



A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems



Adel Sabry Eesa^a, Zeynep Orman^{b,*}, Adnan Mohsin Abdulazeez Brifcani^c

^a Computer Science Department, Faculty of Science, Zakho University, Duhok City, KRG, Iraq

^b Department of Computer Engineering, Faculty of Engineering, Istanbul University, 34320 Avcilar, Istanbul, Turkey

^c Department of IT, Duhok Technical Institute, Duhok Polytechnic University, Duhok City, KRG, Iraq

ARTICLE INFO

Article history:

Available online 15 November 2014

Keywords:

Feature selection
Cuttlefish algorithm
Intrusion detection systems
Decision trees
ID3 algorithm

ABSTRACT

This paper presents a new feature-selection approach based on the cuttlefish optimization algorithm which is used for intrusion detection systems (IDSs). Because IDSs deal with a large amount of data, one of the crucial tasks of IDSs is to keep the best quality of features that represent the whole data and remove the redundant and irrelevant features. The proposed model uses the cuttlefish algorithm (CFA) as a search strategy to ascertain the optimal subset of features and the decision tree (DT) classifier as a judgement on the selected features that are produced by the CFA. The KDD Cup 99 dataset is used to evaluate the proposed model. The results show that the feature subset obtained by using CFA gives a higher detection rate and accuracy rate with a lower false alarm rate, when compared with the obtained results using all features.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Due to the expansion of computer networks, the number of hacking and intrusion incidents is increasing year by year as technology rolls out, which has made many researchers focus on building systems called intrusion detection systems (IDSs). These systems are used to protect computer systems from the risk of theft and intruders (Liao, Lin, Lin, & Tung, 2013). IDSs can be categorized as anomaly detection and misuse detection or signature detection systems (Depren, Topallar, Anarim, & Ciliz, 2005; Wang, Hao, Ma, & Huang, 2010). In anomaly detection, the system builds a profile of that which can be considered as normal or expected usage patterns over a period of time and triggers alarms for anything that deviates from this behaviour. On the other hand, in misuse detection, the system identifies intrusions based on known intrusion techniques and triggers alarms by detecting known exploits or attacks based on their attack signatures.

Dimensionality reduction is a commonly used step in machine learning, especially when dealing with a high dimensional space of features (Fodor, 2002; Van der Maaten, Postma, & van den Herik, 2008). Feature selection (FS) is a part of dimensional reduction which is known as the process of choosing an optimal subset of features that represents the whole dataset. FS has been used in

many fields, such as classification, data mining, object recognition and so forth, and has proven to be effective in removing irrelevant and redundant features from the original dataset. Given a feature set of size n , the FS problem tries to find a minimal feature subset of size m ($m < n$) that enables the construction of the best classifier with high accuracy (Basiri, Ghaseem-Aghaee, & Aghdam, 2008).

FS has been a fertile field of research and development since the 1970s, and it is used successfully in the IDSs domain. Stein, Chen, Wu, and Hua (2005) proposed a hybrid genetic-decision tree (DT) model. They used the genetic algorithm (GA) as a generator to produce an optimal subset of features, and then the produced features were used as an input for the DT that was constructed using the C4.5 algorithm. Bolon-Canedo, Sanchez-Marono, and Alonso-Betanzos (2011) proposed a new combinational method of discretization, filtering and classification which is used as an FS to improve the classification task, and they applied this method on the KDD Cup 99 dataset. Lin, Ying, Lee, and Lee (2012) presented an intelligent algorithm which was applied to anomaly intrusion detection. The paper proposed simulated annealing (SA) and support vector machine (SVM) to find the best feature subsets, while SA and DT were proposed to generate decision rules to detect new attacks. Tsang, Kwong, and Wang (2007) proposed an intrusion detection approach to extract accurate and interpretable fuzzy IF–THEN rules from network traffic data for classification. They also used a wrapper genetic FS to produce an optimal subset of features. Lassez, Rossi, Sheel, and Mukkamala (2008) proposed a new method for FS and

* Corresponding author.

E-mail addresses: adelsabryissa@gmail.com (A.S. Eesa), ormananz@istanbul.edu.tr (Z. Orman), president@dpu.ac (A.M.A. Brifcani).

extraction by using the singular value decomposition paired with the notion of latent semantic analysis, which could discover hidden information to design signatures for forensics and eventually real-time IDSs. They used three automated classification algorithms (Maxim, SVM, LGP). Nguyen, Franke, and Petrovic (2010) presented a generic-feature-selection (GeFS) measure to find global optimal feature sets by using two methods: the correlation feature-selection (CFS) measure and the minimal redundancy-maximal-relevance (mRMR) measure. This approach is based on solving a mixed 0–1 linear programming problem by using the branch-and-bound algorithm, and the authors applied the proposed method to design IDSs. A hybrid model based on the information gain ratio and K-means is proposed by Neelakantan, Nagesh, and Tech (2011) to detect 802.11-specific intrusions. They used the information gain ratio as the FS and the K-means algorithm as the classifier. Mohanabharathi, Kalaikumar, and Karthi (2012) proposed a new method which was a combination of the information gain ratio measure and the K-means classifier used for FS. The back-propagation algorithm was also used for the learning and testing processes. Datti and Lakhina (2012) compared the performance of two feature reduction techniques: principal component analysis and linear discriminate analysis. As a classifier, they used the back-propagation algorithm to test these techniques.

Since IDSs deals with a large amount of data, FS is a critical task in IDSs. In this paper, we propose an FS model based on the cuttlefish optimization algorithm (CFA) to produce the optimal subset of features. DT is also used as a classifier to improve the quality of the produced subsets of features. The rest of this paper is organised as follows: Section 2 presents an introduction and a brief overview of DT and CFA. The proposed feature-selection approach is discussed in Section 3. Section 4 reports on the experimental results of the proposed cuttlefish feature-selection approach and a brief discussion on the obtained results. Finally, the conclusions and future work are stated in Section 5.

2. Introduction to DT and the cuttlefish optimization algorithm

2.1. Decision tree (DT)

DT is one of the most well-known machine learning techniques produced by Quinlan (Salzberg, 1994). DT has three main components: nodes, arcs, and leaves. Each node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values. The main node (root node) is also called the test node which has no incoming edges. Each arc out of a node is labelled with an attribute value and each leaf is labelled with a category or a class. The tree is constructed during a training phase by using the training data. In the test phase, each instance of the test data is classified by the navigation from the root of the tree down to a leaf, according to the outcome of the test data along the path. There are two popular algorithms which are used for constructing the DT: ID3 and C4.5 (Salzberg, 1994). In this paper we use the ID3 algorithm.

2.2. Cuttlefish algorithm (CFA)

In previous work, we produced a novel optimization algorithm called the CFA (Eesa, Abdulazeez, & Orman, 2013). The algorithm mimics the mechanisms behind a cuttlefish that are used to change its colour. The patterns and colours seen in cuttlefish are produced by reflected light from different layers of cells including chromatophores, leucophores and iridophores. The CFA considers two main processes: reflection and visibility. The reflection process is used to simulate the light reflection mechanism, while visibility is used to simulate the visibility of matching patterns. These two processes

are used as a search strategy to find the global optimal solution. The diagram in Fig. 1 of cuttlefish skin, detailing the three main skin structures (chromatophores, iridophores and leucophores), two example states (a, b) and three distinct ray traces (1, 2, 3), shows the sophisticated means by which cuttlefish can change reflective colour (Eric et al., 2012).

CFA reorders these six cases shown in Fig. 1 to be as shown in Fig. 2. The formulation for finding the new solution (*newP*) using reflection and visibility is described in Eq. (1):

$$newp = reflection + visibility \tag{1}$$

For Cases 1 and 2 shown in Fig. 2, CFA uses the two processes reflection and visibility to find a new solution. These cases work as a global search using the value of each point to find a new area around the best solution with a specific interval. The formulations of these processes are described in Eqs. (2) and (3), respectively:

$$reflection_j = R * G_1[i].Points[j] \tag{2}$$

$$visibility_j = V * (Best.Points[j] - G_1[i].Points[j]) \tag{3}$$

where, G_1 is a group of cells, i is the i th cell in G_1 , $Points[j]$ represents the j th point of the i th cell, $Best.Points$ represents the best solution points, R represents the degree of reflection, and V represents the visibility degree of the final view of the pattern. R and V are found as follows:

$$R = random() * (r_1 - r_2) + r_2 \tag{4}$$

$$V = random() * (v_1 - v_2) + v_2 \tag{5}$$

where, $random()$ function is used to generate random numbers between (0,1) and r_1, r_2, v_1, v_2 are four constant values specified by the user. As a local search, CFA uses Cases 3 and 4 to find the difference between the best solution and the current solution to

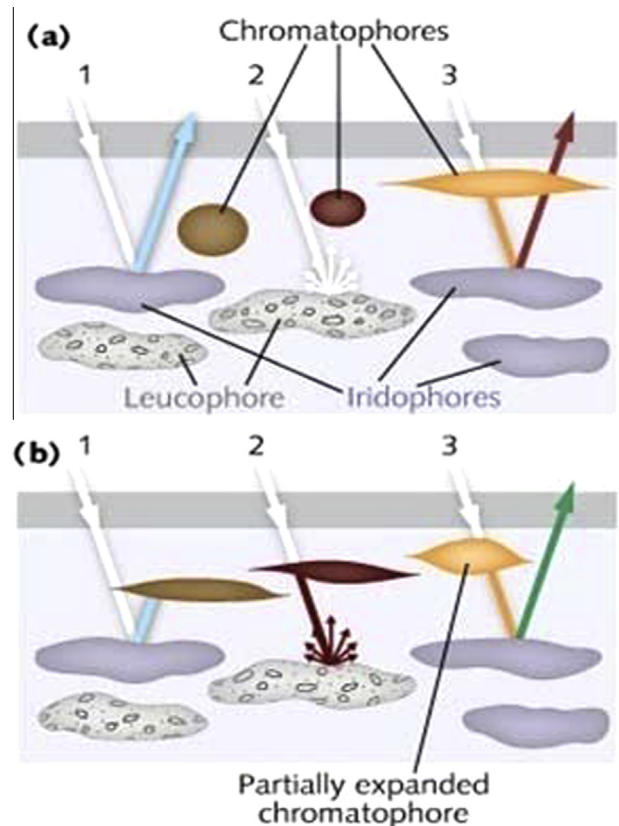


Fig. 1. Diagram of cuttlefish skin detailing the three main skin structures (chromatophores, iridophores and leucophores).

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات