Pattern Recognition 44 (2011) 1658–1670

Contents lists available at ScienceDirect

Pattern Recognition

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

An efficient accelerator for attribute reduction from incomplete data in rough set framework ☆

Yuhua Qian [a,b], Jiye Liang [a,*], Witold Pedrycz [c], Chuangyin Dang [b]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, Shanxi, China
[b] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

ARTICLE INFO

Article history:
Received 7 December 2009
Received in revised form
16 February 2011
Accepted 18 February 2011
Available online 24 February 2011

Keywords:
Feature selection
Rough set theory
Incomplete information systems
Positive approximation
Granular computing

ABSTRACT

Feature selection (attribute reduction) from large-scale incomplete data is a challenging problem in areas such as pattern recognition, machine learning and data mining. In rough set theory, feature selection from incomplete data aims to retain the discriminatory power of original features. To address this issue, many feature selection algorithms have been proposed, however, these algorithms are often computationally time-consuming. To overcome this shortcoming, we introduce in this paper a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. As an application of the proposed accelerator, a general feature selection algorithm is designed. By integrating the accelerator into a heuristic algorithm, we obtain several modified representative heuristic feature selection algorithms in rough set theory. Experiments show that these modified algorithms outperform their original counterparts. It is worth noting that the performance of the modified algorithms becomes more visible when dealing with larger data sets.

1. Introduction

Feature selection, also called attribute reduction, is a common problem in pattern recognition, data mining and machine learning. In recent years, both the number and dimensionality of items in data sets have grown dramatically. For examples, tens, hundreds, and even thousands of attributes are stored in many real-world application databases [1–3]. It is well known that attributes irrelevant to recognition tasks may deteriorate the performance of learning algorithms [4]. In other words, storing and processing all attributes (both relevant and irrelevant) could be computationally very expensive and impractical. To address this issue, as pointed out in [5], some attributes can be omitted, which will not seriously affect the resulting classification (recognition) accuracy. Therefore, the omission of some attributes could be not only tolerable but also even desirable relative to the computational costs involved [6].

In feature selection, there are two general strategies, namely wrappers [7] and filters. The former employs a learning algorithm to evaluate the selected attribute subsets, and the latter selects attributes according to some significance measure such as information gain [8], consistency [9], distance [10], dependency [11], and others. These measures can be divided into two main categories: distance-based measures and consistency-based measures [5]. Attribute reduction in rough set theory offers a systematic theoretic framework for consistency-based feature selection, which does not attempt to maximize the class separability but rather attempts to retain the discerning ability of original features for the objects from the universe [12,13].

Generally speaking, one always needs to handle two types of data, viz, those that assume numerical values and symbolic values. For numerical values, there are two types of approaches. One relies on fuzzy rough set theory, and the other is concerned with the discretization of numerical attributes. In order to deal with hybrid attributes, several approaches have been developed in the literature [12,14–21]. In classical rough set theory, the attribute reduction algorithm takes all attributes to be symbolic values. After preprocessing original data, one can use classical rough set theory to select a subset of features that is most suitable for a given recognition problem.

Feature selection based on rough set theory starts from a data table, which is also called an information system and contains data about objects of interest that are characterized by a finite set of attributes. According to whether or not there are missing data (null values), information systems are classified into two

☆ This is an extended version of the paper presented at the Eighth IEEE International Conference on Machine Learning and Cybernetics, Baoding, 2009, China.

* Corresponding author. Tel./fax: +86 0351 7018176.
E-mail addresses: jinchengqyh@126.com (Y. Qian), ljy@sxu.edu.cn (J. Liang), pedrycz@ee.ualberta.ca (W. Pedrycz), mecdang@cityu.edu.hk (C. Dang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.020

categories: complete and incomplete. In general, by an incomplete information system, we mean a system with missing data (null values) [22,23]. For an incomplete information system, if condition attributes and decision attributes are distinguished from each other, then it is called an incomplete decision table. Feature selection from incomplete data usually starts from incomplete decision tables.

In the last two decades, many techniques for attribute reduction have been developed in rough set theory [25–29]. Especially £ in order to obtain all attribute reducts of a given data set, Skowron proposed a discernibility matrix method [30]. However, these feature selection algorithms are usually time-consuming to process large-scale data. Aiming at efficient feature selection, many heuristic attribute reduction algorithms have been developed in rough set theory, cf. [5,19,31–37]. Each of these algorithms preserves a particular property of a given information system. To accomplish attribute reduction from incomplete decision tables, similar to the discernibility matrix proposed by Skowron, Kryszkiewicz gave a generalized discernibility matrix to obtain all attribute reducts of an incomplete decision table [24]. To efficiently obtain an attribute reduct, several heuristic attribution reduction approaches have been presented [38–41]. For convenience of our further development, we review several representative heuristic attribute reduction algorithms in the context of incomplete data here. Applying the idea of positive-region reduction, Yang and Shu proposed a heuristic feature selection algorithm in incomplete decision tables, which keeps the positive region of target decision unchanged [39]. Liang et al. defined new information entropy to measure the uncertainty of an incomplete information system [23] and applied the corresponding conditional entropy to reduce redundant features [38]. Qian and Liang [40] presented the combination entropy for measuring the uncertainty of an incomplete information system and used its conditional entropy to obtain a feature subset. As Shannon's information entropy was introduced to search reducts in classical rough set model [34], an extension of its conditional entropy also can be used to calculate a relative attribute reduct of an incomplete decision table. However, the above algorithms are still computationally time-consuming to deal with large-scale data sets.

In this study, we will not consider how to discretize numerical attributes and construct a heuristic function for feature selection. Our objective is how to improve computational efficiency of a heuristic attribute reduction algorithm in the context of incomplete data. A brief version of this work has been published in the literature [42]. In this extended version, we propose a new rough set framework, which is called positive approximation in incomplete information systems. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure of an incomplete rough set using a granulation order. Based on the positive approximation, we develop a common accelerator for improving computational efficiency of a heuristic feature selection, which provides a vehicle of making rough set-based feature selection algorithms faster. By incorporating the accelerator into each of the above representative heuristic attribute reduction algorithms, we obtain their modified versions. Numerical experiments show that each of the modified algorithms can choose the same feature subset as that of the corresponding original algorithm while greatly reducing computational time. Furthermore, we would like to stress that the improvement becomes more visible when the data sets become larger.

The rest of the paper is organized as follows. Some basic concepts are briefly reviewed in Section 2, which include incomplete information systems, incomplete rough set model, incomplete variable precision rough set model and partial relations. In Section 3, we establish a positive approximation framework in incomplete information systems and investigate some of its main properties. In Section 4, by analyzing the rank preservation of several representative significance measures of attributes, a general algorithm based on the positive approximation is first introduced, and a series of experimental studies are then conducted, which focus on comparison of computational efficiency and stability of the selected attributes. Finally, Section 5 concludes the paper with some remarks and discussions.

2. Preliminaries

In this section, we will review several basic concepts such as incomplete information systems, tolerance relation and partial relation. Throughout this paper, we suppose that the universe U is a finite non-empty set.

An information system is a pair $S=(U,A)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, and for every $a \in A$, there is a mapping a, $a: U \to V_a$, where V_a is called the value set of a.

It may happen that some of attribute values for an object are missing. To distinguish such a situation from the other, a so-called null value, denoted by $*$, is usually assigned to those attributes. If V_a contains a null value for at least one attribute $a \in A$, then S is called an incomplete information system; otherwise it is a complete one [24].

Let $S=(U,A)$ be an information system and $P \subseteq A$ an attribute set. We define a binary relation on U as follows:

$$SIM(P)=\{(u,v) \in U \times U \mid \forall a \in P,\ a(u)=a(v) \quad \text{or } a(u)=* \quad \text{or } a(v)=*\}.$$

In fact, $SIM(P)$ is a tolerance relation on U. The concept of a tolerance relation has a wide variety of applications in classification. It can be easily shown that $SIM(P)=\bigcap_{a \in P} SIM(\{a\})$. Let $S_P(u)$ denote the set $\{v \in U \mid (u,v) \in SIM(P)\}$. $S_P(u)$ is the maximal set of objects which are possibly indistinguishable by P with u. Let $U/SIM(P)$ denote the family sets $\{S_P(u) \mid u \in U\}$, which is the classification or the knowledge induced by P. A member $S_P(u)$ from $U/SIM(P)$ will be called a tolerance class or an information granule. It should be noticed that the tolerance classes in $U/SIM(P)$ do not yield a partition of U in general. They form a cover of U, i.e., $S_P(u) \neq \emptyset$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u)=U$.

Let $S=(U,A)$ be an incomplete information system, X a subset of U, and $P \subseteq A$ an attribute set. In the rough set model, based on the tolerance relation [24,43], X can be characterized by $\overline{SIM(P)}X$ and $\underline{SIM(P)}X$, where

$$\begin{cases} \underline{SIM(P)}X = \cup\{Y \in U/SIM(P) \mid Y \subseteq X\}, \\ \overline{SIM(P)}X = \cup\{Y \in U/SIM(P) \mid Y \cap X \neq \emptyset\}. \end{cases}$$

There are two kinds of attributes for a classification problem. Each of them can be characterized by a decision table $S=(U,C \cup D)$ with $C \cap D=\emptyset$, where an element of C is called a condition attribute, C is called a condition attribute set, an element of D is called a decision attribute, and D is called a decision attribute set. Assume the objects are partitioned into r mutually exclusive crisp subsets $\{X_1,X_2,\ldots,X_r\}$ by the decision attribute set D. Given any subset $P \subseteq C$ and the tolerance relation $SIM(P)$ induced by P, one can then define the lower and upper approximations of the decision attribute set D as

$$\begin{cases} \underline{SIM(P)}D = \{\underline{SIM(P)}X_1,\underline{SIM(P)}X_2,\ldots,\underline{SIM(P)}X_r\}, \\ \overline{SIM(P)}D = \{\overline{SIM(P)}X_1,\overline{SIM(P)}X_2,\ldots,\overline{SIM(P)}X_r\}. \end{cases}$$

Let $POS_P(D)=\bigcup_{i=1}^{r} \underline{SIM(P)}X_i$, which is called the positive region of D with respect to the condition attribute set P.