Computers and Electrical Engineering 000 (2017) 1–12

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Development of Rough Set – Hypergraph Technique for Key Feature Identification in Intrusion Detection Systems☆

M R Gauthama Raman[a], Kannan Kirthivasan[b], V S Shankar Sriram[a,*]

[a] Centre for Information Super Highway (CISH), School of Computing, SASTRA University, Thanjavur, Tamil Nadu, India
[b] Discrete Mathematics Research Laboratory (DMRL), Department of Mathematics, SASTRA University, Thanjavur, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 12 January 2016
Revised 6 January 2017
Accepted 6 January 2017
Available online xxx

Keywords:
Intrusion Detection System (IDS)
Hypergraph
Vertex linearity property
Minimal transversal property
Rough Set Theory (RST)
Optimal feature subset

ABSTRACT

'Curse of dimensionality' - an unresolved challenge in the design of an intelligent system makes dimensionality reduction a significant topic of research for the identification of informative features from high-dimensional data sets. This paper presents a novel feature selection technique based on Rough Sets (RS) and few interesting properties of Hypergraph (RSHGT), such as minimal transversal and vertex linearity for the identification of the optimal feature subset. Experiments were carried out using KDD cup 1999 intrusion dataset obtained from the UCI repository. Validation using Weka tool shows the dominance of RSHGT over the existing feature selection techniques with respect to the reduct size, classifier accuracy and time complexity. To summarize, RSHGT was found to be flexible, accommodative and computationally attractive for high dimensional data sets.

1. Introduction

In this digital era, the remarkable increase in the usage of internet and internet based applications has resulted in the massive generation and exchange of information across the globe. As information systems provided by various organizations are open to all types of internet users, it is mandatory to protect them from the cyber threats [1]. Traditional methods like user authentication, access control, data encryption, firewall, etc. are considered to be immature, since they are frail to the new types of vulnerabilities [2]. Hence, to have a strong line of defense against various intrusions, firewalls have been incorporated with intrusion detection capabilities. According to NIST, "*Intrusion is an attempt to compromise Confidentiality, Integrity and Availability (CIA) or to bypass the security mechanisms of a computer or a network; Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions, defined as attempts to compromise the CIA or to bypass the security mechanisms of a computer or network. Intrusion Detection Systems (IDSs) are software or hardware products that automate this monitoring and analysis process*" [3].

In recent years, a substantial amount of research work has been carried out in the design of an intelligent intrusion detection system. These works focus on the integration of various machine learning techniques (Neural Networks (NN), Support Vector Machine (SVM), K – Nearest Neighbor (KNN), Decision Trees, etc.) with data mining and statistical analytic techniques to enhance the detection rate and minimize the false alarm rate [1,4]. However, there exist a few limitations like

☆ Reviews processed and approved for publication by Editor-in-Chief.

* Corresponding author.
E-mail addresses: gauthamaraman_mr@sastra.ac.in (M.R. Gauthama Raman), kkannan@maths.sastra.edu (K. Kirthivasan), sriram@it.sastra.edu (V.S. Shankar Sriram).

http://dx.doi.org/10.1016/j.compeleceng.2017.01.006
0045-7906/

Table 1
Related works.

Authors	Technique developed
Wroblewski et al. [7]	Obtains the minimal feature subset using genetic algorithm
Jensen et al. [8]	Hybridization of RST and ant colony optimization for text processing and web content classification
Sengupta et al. [9]	Integration of RST and Q - learning to identify the reduct with maximum classification accuracy.
Jensen et al. [10]	Fuzzy based RST applied to discretized data set for the identification of features with minimal information loss.
Hu et al. [11]	Attribute-oriented induction combined with RST for knowledge discovery in the relational database.
K.Y. Hu et al. [12]	Feature ranking algorithm based on discernibility matrix in RST.
Mac Parthalain et al. [13]	Tolerance rough set model which utilizes the information in the boundary region
Slezak et al. [14]	Approximate entropy reduction principle - An extension of RST
Xu et al. [15]	Radix sorting technique to reduce the time complexity during attribute reduction based on RST
Wang et al. [16]	Feature selection based on RST and particle swarm optimization.
Jiang et al. [6]	Feature selection based on relative decision entropy.

computational complexity, curse of dimensionality, etc. that degrades the performance of an IDS. An important note is that the classifier accuracy can determine the performance of an IDS and in turn, the design of a preeminent feature selection technique enhances the effectiveness of the classifier. This scenario motivates the need for the design of a suitable feature selection technique which improves the performance of the classifier by minimizing the search space of the learning model through the identification of the most informative and optimal feature subset (reduct) from high dimensional datasets [5].

RST, an expansion of naive set theory has been an excellent tool for knowledge discovery from inconsistent, uncertain and incomplete datasets. The essence of RST is to obtain the minimal reduct from all possible feature sets, which is an NP-hard problem [6]. Various research works carried out on RST based feature selection techniques predominantly use heuristic functions to avoid the exponential computations in exhaustive methods, leaving out the problem of intensive computation unaddressed (Table 1). Therefore, to overcome these challenges, it is necessary to hybridize RST with a generic data representation tool. Hence, this paper puts forth a Rough Set - Hypergraph (RSHGT) feature selection technique for the identification of the optimal feature subset to enhance the performance of IDS, in terms of improved classification accuracy and reduced time complexity. RSHGT hybridizes the strength of RST with the interesting properties of hypergraph (minimal transversal and vertex linearity) for the construction of the optimal feature subset.

The rest of the paper is structured as follows: Section 2 enlightens the basics of RST and Hypergraph along with its properties. Section 3 describes the proposed feature selection technique - RSHGT. Section 4 discusses the experimental setup and performance analysis of RSHGT. Section 4 concludes the paper with future works.

2. Preliminaries

2.1. Rough Set Theory

Rough set theory was introduced during the year 1982 by Zdsislaw Pawlak [17] to handle imperfect and vague dataset. RST differs from other theories like fuzzy sets, Dempster-Shafer sets, etc. as it does not require any additional information for data analysis [18]. RST proves its efficiency through successful implementation in various domains such as machine learning, data mining and expert systems for knowledge discovery. The major assumptions behind RST are (i) objects in the universe of discourse represented by attributes with some information associated with them (ii) objects with similar information are indiscernible [18]. The basic concepts and fundamental definitions of RST are as follows.

Let U be a non-empty finite set of universe of discourse and R be the binary relation defined on U. The collection of (U, R) is known as approximation space and (U, A) is called as information system, where A is the finite set of attributes; a: $U \rightarrow V_a$ for $a \varepsilon A$. (V_a is the domain of attribute a); f: $U \times A \rightarrow V$ is a decision function for every $a \in A$ and $x \in U$ [19]

$$V = \bigcup_{a \in A} V_a \tag{1}$$

The data analysis carried out using RST starts from the decision table (D_T), which consists of rows and columns representing objects and attributes respectively. The attributes in D_T represented by class labels are known as decisional attributes (D), while the rest of the attributes are known as conditional attributes (C), such that $C \cap D =$"{}"

Let N be the subset of A, then the binary relation (indiscernibility relation) $IND(N)$ can be defined as

$$IND(N) = \{(x, y) \in U \times U \mid \forall\, a \in N, f(x, a) = f(y, a)\} \tag{2}$$

Through the above equivalence relation ($IND(N)$), the set U partitioned into disjoint subsets and family of all equivalent classes are denoted by $U/IND(N)$ or U/N. The equivalent class $IND(N)$ i.e. part of $U/N = \{N_1, N_2, \ldots, N_i, \ldots\}$ consists of object y denoted by $[y]_N$. The equivalent condition and decision classes can be represented by U/C and U/D respectively.

Definition 2.1. Given a five tuple decision table $D_T = (U, C, D, V, f)$. Let $Y \subseteq U$ & $N \subseteq C \cup D$. The N-lower approximation ($AR_N(Y)$), N - upper approximation ($AR^N(Y)$) and N - boundary region of the set Y can be defined as [19]

$$AR_N(Y) = \cup\{[y]_N \in U/IND(N) : [y]_N \in Y\} \tag{3}$$

Please cite this article as: M.R. Gauthama Raman et al., Development of Rough Set – Hypergraph Technique for Key Feature Identification in Intrusion Detection Systems, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.01.006