Applied Soft Computing 48 (2016) 729–734

Contents lists available at ScienceDirect

Applied Soft Computing

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

A new fast associative classification algorithm for detecting phishing websites

Wa'el Hadi [a,*], Faisal Aburub [a], Samer Alhawari [b]

[a] University of Petra, MIS Department, Jordan
[b] The World Islamic Science & Education University, MIS Department, Jordan

ARTICLE INFO

Article history:
Received 18 April 2016
Received in revised form 27 July 2016
Accepted 2 August 2016
Available online 6 August 2016

Keywords:
Associative classification
Phishing websites
Classification
Data mining

ABSTRACT

Associative classification (AC) is a new, effective supervised learning approach that aims to predict unseen instances. AC effectively integrates association rule mining and classification, and produces more accurate results than other traditional data mining classification algorithms. In this paper, we propose a new AC algorithm called the Fast Associative Classification Algorithm (FACA). We investigate our proposed algorithm against four well-known AC algorithms (CBA, CMAR, MCAR, and ECAR) on real-world phishing datasets. The bases of the investigation in our experiments are classification accuracy and the F1 evaluation measures. The results indicate that FACA is very successful with regard to the F1 evaluation measure compared with the other four well-known algorithms (CBA, CMAR, MCAR, and ECAR). The FACA also outperformed the other four AC algorithms with regard to the accuracy evaluation measure.

1. Introduction

The rapid growth in the number of internet users has dramatically increased e-shopping as a practical alternative to traditional shopping. According to Ingham et al. [11], worldwide sales from e-shopping increased by 20.1% in 2014 to $1500 trillion. Therefore, the main objective of e-shops is to attract more and more consumers. This has led to increased competition among e-shops to introduce quality services for consumers. However, the increasing number of e-shops and websites has been accompanied by a rapid growth in the number of phishing websites. Although internet users are aware of phishing, many fall victim to such attacks. The aim of phishers is to make internet users believe that they are interacting with trusted online sites. Phishing websites can appear to be any type of website, including online payment or auction websites. An efficient method of identifying phishing websites is required in order to protect users' sensitive data.

Phishing applies both technical tricks and social engineering to access private information illegally. The purpose of the phishing is to take private information by publishing a forged website that appears to be a legitimate one, such as a real website of a company or bank, and requesting a person to input private information, such as account number, credit card number, username, and password. As well as being harmful to customers, phishing attacks also damage the reputation of the financial institutions concerned, since customers become less confident that they can securely access their accounts. Phishing websites are considered to be one of the most common electronic crimes [7,13]. According to a report from APWG [7], the number of distinct phishing reports submitted to the organization during quarter 4 of 2014 was 197,252. This was an increase of 18% on the 163,333 received in quarter 3 of 2014.

Data mining is a field of study that aims to find useful information in large databases in order to help decision makers to make correct decisions. Data mining involves many tasks, such as classification, association rules and clustering. Classification is the task of forecasting, assigning or predicting unseen instances to their predefined classes, for example, forecasting incoming email as either inbox or spam. Association rules is the task of finding relationships between attributes (features) in a large database. For example, if a consumer buys soda and potatoes together, they are also likely to buy meat; this relationship is represented as a rule of the form (*soda, potatoes* → *meat*), where "*soda and potatoes*" is called the rule body and "*meat*" the head of the rule. Associative classification (AC) is a new task in data mining and machine learning, and aims to forecast unseen instances based on association rules. AC is a promising approach because many researchers have indicated that it produces more accurate results than other traditional data mining classification techniques [19,2,1,5].

* Corresponding author.
E-mail addresses: whadi@uop.edu.jo (W. Hadi), faburub@uop.edu.jo (F. Aburub), samer.alhawari@yahoo.com (S. Alhawari).

http://dx.doi.org/10.1016/j.asoc.2016.08.005
1568-4946/

The main goals of this paper are to present a new, fast, and very efficient AC classifier, and to compare this new AC classifier with four well-known AC algorithms with reference to classification accuracy and F1 evaluation measures on a new phishing dataset proposed by Mohammad et al. [16].

The rest of this paper is organized as follows. Related works are explained in Section 2 and the proposed AC classifier is described in Section 3. In Section 4, the experimental results are discussed, and finally, in Section 5, conclusions are presented.

2. Related works

Over the past decade, many researchers have investigated the problem of detecting phishing websites using data mining techniques, but there are a limited number of research articles relating to the AC approach. In this section, we shed light on both traditional data mining techniques and AC approaches.

Abdelhamid et al. [3] investigated the problem of website phishing using a new proposed multi-label classifier-based associative classification, MCAC. The main goal of the MCAC algorithm developed is to recognize attributes or features that distinguish phishing websites from legitimate ones. The results showed that the MCAC algorithm forecasted phishing websites better than traditional data mining algorithms.

Dadkhah et al. [8] developed a new method to forecast and detect phishing websites using classification algorithms based on the weight of web page features. The results showed that the proposed method produced a lower error rate than other data mining methods.

Abdelhamid [1] proposed an enhanced multi-label classifier-based associative classification algorithm, eMCAC. This generates rules associated with a set of classes from single-label datasets using the transaction ID list (Tid-list) vertical mining approach. The algorithm employs a novel classifier building method that reduces the number of generated rules. The experiments indicated that the eMCAC algorithm outperformed other algorithms with regard to the accuracy evaluation measure.

Jabri and Ibrahim [12] proposed an enhanced PRISM algorithm for forecasting phishing websites. The experimental results revealed that the modified PRISM algorithm outperformed the original PRISM algorithm in terms of the number of rules, accuracy (87%), and lower error rate (0.1%).

Alazaidah et al. [5] proposed a new multi-label classification algorithm based on correlations among labels, MLC-ACL. The MLC-ACL utilizes both problem transformation techniques and algorithm adaptation techniques. The proposed algorithm starts by converting a multi-label dataset into a single-label dataset using the least frequent label criteria, and then employs the PART machine learning classifier on the converted dataset. The output of the classifier is multi-label rules. In addition, MLC-ACL attempts to gain advantage from positive correlations among labels using the predictive Apriori algorithm. The MLC-ACL algorithm was investigated using two multi-label datasets named Emotions and Yeast. The experiments revealed that the MLC-ACL algorithm outperformed other machine learning algorithms in terms of three well-known evaluation measures (Hamming Loss, Harmonic Mean, and Accuracy).

Taware et al. [17] proposed a new MCAC that aims to recognize attributes that differentiate phishing websites from legitimate ones. The MCAC algorithm produced better results than other data mining algorithms with regard to accuracy.

Antonelli et al. [6] developed a new efficient AC algorithm using a fuzzy frequent pattern method. The experiment results showed that the new fuzzy AC algorithm outperformed the well-known CMAR algorithm and generated accuracies similar to two recent AC algorithms, namely FARC-HD and D-MOFARC, on 17 real-world datasets.

The problem with the current AC algorithms is that the set of candidate rules produced from the training data is typically large, which consumes time and Input/Output resources. This problem motivated us to propose a new AC algorithm that generates all frequent rules using an efficient association rule mining method that reduces the time and memory required. Moreover, we propose a new prediction method to forecast unseen instances more accurately than other methods. Unlike most of the current AC prediction methods, our proposed method considers multiple rules to assign the class in the prediction step.

Table 1
Training data.

Tid	Age	Income	Has a car	Buy/class
1	Senior	middle	yes	yes
2	Youth	low	yes	no
3	Junior	high	yes	yes
4	Youth	middle	yes	yes
5	Senior	high	no	yes
6	Junior	low	no	no
7	Senior	middle	no	no

3. Proposed algorithm

AC mining is a new branch of data mining and machine learning that classifies unseen instances based on association rules. AC mining algorithms have been investigated during the past two decades by many specialist researchers in real-world fields such as phishing websites, text classification, medical diagnoses, and fraud detection. The AC mining approach is widely used for a number of reasons. First, it produces higher classification accuracy rates for the outputting classifier than other data mining and machine learning approaches. Second, rules produced by AC mining classifiers are simple and illustrated by simple "*if–then*" rules, so the user can easily read, remove, modify, and understand the produced rules. The primary goal of an AC mining approach is to build a classifier (model) from a huge database (training data) to forecast (detect or predict) the type of unseen instances (testing data).

In this paper, a new, fast, and efficient AC mining classifier called the Fast Associative Classification Algorithm (FACA) is developed. FACA is an efficient AC mining classifier, the difference between FACA and all other AC classifiers in the literature. It employs a vertical mining approach called Diffset [20] for discovering all frequent itemsets, and utilizes a new prediction method to classify unseen instances more accurately than other methods. To the best of the author's knowledge, there is no AC mining algorithm that implements the Diffset method.

Before we discuss the FACA algorithm steps, more elaboration of the Diffset technique is required in order to ensure a better understanding. Diffset is a vertical data approach that keeps track of only the transaction IDs in which a ruleitem does not occur. For example, the Diffset for <Age, senior> → yes is {2, 3, 4, 6, 7}, according to the training data shown in Table 1, because the ruleitem <Age, senior> → yes occurs in two transactions {1, 5}. The support of a candidate K-ruleitem (rule with K items in its body) is the cardinality of the Diffset for the (K − 1)-ruleitem subtracted from the cardinality of the Diffset for the K-ruleitem itself. If a candidate rule is a single-ruleitem, the support is computed by subtracting the number of transactions in the training data from the cardinality of the Diffset for the single-ruleitem itself, i.e., the support for the single-ruleitem <Age, senior> → yes = 7 − 5 = 2. The confidence of the ruleitem A → B is the conditional probability that a transaction contains B, given that it contains A, and is computed by dividing the support of (A$\bigcup$B) by the support of (A), i.e., the con-