European Journal of Operational Research 251 (2016) 904–917

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Computational Intelligence and Information Management

Two Bayesian approaches to rough sets

Yiyu Yao[a], Bing Zhou[b,*]

[a] Department of Computer Science, University of Regina, Regina, Saskatchewan S4S 0A2, Canada
[b] Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, USA

ARTICLE INFO

Article history:
Received 18 June 2014
Accepted 31 August 2015
Available online 7 September 2015

Keywords:
Bayesian classification rough sets
Bayesian confirmation rough sets
Confirmation-theoretic rough sets
Decision-theoretic rough sets
Probabilistic rough sets

ABSTRACT

Bayesian inference and probabilistic rough sets (PRSs) provide two methods for data analysis. Both of them use probabilities to express uncertainties and knowledge in data and to make inference about data. Many proposals have been made to combine Bayesian inference and rough sets. The main objective of this paper is to present a unified framework that enables us (a) to review and classify Bayesian approaches to rough sets, (b) to give proper perspectives of existing studies, and (c) to examine basic ingredients and fundamental issues of Bayesian approaches to rough sets. By reviewing existing studies, we identify two classes of Bayesian approaches to PRSsand three fundamental issues. One class is interpreted as Bayesian classification rough sets, which is built from decision-theoretic rough set (DTRS) models proposed by Yao, Wong and Lingras. The other class is interpreted as Bayesian confirmation rough sets, which is built from parameterized rough set models proposed by Greco, Matarazzo and Słowiński. Although the two classes share many similarities in terms of making use of Bayes' theorem and a pair of thresholds to produce three regions, their semantic interpretations and, hence, intended applications are different. The three fundamental issues are the computation and interpretation of thresholds, the estimation of required conditional probabilities, and the application of derived three regions. DTRS models provide an interpretation and a method for computing a pair of thresholds according to Bayesian decision theory. Naive Bayesian rough set models give a practical technique for estimating probability based on Bayes' theorem and inference. Finally, a theory of three-way decisions offers a tool for building ternary classifiers. The main contribution of the paper lies in weaving together existing results into a coherent study of Bayesian approaches to rough sets, rather than introducing new specific results.

1. Introduction

In Bayesian data analysis, a prior probability is used to capture our belief about an event or a hypothesis before observing the evidence or data, and the Bayes' theorem is used to update the prior probability into a posterior probability through a likelihood when evidence becomes available. Bayesian methods have been widely applied in many fields for decision making and classification under uncertainty (Liu, Hua, & Lim, 2015). Pawlak rough set theory provides another approach to data analysis (Pawlak, 1982, 1991). Rough set analysis identifies decision rules and dependencies from data for decision making and classification. In generalized probabilistic approaches to rough sets, rules are typically studied and characterized in probabilistic terms (Grzymala-Busse, Marepally, & Yao, 2010; Pawlak, 1999; Tsumoto, 2002; Yao, 2003, 2008). There are close connections between Bayesian inference and rough set theory. A reviewer of this paper concisely summarized, "The relationship of rough sets with probability theory has been a matter of debate ever since the rough sets were first proposed in 1980's. Over last three decades a number of researchers have made significant contributions to the study of relationship between the two theories. These contributions have not only helped us understand the rough sets better, but they have also provided useful extension of the rough set theory and in some cases created new ways of reasoning that combine concepts from rough sets and probability theory."

The first probabilistic rough set (PRS) model, called the 0.5-probabilistic rough set model (Yao, 2007a), was proposed by Wong and Ziarko (1985) and Pawlak, Wong, and Ziarko (1988). A threshold of 0.5 on probability is used to define probabilistic lower and upper approximations, or equivalently three probabilistic regions, of a set. The threshold 0.5 can be intuitively interpreted based on the notion of the majority rule. Wong and Ziarko (1986b) late generalized the model by using a pair of thresholds (α, 0.5). Based on Bayesian decision theory, Yao (2007a), Yao and Wong (1992), and Yao, Wong, Lingras, Ras, Zemankova, and Emrich (1990) proposed a generalized probabilistic model, called a decision-theoretic rough set (DTRS) model, by considering a pair of thresholds (α, β) on probabilities for defining probabilistic approximations. The pair of

* Corresponding author. Tel.: +001 9364458637.
E-mail addresses: yyao@cs.uregina.ca (Y. Yao), zhou@shsu.edu (B. Zhou).

http://dx.doi.org/10.1016/j.ejor.2015.08.053
0377-2217/

thresholds can be systematically calculated based on the well established Bayesian decision theory, and interpreted in terms of more practically operable notions such as cost, risk, benefit etc. Herbert and Yao (2008) integrated game theory and DTRS model to introduce a new PRS model known as a game-theoretic rough set (GTRS) model.

Based on the notion of graded set inclusion, Ziarko (1993) introduced a variable precision rough set (VPRS) model by using a pair of thresholds on a set-inclusion function. The derived approximations are equivalent to a special case of the DTRS model. The main results of the model were later explicitly re-expressed in terms of thresholds on probability instead of a set-inclusion function (Ziarko, 2002).

Ślęzak and Ziarko (2002, 2005) introduced a Bayesian rough set (BRS) model. A prior probability is used as a threshold for defining three regions. They also suggested to compare two likelihoods directly when neither posterior probability nor prior probability is derivable from data. Ślęzak (2005) further drew a natural correspondence between the fundamental notions of rough sets and statistics. The set to be approximated corresponds to a hypothesis and an equivalence class to a piece of evidence; the three probabilistic regions correspond to the cases that the hypothesis is verified positively, negatively, or undecided based on the evidence. Based on such a correspondence, Ślęzak introduced a rough Bayesian model (Ślęzak, 2005), in which probabilistic approximations are defined based on a pair of thresholds on the ratio of the prior and posterior probabilities.

Greco, Pawlak, and Słowiński (2004a, 2004b) first introduced Bayesian confirmation measures into rough set theory to study decision rules and decision algorithms. Greco, Matarazzo, and Słowiński (2005; 2008) argued that it may be insufficient to consider only probability values when formulating a PRS model. As a result, they introduced a parameterized rough set model by considering a pair of thresholds on a Bayesian confirmation measure, in addition to a pair of thresholds on probability. They presented and analyzed systematically Bayesian confirmation measures in constructing PRSs. Moreover, they explicitly showed that Pawlak rough sets, VPRSs (Ziarko, 1993), BRSs (Ślęzak & Ziarko, 2002, 2005) and rough Bayesian sets (Ślęzak, 2005) are special cases of their parameterized model. A problem to be solved is how to systematically determine a pair of thresholds on a Bayesian confirmation measure.

An important problem of PRSs is the estimation of the conditional probability. The rough membership function (Pawlak, Skowron, Yager, Fedrizzi, & Kacprzyk, 1994) is a simple way to do it, but of limited value due to the requirement of a large-sized sample. Kotlowski, Dembczynski, Greco, and Słowiński (2008) and Kotlowski and Słowiński (2013) suggested a statistical model in which probabilities are estimated based on the maximization of a likelihood function. Yao and Zhou (2010) introduced a naive Bayesian rough set (NBRS) model by slightly modifying results from the standard naive Bayesian model for classification. An equivalence class is described by a vector of attribute values. Through an application of Bayes' theorem, the estimation of the posterior probability is turned into the estimation of the likelihood based on the naive probabilistic independence assumption of attributes.

Motivated by rough set classification with three regions, Yao (2009, 2010, 2011) introduced the notion of three-way decisions. Specifically, similar to the concepts of accepting a hypothesis, rejecting a hypothesis, or further testing in statistical testing (Wald, 1945), the three regions can induce positive rules for accepting an object to be an instance of a concept, negative rules for rejecting, and boundary rules for deferring a definite decision.

In a series of papers, Pawlak (Greco et al., 2004b; Pawlak, 1999; 2002) advocated another research direction in studying connections between Bayesian methods and rough set approaches. He reinterpreted some of the results of Bayesian data analysis in the context of rough sets. Every decision rule is associated with two conditional probabilities, called accuracy and coverage (Tsumoto, 2002). While the accuracy corresponds to posterior probability, the coverage corresponds to the likelihood in Bayesian methods. In other words, Pawlak used Bayes' theorem to explain the probabilistic relationship between conditions and decisions in decision rules.

Each of these studies focuses on a specific perspective on PRSs. They are complementary to each other and, working together, they provide the main ingredients of a general framework for studying Bayesian probabilistic approaches to rough sets. Although such a framework emerges from the vast amount of studies, the specifics of the framework have not been fully examined, discussed and analyzed. The main objective of this paper is therefore to present one such general framework.

There are two parts of our framework. The first part is a classification of existing PRSs into two categories, namely, Bayesian classification rough sets and Bayesian confirmation rough sets, as shown in Fig. 1. There are connections and differences between the parameterized model (Greco et al., 2005, 2008) and our framework. While the parameterize model is a single model with two components, we explicitly divide them into two classes of models, with each class intended for a different type of applications. The second part is the identification of three fundamental issues: (a) determination of thresholds, (b) estimation of conditional probability, and (c) application of three regions. The framework enables us to have a comprehensive understanding of the evolution of PRSs, perspectives of different PRS models, differences between these models and their intended applications.

It should be pointed out that specific results in this paper are not entirely new and have been examined and discussed in many other papers. The main contribution of the paper is to integrate these results coherently into a complete whole within a common framework and to present them, first time, in a single paper. To achieve our goal, the rest of the paper is organized as follows. Section 2 briefly summarizes the main results from existing studies of Bayesian approaches to rough sets and identifies three fundamental issues. We give reasons for our preference to a categorization consisting of two classes of models, namely, Bayesian classification rough sets and Bayesian confirmation rough sets. Following the formulation of parameterized rough sets given by Greco et al. (2005; 2008), Section 3 presents a formulation of Bayesian confirmation rough sets, or confirmation-theoretic rough sets, and discusses their differences from DTRS models. Section 4 presents a different type of formulation of Bayesian classification rough sets based on the result of the DTRS models. This section focuses on the issue of interpreting and determining the required thresholds based on Bayesian decision theory. Section 5 examines the problem of estimating probabilities based on Bayes' theorem. After discussing the general form of a BRS model, an NBRS model is presented. A special case of the NBRS model, called a binary probabilistic independence rough set (BPIRS) model, is derived. Section 6 interprets the three regions of Bayesian classification models based on the notion of three-way decisions. We demonstrate applications of Bayesian classification rough sets for building ternary classifiers. Section 7 provides concluding remarks.

2. Models and basic issues in Bayesian approaches to rough sets

In this section, we briefly summarize fundamental results of Pawlak rough sets and PRSs. We identify two basic classes of models, namely, Bayesian classification rough sets and Bayesian confirmation rough sets. We also point out three basic issues in studying Bayesian approaches to rough sets. To emphasize the semantic interpretation of PRSs with three-way decisions (Yao, 2009; 2010), our formulation directly uses three pair-wise disjoint positive, boundary, and negative regions, instead of a pair of lower and upper approximations.