Pattern Recognition 42 (2009) 655 – 667

Contents lists available at ScienceDirect

Pattern Recognition

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

Exploring the boundary region of tolerance rough sets for feature selection

Neil Mac Parthaláin, Qiang Shen*

Department of Computer Science, Aberystwyth University, Wales, UK

ARTICLE INFO

Article history:
Received 29 November 2007
Received in revised form 7 August 2008
Accepted 13 August 2008

Keywords:
Feature selection
Attribute reduction
Rough sets
Classification

ABSTRACT

Of all of the challenges which face the effective application of computational intelligence technologies for pattern recognition, dataset dimensionality is undoubtedly one of the primary impediments. In order for pattern classifiers to be efficient, a dimensionality reduction stage is usually performed prior to classification. Much use has been made of rough set theory for this purpose as it is completely data-driven and no other information is required; most other methods require some additional knowledge. However, traditional rough set-based methods in the literature are restricted to the requirement that all data must be discrete. It is therefore not possible to consider real-valued or noisy data. This is usually addressed by employing a discretisation method, which can result in information loss. This paper proposes a new approach based on the tolerance rough set model, which has the ability to deal with real-valued data whilst simultaneously retaining dataset semantics. More significantly, this paper describes the underlying mechanism for this new approach to utilise the information contained within the boundary region or region of uncertainty. The use of this information can result in the discovery of more compact feature subsets and improved classification accuracy. These results are supported by an experimental evaluation which compares the proposed approach with a number of existing feature selection techniques.

1. Introduction

Quite often, at the point of data collection every single aspect of a domain may be recorded such that complete representation can be achieved. The problems associated with such large dimensionality, however, mean that any attempt to use machine learning tools to extract knowledge, results in very poor performance. Feature selection (FS) [1–7] is a process which attempts to select features which are information-rich and also retain the original meaning of the features following reduction. It is not surprising therefore that FS has been applied to problems which have very large dimensionality (> 10 000) [8].

Problems of such scale are outside the scope of most learning algorithms, and in cases where they are not, the learning algorithm will often find patterns that are spurious and invalid. As mentioned previously, it may be expected that the inclusion of an increasing number of features would also improve the likelihood of the ability to distinguish between classes. This may not be the case, however, if the training data size does not also increase significantly with the addition of each feature. Most learning approaches utilise a reduction step to overcome such problems when dealing with high dimensionality.

Rough set theory (RST) [9] is an approach that can be used for dimensionality reduction, whilst simultaneously preserving the semantics of the features [10]. Also, as RST operates only on the data and does not require any thresholding information, it is completely data-driven. Other useful approaches may also be employed for dimensionality reduction and FS such as Refs. [2,5,7,11], unlike RST, however, these approaches require additional information or transform the data. The main disadvantage of RST is its inability to deal with real-valued data. In order to tackle this problem, methods of discretising the data were employed prior to the application of RST. The use of such methods can result in information loss, however, and a number of extensions to RST have emerged [12–14] which have attempted to address this inability to operate on real-valued domains. Perhaps the most significant of these is the tolerance rough set model (TRSM) [13]. TRSM has the ability to operate effectively on real-valued (and crisp) data, thus minimising any information loss.

This paper presents a new method for FS which is based on the TRSM. It employs a distance metric to examine the uncertain information contained in the boundary region of tolerance rough sets, and uses this information to guide the FS process. This uncertain information is normally ignored in the traditional RST and TRSM approaches to FS which can result in information loss. The remainder of this paper is structured as follows. Section 2 introduces the theoretical background to RST and TRSM and their application to FS. Section 3 presents the new distance metric-assisted tolerance rough set selection method with a worked example to demonstrate the

* Corresponding author. Tel.: +44 1970 621825; fax: +44 1970 628536.
E-mail address: qqs@aber.ac.uk (Q. Shen).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.029

approach fully. All experimental evaluation and results for both approaches are presented in Section 4, as well as a comparison with the principal component analysis (PCA) dimensionality reduction technique [15], and also four additional FS techniques correlation-based feature selection (CFS) [16], consistency-based FS [17], ReliefF [18], and a wrapper FS approach which employs J48 [19] as an evaluation metric. The paper is then concluded with a brief discussion of future work in Section 5.

2. Background

Although the principal focus of this paper lies in the examination of the information contained in the boundary region of tolerance rough sets, an in-depth view of both the RST and TRSM methodologies is necessary in order to demonstrate the motivation for the investigation of the information in the boundary region.

RST [9] is an extension of conventional set theory which supports approximations in decision making. A rough set is the approximation of a vague concept by a pair of precise concepts which are known as upper and lower approximations. These concepts are illustrated in Fig. 1. The lower approximation is a definition of the domain objects which are known with absolute certainty to belong the concept of interest (set X), whilst the upper approximation is the set of those objects which possibly belong to the concept of interest. The boundary region or region of uncertainty is the difference between the upper and lower approximations. Equivalence classes are groups of objects which are indiscernible from each other, such as a group of objects in which all of the condition features are the same for each object.

2.1. Rough set attribute reduction

At the heart of the RSAR approach is the concept of indiscernibility. Let $I=(\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe) and $\mathbb{A}$ is a non-empty finite set of attributes so that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that a can take. For any $P \subseteq \mathbb{A}$, there exists an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P, a(x) = a(y)\} \tag{1}$$

The partition generated by $IND(P)$ is denoted as $\mathbb{U}/IND(P)$ or abbreviated to $\mathbb{U}/P$ and is calculated as follows:

$$\mathbb{U}/IND(P) = \otimes\{a \in P | \mathbb{U}/IND(\{a\})\} \tag{2}$$

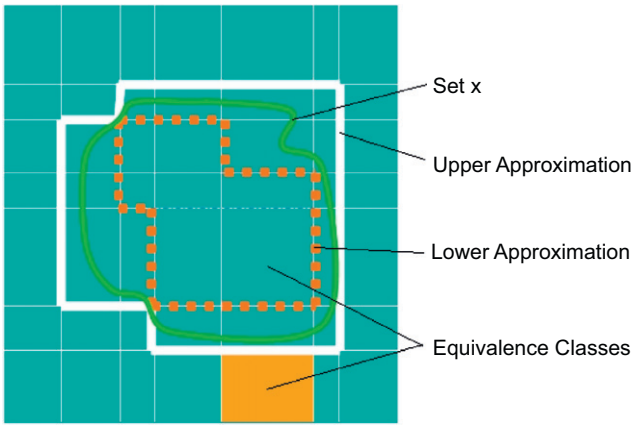

Fig. 1. Rough set representation.

where

$$\mathbb{U}/IND(\{a\}) = \{\{x | a(x) = b, x \in \mathbb{U}\} | b \in V_a\} \tag{3}$$

and

$$A \otimes B = \{X \cap Y | \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \tag{4}$$

where A and B are families of sets.

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted by $[x]_p$. Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained in P by constructing the *P-lower* and *P-upper* approximations of X:

$$\underline{P}X = \{x | [x]_p \subseteq X\} \tag{5}$$

$$\overline{P}X = \{x | [x]_p \cap X \neq \emptyset\} \tag{6}$$

Let P and Q be attribute sets that induce equivalence relations over $\mathbb{U}$, then the positive, negative and boundary regions can be defined as

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{7}$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \tag{8}$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{9}$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of a set of attributes Q on a set of attributes P. This can be achieved as follows: For P,$Q \subseteq A$, it can be said that Q depends on P in a degree k $(0 \leqslant k \leqslant 1)$, this is denoted as $(P \Rightarrow_k Q)$ if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \tag{10}$$

The reduction of attributes or selection of survival features can be achieved through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features. A *reduct* (R) can be defined as a subset of minimal cardinality of the conditional attribute set (C) where $\gamma_R(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})$, where D is the decision attribute set.

The *QuickReduct* algorithm in Ref. [10] also shown below searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until the search produces its maximum possible dependency value for that dataset ($\gamma_C(\mathbb{D})$). Note that this type of hill-climbing search does not guarantee a minimal subset and may only discover a local minimum.

Algorithm. *QuickReduct*
Input: C, the set of all conditional features
Input: D, the set of all decisional features
Output: R, a feature subset
1. $R \leftarrow \{\ \}$
2. **repeat**
3. $T \leftarrow R$
4. $\forall x \in (C - R)$
5. (**if** $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
6. $T \leftarrow R \cup \{x\}$
7. $R \leftarrow T$
8. **until** $\gamma_R(D) = \gamma_C(D)$
9. **return** R