Applied Computing and Informatics (2015) **11**, 29–46

Saudi Computer Society, King Saud University

Applied Computing and Informatics

(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Multi-label rules for phishing classification

Neda Abdelhamid *

Computing and Informatics Department, De Montfort University, Leicester, UK

Received 3 January 2014; revised 13 May 2014; accepted 3 July 2014
Available online 15 July 2014

KEYWORDS

Associative rule;
Associative classification;
Data mining;
Website phishing;
On-line security

Abstract Generating multi-label rules in associative classification (AC) from single label data sets is considered a challenging task making the number of existing algorithms for this task rare. Current AC algorithms produce only the largest frequency class connected with a rule in the training data set and discard all other classes even though these classes have data representation with the rule's body. In this paper, we deal with the above problem by proposing an AC algorithm called Enhanced Multi-label Classifiers based Associative Classification (eMCAC). This algorithm discovers rules associated with a set of classes from single label data that other current AC algorithms are unable to induce. Furthermore, eMCAC minimises the number of extracted rules using a classifier building method. The proposed algorithm has been tested on a real world application data set related to website phishing and the results reveal that eMCAC's accuracy is highly competitive if contrasted with other known AC and classic classification algorithms in data mining. Lastly, the experimental results show that our algorithm is able to derive new rules from the phishing data sets that end-users can exploit in decision making.

* Tel.: +44 (0)116 2 50 60 70.
E-mail address: P09050665@myemail.dmu.ac.uk.
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2014.07.002

1. Introduction

Generally and according to (Tsoumakas and Katakis, 2007), there are two types of classification problems, these are termed as single label and multi-label. In a single label classification, each training case in the input data is associated with only one class. In cases where the input data set contains just two class labels, the problem is called binary classification. However, if more than two classes are available, the problem is named multi-class classification. The majority of the research works conducted in classification data mining are concerned with single label classification, i.e. (Li et al., 2008; Chien and chen, 2010; Wang et al., 2011). However, domain applications like medical diagnoses, website phishing detection, text categorisation (TC) and bioinformatics may necessitate the production of multiple label rules. This is since there is a class overlapping among the training cases in these applications data. Meaning a set of attribute values in the rule's body may link with more than one class in the data set and thus producing all these classes in the rule's consequent (right hand side). It is advantageous to generate the other classes besides the largest frequency class with the rule's body since they contain valuable information having a sufficient representation in the data set.

In the last few years, a learning strategy which applies the association rule in classification data called associative classification (AC) emerged (Thabtah et al., 2010; Thabtah et al., 2011; Wang et al., 2011). Most AC algorithms like MAC (Abdelhamid et al., 2012), CMAR (Li et al., 2001) and others usually apply an association rule technique to discover the rules, and then filter out the rules to include only those which their consequent is the class attribute. Experimental research works (Jabbar et al., 2013; Jabez, 2011) indicated that AC algorithms frequently build more accurate classifiers than classic classification approaches such as the probabilistic approach (Witten and Frank, 2002), decision tree (Quinlan, 1993) and rule induction (Cohen, 1995). The algorithm proposed in this article is part of the AC family.

Limited research attempts in AC have been conducted to produce rules with more than one class, i.e. Lazy AC (CLAC) (Veloso et al., 2011) and Multi-label Multi-class AC (MMAC) (Thabtah et al., 2004). The rest of the existing AC algorithms is unable to deal with discovering multi-label rules from single label data sets, and normally derive only the largest frequency class connected with the attribute value(s) and ignore all other classes even if these classes have large frequencies with the attribute value(s). For instance, this condition occurs if an attribute value such as $<A>$ is associated with two class labels (cl_1, cl_2) in different places (examples) within the training data set with frequencies equal to 44 and 45 respectively. A typical AC algorithm like CBA will only take on class "cl_2" simply because it has a larger frequency than cl_1 with $<A>$ and ignores class cl_1 even if this class is statistically significant with $<A>$. This surely makes the selection of ($<A>$, cl_2) questionable. In the proposed algorithm we pick the two class labels and construct a multi-label rule rather than removing class cl_1. This enables