [44,49,65]. A main result in PRS is parameterized probabilistic approximations, which is similar to the notion of α-cuts of fuzzy sets [53]. This can be done by replacing the values 1 and 0 in Eq. (5) by a pair of parameters α and β. The (α, β)-lower approximation and (α, β)-upper approximation are defined respectively as follows.

$$\begin{aligned} &\underline{R}_{(\alpha,\beta)}(X) = \{x \in U | Pr(X|[x]) \geqslant \alpha\}, \\ &\overline{R}_{(\alpha,\beta)}(X) = \{x \in U | Pr(X|[x]) > \beta\}. \end{aligned} \tag{5}$$

In (5), an equivalence class $[x]$ is a part of the lower approximation if the conditional probability $Pr(X|[x])$ is above α, and it is a part of the upper approximation if the conditional probability $Pr(X|[x])$ is above β. The (α, β)-probabilistic positive, boundary and negative regions can be defined by the (α, β)-probabilistic lower and upper approximations.

$$\begin{aligned} &\mathrm{POS}_{(\alpha,\beta)}(X) = \{x \in U | Pr(X|[x]) \geqslant \alpha\}, \\ &\mathrm{BND}_{(\alpha,\beta)}(X) = \{x \in U | \beta < Pr(X|[x]) < \alpha\}, \\ &\mathrm{NEG}_{(\alpha,\beta)}(X) = \{x \in U | Pr(X|[x]) \leqslant \beta\}. \end{aligned} \tag{6}$$

The parameters α and β allow certain acceptable levele of errors, and the PRS makes the process of decision making more reasonable. Specially, when we set $\alpha = \beta = 0.5$, it becomes the 0.5-probabilistic rough set model [35]; when we set $\alpha + \beta = 1$ and $\alpha > 0.5$, it becomes the symmetric variable precision rough set model [65]; when we set $0 \leqslant \beta < \alpha \leqslant 1$, it becomes the asymmetric variable precision rough set model [17]. In addition, the values of α and β can be automatically computed by using the bayesion minimum conditional risk criterion in decision-theoretic rough set models [28,49]. In our following discussions, we focus on investigating the incremental updating approximations approaches in PRS.