Pattern Recognition Letters 36 (2014) 81–88

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A fast feature selection approach based on rough set boundary regions

Zhengcai Lu [a,*], Zheng Qin [a,b], Yongqiang Zhang [c], Jun Fang [a]

[a] *Department of Computer Science & Technology, Tsinghua University, Beijing 100084, PR China*
[b] *School of Software, Tsinghua University, Beijing 100084, PR China*
[c] *Institute of Tracking & Telecommunication Technology, Beijing 100094, PR China*

ARTICLE INFO

Article history:
Received 16 September 2012
Available online 29 September 2013

Communicated by A. Heyden

Keywords:
Feature selection
Rough set theory
Boundary region
Tolerance relation
Incomplete decision systems

ABSTRACT

Dataset dimensionality is one of the primary impediments to data analysis in areas such as pattern recognition, data mining, and decision support. A feature subset that possesses the same classification power as that of the whole feature set is expected to be found prior to performing a classification task. For this purpose, many rough set algorithms for feature selection have been developed and applied to incomplete decision systems. When they address large data, however, their undesirable efficiencies could be intolerable. This paper proposes a boundary region-based feature selection algorithm (BRFS), which has the ability to efficiently find a feature subset from a large incomplete decision system. BRFS captures an inconsistent block family to construct a rough set boundary region and designs a positive stepwise mechanism for the construction of boundary regions with respect to multiple attribute subsets, making the acquisition of boundary regions highly efficient. The boundary regions are used to build significance measures as heuristics to determine the optimal search path and establish an evaluation criterion for rules to identify feature subsets. These arrangements make BRFS capable of locating a reduct more efficiently than other available algorithms; this finding is supported by experimental results.

1. Introduction

We are currently overwhelmed with large-scale datasets from which pattern induction with machine learning tools results in very poor performance. For a specific task in data analysis, however, it can be observed that not all features are always necessary; in fact, some of them are actually superfluous. These irrelevant features can deplete the storage space, deteriorate the computational performance, and even decrease the generalization power of the induced patterns. It is, thus, desirable to search for a feature subset that has the same predictive capability as that of the original feature set. Such a process is called feature selection or attribute reduction.

Rough set theory (RST) (Pawlak, 1982) is a powerful tool for data analysis in areas such as pattern recognition, data mining, and decision support. RST has been widely used for feature selection because it is completely data-driven and does not require any auxiliary information. Based on the classical rough set model (CRSM), feature selection has been the subject of extensive research, such as in the indiscernibility matrix technique (Ye and Chen, 2002; Yang, 2007a; Yao and Zhao, 2009), the discernibility and indiscernibility technique (Zhao et al., 2007; Qian et al., 2011b), the positive region technique (Liu et al., 2003; Ge et al., 2009; Liu et al., 2009), and the information entropy technique (Wang et al., 2002; Yang, 2007b). These studies have offered interesting insights into feature selection and have been successfully applied to complete decision systems.

Regrettably, CRSM-based techniques are invalid in the context of incomplete decision systems. This defect arises from the fact that CRSM is inapplicable in the context of incomplete data. Concomitant solutions to the problem can be classified into two categories: indirect and direct. The indirect solutions transform incomplete systems into complete systems prior to the application of CRSM, such as deleting objects that carry missing values (Chmielewski et al., 1993), filling in null values with certain values (Li and Wu, 2009; Zhao et al., 2011), and assigning all of the possible values to an unknown value (Grzymala-Busse, 1991; Grzymala-Busse and Fu, 2000). However, as noted in the literature (Meng and Shi, 2009), preprocessing technologies that are employed by indirect solutions can change the information of the original systems to some degree and cause uncertainty in the induced patterns.

The direct solutions are capable of preserving the original structure of the information system as well as making feature selection more effective and reliable by handling incomplete data with extensions to CRSM, such as by using a tolerance rough set model (TRSM) (Kryszkiewicz, 1998a), an unsymmetrical tolerance rough

* Corresponding author. Tel.: +86 010 62795399.
E-mail addresses: luzc09@mails.tsinghua.edu.cn (Z. Lu), qingzh@mail.tsinghua.edu.cn (Z. Qin), zyqjianxia@yahoo.com.cn (Y. Zhang), fangjun06@mails.tsinghua.edu.cn (J. Fang).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.012

set model (UTRSM) (Stefanowski and Tsoukias, 1999), a quantitative tolerance rough set model (QTRSM) (Stefanowski and Tsoukias, 2001), or a limited tolerance rough set model (LTRSM) (Wang, 2002). Among these options, TRSM, which substitutes a tolerance relation for an equivalence relation, is the most prevalent and extensively used to perform feature selection from incomplete data (Kryszkiewicz, 1998b; Liang and Xu, 2002; Leung and Li, 2003; Huang et al., 2004; Huang et al., 2005; Yang and Shu, 2006; Li et al., 2007; Wu, 2008; Meng and Shi, 2009; Qian et al., 2010; Zhang et al., 2010; Qian et al., 2011a; Sun et al., 2012).

One of the pioneering approaches employs discernibility functions, which are constructed by pairwise comparisons of objects (Kryszkiewicz, 1998b). Feature subsets can be derived from prime implicants in simplified discernibility functions. A later study (Leung and Li, 2003) presented maximal consistent blocks as units for creating discernibility functions, making the simplification of discernibility functions more efficient. Another study (Qian et al., 2010) utilized the maximal consistent block technique to construct discernibility matrices for lower/upper approximation reducts, which costs less memory and less time. Although these algorithms are intuitive, concise and sufficient, they expose serious deficiencies when addressing large data: a large number of items that occur in discernibility functions could worsen the efficiency of simplification dramatically, and a large number of duplicate elements that appear in discernibility matrices could result in massive cost to both the storage space and the computational time. Consequently, this approach requires a time complexity of no less than $O(|C|^2|U|^2)$ to complete the feature selection tasks.

There are some other typical feature selection techniques in the context of incomplete data. Yang and Shu (2006) proposed a positive region-based feature selection algorithm that considers a positive region as an evaluation criterion to identify the feature subsets. Following this idea, Meng and Shi (2009) improved the algorithm by employing decomposition techniques to compute the positive region. From the perspective of information, Huang et al. (2005) introduced information entropy and conditional information entropy as heuristic information and proposed an information entropy-based feature selection algorithm for incomplete decision tables. Zhang et al. (2010) designed a new method of computing information entropy to make the algorithm more efficient. Sun et al. (2012) utilized rough entropy-based uncertainty measures to evaluate the roughness and accuracy of the knowledge, and then, they constructed a heuristic search algorithm that has a lower computational complexity for feature selection in context of incomplete decision systems. Unfortunately, the time complexities of these algorithms are no less than $O(|C|^2|U|\log|U|)$.

One can observe that the existing approaches to feature selection from incomplete decision systems have time complexities of no less than $O(|C|^2|U|\log|U|)$; thus, these approaches are computationally costly and inefficient for large amounts of data. An efficient and feasible feature selection approach is truly desirable. This paper concentrates on providing such a solution.

It has been demonstrated that almost all of the feature selection algorithms based on TRSM have focused on the low approximation of a rough set, and little work has focused on the boundary region of a rough set (Inuiguchi and Tsurumi, 2006; Parthalain and Shen, 2009). In fact, the boundary region defined by TRSM in incomplete decision systems is so informative that we are inspired to explore a fast feature selection algorithm based on boundary regions.

Our main contributions are as follows. First, we propose a new approach to construct a boundary region, which bypasses the lower and upper approximations. Second, a positive stepwise mechanism is designed to lead efficient computation to boundary regions with respect to multiple attribute subsets. Third, a boundary region is applied to build a significance measure as well as an evaluation criterion. Finally, a boundary region-based feature selection algorithm is proposed for incomplete decision systems and achieves a time complexity of no more than $O(|C|^2|U|)$. This approach is capable of locating a reduct more efficiently than the other available algorithms, which is a finding that is also supported by experimental results.

The remainder of this paper is organized as follows. In Section 2, we review some basic concepts that are related to incomplete decision systems and the tolerance rough set model. Section 3 explores a new approach to the boundary region as well as an efficient mechanism for boundary regions with respect to multiple attribute subsets. In Section 4, a feature selection algorithm using a boundary region technique is built. Some experiments are provided to validate the effectiveness of the proposed algorithm in Section 5. Finally, we give a concise conclusion and prospectives for further work in Section 6.

2. Preliminaries

A decision system is a 2-tuple $S = (U, C \cup D)$, where U, called the universe of discourse, is a non-empty finite set of objects, C is a condition attribute set, and D is a decision attribute set. For any $a \in C \cup D$, there is a mapping f, $f: U \to V_a$, where V_a is the value domain of a. If there are missing values (usually denoted by $*$) in the value domain of C, we say that the decision system is incomplete. An incomplete decision system is usually formulated by a data table, where the columns are referred to as attributes and the rows are referred to as objects of interest.

Let $S = (U, C \cup D)$ be an incomplete decision system. For any subset $P \subseteq C, P$ determines a binary relation, denoted by $SIM(P)$, which is defined as

$$SIM(P) = \{(u, v) \in U \times U | \forall\ a \in P, f(u,a) = f(v,a) \text{ or } f(u,a) = * \text{ or } f(v,a) = *\}. \tag{1}$$

It is easily known that $SIM(P)$ is reflexive, symmetric, and intransitive; as a result, it is a tolerance relation. For any object $u \in U, SIM(P)$ determines the tolerance class of u, which is denoted by $S_P(u)$. $S_P(u)$ describes the maximal set of objects that are possibly indistinguishable to u with respect to P. In other words, $S_P(u) = \{v \in U | (u, v) \in SIM(P)\}$. Clearly, the tolerance classes of all of the objects from U constitute a cover of U such that $\bigcup_{u \in U} S_P(u) = U$.

Consider a partition $\pi_D = \{D_i | i = 1, 2, \ldots, j\}$ of U that is determined by D. D_i is called a decision class and is approximated by a pair of precise concepts that are known as lower and upper approximations. The dual approximations of D_i are defined, respectively, as

$$\underline{P}(D_i) = \{u \in U | S_P(u) \subseteq D_i\}, \quad \bar{P}(D_i) = \{u \in U | S_P(u) \cap D_i \neq \emptyset\}. \tag{2}$$

$\underline{P}(D_i)$ is the maximal P-definable set that is contained in D_i, whereas $\bar{P}(D_i)$ is the minimal P-definable set that contains D_i. If $\underline{P}(D_i) = \bar{P}(D_i)$, then D_i is a P-exact set; otherwise, it is a P-rough set.

By the dual approximations, the universe of the decision system is partitioned into two mutually exclusive crisp regions: the positive region and the boundary region, which are defined, respectively, as

$$POS_P(\pi_D) = \bigcup_{i=1}^{j} \underline{P}(D_i), \tag{3}$$

$$BND_P(\pi_D) = \bigcup_{i=1}^{j} (\bar{P}(D_i) - \underline{P}(D_i)). \tag{4}$$