Simulation Modelling Practice and Theory 64 (2016) 18–29

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

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

Towards scalable rough set based attribute subset selection for intrusion detection using parallel genetic algorithm in MapReduce

El-Sayed M. El-Alfy[a,*], Mashaan A. Alshammari[b]

[a] College of Computer Sciences and Engineering, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[b] Information and Computer Science Department, University of Ha'il, Ha'il, Saudi Arabia

ARTICLE INFO

Article history:
Available online 19 February 2016

Keywords:
Attribute subset selection
Rough sets
Hybrid methods
Minimum reduct
Big data
MapReduce
Parallel genetic algorithms

ABSTRACT

Attribute subset selection based on rough sets is a crucial preprocessing step in data mining and pattern recognition to reduce the modeling complexity. To cope with the new era of big data, new approaches need to be explored to address this problem effectively. In this paper, we review recent work related to attribute subset selection in decision-theoretic rough set models. We also introduce a scalable implementation of a parallel genetic algorithm in Hadoop MapReduce to approximate the minimum reduct which has the same discernibility power as the original attribute set in the decision table. Then, we focus on intrusion detection in computer networks and apply the proposed approach on four datasets with varying characteristics. The results show that the proposed model can be a powerful tool to boost the performance of identifying attributes in the minimum reduct in large-scale decision systems.

1. Introduction

Considering the characteristics of data in the new era of distributed information systems of being of massive scale and from multiple heterogeneous sources, data mining and analytics become challenging tasks. A central preprocessing step for extracting meaningful compact information in these systems is attribute selection. It has a plethora of decision-making applications in scientific, engineering and business domains [1–3]. The aim is to minimize the overhead caused by redundant and/or irrelevant attributes. This results in significantly reducing the search space and hence reducing the complexity and/or improving the effectiveness and interpretation of the developed predictive computational models.

Several attribute selection techniques have been proposed and categorized as being filter-, wrapper- or embedded-based methods [4,5]. The filter-based methods use some statistical measure, such as entropy or correlation, to rank the merit of individual attributes before the construction of a predictive model. On the other hand, both wrapper-based and embedded methods search for an optimal attribute subset that is tailored to a particular induction algorithm and domain. The wrapper methods differ from the embedded methods in the sense that their search strategy and evaluation function are two separate steps [6]. Unlike filter based methods, the number of candidate attribute subsets is gowning exponentially with the number of attributes. A useful benchmark study to compare several major attribute selection methods for supervised classification can be found in [7].

* Corresponding author. Tel.: +966 138601930; fax: +966 138602174.
E-mail addresses: alfy@kfupm.edu.sa (E.M. El-Alfy), maaw.alshammari@uoh.edu.sa (M.A. Alshammari).

http://dx.doi.org/10.1016/j.simpat.2016.01.010
S1569-190X(16)00017-4/

Rough sets have been applied successfully to attribute selection and rule induction in various fields, e.g. [8–12]. The rough set theory (RST) was first introduced by Pawlak in 1982 as a tool for handling uncertainty resulting from noisy or incomplete information systems [13]. In contrast to other methods, attribute selection based on rough set theory detects the attribute dependencies using the decision table [14–16]. A core part of rough set theory is the identification of the minimum reduct (i.e. minimal subset of conditional attributes) that has similar discernibility properties of the full attribute set. This problem has been described as an optimization problem, which is known to be NP-hard [17–19]. The brute-force search approach checks the merits of all possible combinations of attributes, which is very time consuming and hence becomes ineffective as the data size grows. To tackle this problem, a heuristic or metaheuristic search method is typically used with attribute evaluation [20,21]. Among these methods is genetic algorithms (GA) [22]. However, the traditional implementations of GA do not demonstrate the full potential of GA capabilities.

Although rough sets have been intensively studied in the past, new approaches need to be explored to cope with the emerging trend of big data [23]. This study is motivated with the inherent structure of GA for parallelism and the introduction of MapReduce framework for processing intensive datasets [24]. We propose a MapReduce approach for determining the minimum rough set reduct with a parallel genetic algorithm implementation. This paper builds on and extends our previous work [25]. It provides a more solid study with intensive review of the state-of-the-art methodologies on rough set based attribute selection. It also provides additional experiments to demonstrate the scalability of the proposed approach with focus on intrusion detection.

The remainder of this paper is organized as follows. Section 2 explains the concept of rough sets and describes the minimum reduct optimization problem. It also provides a brief background on the MapReduce programming paradigm. Section 3 reviews related work. Then, the proposed approach is explained in Section 4. Afterwards, we present the experiments, findings, and discussions in Section 5. Finally, Section 6 concludes the paper.

2. Preliminaries

2.1. Basics of rough sets and minimum reduct problem

The rough set theory is one of the powerful soft computing methods for the analysis and knowledge discovery in uncertain decision systems [13,18,26]. It has been extensively studied and applied to datasets from various sources such as bioinformatics, social networks, and meteorology. In such systems, a dataset or decision system is represented by a table where rows represent objects or instances and columns represent attributes or features. This table is formally described as a pair $S = (U, A)$ where $U = \{u_1, u_2, ..., u_N\}$ is a non-empty finite set of N objects or instances (called universe) and A is a non-empty set of $(n + k)$ attributes. The attribute set, $A = C \cup D$, consists of n conditional attributes or predictors, $C = \{a_1, a_2, ..., a_n\}$, and k decision attributes or output variables, $D = \{d_1, d_2, ..., d_k\}$. Each attribute $a \in A$ has an associated set of possible values V_a, known as domain of a.

A central problem in rough set is the determination of most relevant conditional attributes or minimal reduct. In order to formulate this problem, we start with some definitions. For each non-empty subset of attributes $P \subseteq C$, a binary relation called P-indiscernibility relation is defined as follows:

$$IND(P) = \{(u_1, u_2) \in U^2 : \forall a \in P, a(u_1) = a(u_2)\} \tag{1}$$

where $a(u_i)$ means the value of attribute a for the instance u_i. This means if $(u_1, u_2) \in IND(P)$, then u_1 is indistinguishable (indiscernible) from u_2 by the attributes P. Note that this relation is reflexive, symmetric, and transitive. The induced set of equivalence classes is denoted as $[u]_P$ where $u \in U$, and it partitions U into different blocks denoted as U/P.

The rough set approximates a concept or a target set of objects $X \subseteq U$ using the equivalence classes induced using P as follows:

$$\underline{P}X = \{u : [u]_P \subseteq X\} \tag{2}$$

$$\overline{P}X = \{u : [u]_P \cap X \neq \varphi\ \} \tag{3}$$

where $\underline{P}X$ and $\overline{P}X$ denote the P-lower (certainly classified as members of X) and P-upper (possibly classified as members of X) approximations of X, respectively. The notation $\cap$ denotes the intersection operation. The difference between the two approximations is known as the boundary region which represents uncertain objects. X is a crisp set if the boundary region is an empty set (i.e. accurate approximation); otherwise it is a rough set. To compare subsets of attributes, a dependency measure is defined. For instance, the dependency measure of an attribute subset Q on another attribute subset P is given as:

$$\gamma_P(Q) = \frac{|\bigcup_{X \in [u]_Q} \underline{P}X|}{|U|} \tag{4}$$

where $0 \leq \gamma_P(Q) \leq 1$, $\bigcup$ denotes the union operation, and $|.|$ denotes the set cardinality. The numerator of Eq. (4) is known as the positive region of Q with respect to P, and denoted as $POS_P(Q)$. The closer is $\gamma_P(Q)$ to 1, the more Q depends on P. Using this formulation of the information system, a reduct, R, is a subset of conditional attributes, C, such that $\gamma_R(D) = \gamma_C(D)$.