Applied Soft Computing 62 (2018) 592–601

Contents lists available at ScienceDirect

Applied Soft Computing

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

Large-scale cyber attacks monitoring using Evolving Cauchy Possibilistic Clustering

Igor Škrjanc[a,*], Seiichi Ozawa[b], Tao Ban[c], Dejan Dovžan[a]

[a] Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[b] Kobe University, Kobe, Japan
[c] National Institute of Information and Communications Technology, Tokyo, Japan

ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form 30 October 2017
Accepted 4 November 2017
Available online 16 November 2017

Keywords:
Big-data
Data stream
Evolving clustering
Cauchy density
Cyber security

ABSTRACT

We are living in an information age where all our personal data and systems are connected to the Internet and accessible from more or less anywhere in the world. Such systems can be prone to cyber-attacks; therefore the monitoring and identification of cyber-attacks play a significant role in preventing the abuse of our data and systems. The majority of such systems proposed in the literature are based on a model/classifiers built with the help of classical/off-line learning methods on a learning data set. Since cyber-attacks evolve over time such models or classifiers sooner or later become outdated. To keep a proper system functioning the models need to be updated over a period of time. When dealing with models/classifiers learned by classical off-line methods, this is an expensive and time-consuming task. One way to keep the models updated is to use evolving methodologies to learn and adapt the models in an on-line manner. Such methods have been developed, extensively studied and implemented for regression problems. The presented paper introduces a novel evolving possibilistic Cauchy clustering (eCauchy) method for classification problems. The given method is used as a basis for large-scale monitoring of cyber-attacks. By using the presented method a more flexible system for detection of attacks is obtained. The approach was tested on a database from 1999 KDD intrusion detection competition. The obtained results are promising. The presented method gives a comparable degree of accuracy on raw data to other methods found in the literature; however, it has the advantage of being able to adapt the classifier in an on-line manner. The presented method also uses less labeled data to learn the classifier than classical methods presented in the literature decreasing the costs of data labeling. The study is opening a new possible application area for evolving methodologies. In future research, the focus will be on implementing additional data filtering and new algorithms to optimize the classifier for detection of cyber-attacks.

1. Introduction

Recent developments in the Information and Computer Technology (ICT) smart technologies such as data processing, pattern or feature extraction and their classification are an important part of many automated operations in the real world. By sensing and processing various kinds of information from the physical world, many applications and tasks can be improved. By understanding the big data, an improvement in decision making can be utilized. In general analysis of data can result in improving the decisions, detection of faults and monitoring of the system operation. It can also be used for detection of frauds and monitoring, augmenting and enhancing the cyber security in real-time. The big data technologies enable the analysis of new types of social behavior which can significantly improve intelligence, comfort, security, etc. Analysis of big data can be efficiently performed by the use of clustering and classification techniques. These techniques are already very well established. Usually, these methods are used for preprocessing data in a batch form. In a modified form they can be efficiently used also for problems where the data are coming continuously as a data stream. In this case, the analytics should be done in real time based on obtained models. On-line analytics can be done using evolving identification methods.

The concept of evolving identification methods allows the adaptation of the parameters and the structure simultaneously. Since they are one-sweep algorithms, they are also usefully for big data

* Corresponding author.
E-mail address: igor.skrjanc@fe.uni-lj.si (I. Škrjanc).

https://doi.org/10.1016/j.asoc.2017.11.008
1568-4946/

analytics. The advantages of evolving methods over the batch are that they are faster and can simply include new information into the model without retraining the whole model [1]. On the other hand, they usually generate a bit less accurate models. The evolving methods were first based on a neuro-fuzzy principle. The adaptation algorithms were based on gradient decent and chain rule approach [2]. Further the development focused on evolving fuzzy models. Where on-line clustering methods derived from the off-line versions are used for space partitioning and recursive least squares are used for local parameter identification. This recently transitioned in the idea of cloud based identification where instead of membership degrees the densities, and instead of clusters clouds are used. The presented method can be classified as a cloud based identification method.

In the digitalized world majority of critical data and systems are connected to the Internet and can be accessed from practically anywhere in the world. To protect the data and systems from unauthorized access and attacks systems for monitoring the network connections are used. Such systems should be able to distinguish between a normal connection and a cyber-attack. Since the cyber-attacks evolve over time [3] and new attacks emerge, the use of an evolving algorithm seems a natural solution for monitoring cyber-attacks. The approaches found in the literature are usually based on off-line learning methods. The classifier structure is learned beforehand based on training data. The classifier is then used for detection of intrusion. The classifier is fixed and does not take into account possible new information about the novel attacks. To include the new information the whole classifier must be retrained. Most of the currently available evolving methods are proposed for regression problems. In this paper, a novel on-line learning algorithm is presented that can adapt the classifier in an on-line manner. Further, the paper introduces the idea of using evolving methodologies for designing a cyber-attack detection system. As it will be seen from the obtained results, such system could in practice decrease the cost of labeling the learning data. Since it is an on-line approach, the inclusion of definitions for new attacks is much faster and simpler than with the batch learning algorithms. The cloud-based implementation of such system would decrease the reaction time of the detection system to new attacks.

The paper is organized as follows: after the introduction and related work review, the problem of intrusion detection is shortly described. Follows the methodology description where the Cauchy density for the data stream is presented. The general density with inner matrix norm is presented and extension to the recursive computation is given. At the end results of network intrusion detection are presented using the proposed methodology. The results are discussed and compared to other results found in the literature.

1.1. Related work

The evolving methods can be divided into two groups based on the mechanisms that they implement: methods based on fuzzy logic, which employ unsupervised space partitioning with on-line clustering (eTS [1], exTS [4,5], simpl_eTS [6], +eTS [7], FLEXFIS [8], FLEXFIS+ [9]); and methods based on neural networks that exploit the functional equivalence of neural network and fuzzy reasoning [10] (DENFIS [11], SAFIS [12], NeuroFAST [13], ENFM [14]). Identification based on data clouds was introduced recently [15,16], which is very similar to the fuzzy logic concept.

Depending on the learning abilities, the evolving methods can be divided into: *Adaptive methods* (e.g., ANFIS [17], rFCM [18], rGK [19], rPFM [20], AFPC [21]), where the initial structure of the fuzzy model must be given. The number of space partitions/clusters does not change over time, only the parameters of the membership functions and local models are adapted; *Incremental methods* (e.g., NeuroFAST [13], DENFIS [11], eTS [1], FLEXFIS [8], PANFIS [22]), Incremental Granual Fuzz Model [23], FCRM [24], eFT [25] where only adding mechanisms are implemented; *Evolving methods* (e.g., SAFIS [12], SOFNN [26], ENFM [14], eTS+ [7], ENFM [14], FLEXFIS++ [27], AHLTNM [28], SOFMLS [29], aFCR [30], ePL [31], ENF [32]) which, besides an adding mechanism, implement removing and some of them also merging and splitting mechanisms. The above-mentioned methods were mainly developed for on-line learning of regression models. They use recursive least squares method for model parameters estimation and recursive/on-line clustering for space partitioning. On-line clustering can also be used for solving the classification problems as in our case.

When dealing with the on-line clustering algorithms, the clustering of the input-output space is updated using a new data sample from the stream. The estimation of the clusters' parameters is calculated by the recursive clustering algorithm. The on-line clustering algorithms are usually derived from off-line clustering algorithms. Recursive version of Gustafson-Kessel clustering is presented in [33,19]. In [1] an on-line version of subtractive clustering is presented. In [14] a recursive method based on Gath-Geva clustering algorithm is introduced, and in [34], a recursive possibilistic fuzzy modeling approach is given. The basic on-line clustering algorithms are based on the Euclidian distance, which results in hyperspherical shapes of clusters [35]. By introducing a fuzzy covariance matrix and its inverse, by using Mahalanobis distance, also the clusters of hyper-ellipsoid shapes can be detected as shown in [19,33,36]. More on evolving approaches and on-line clustering can be found in [37].

The presented idea of classification algorithm merges the concepts of possibilistic clustering (PCM) [38] and possibilistic fuzzy clustering (PFCM) [39–41] algorithms together with the idea of density and clouds given in [42,43]. With this, an on-line clustering approach was obtained, which does not assume the normalized membership values of the data sample, as fuzzy c-means based clustering algorithms. A data sample might have a zero membership degree to clusters when it is far away from it. This fact can be utilized for detecting new patterns in the data or detecting outliers. The space partitioning is made by using an evolving methodology which classifies the data into clusters and adds new clusters when necessary. The density in [42] is based on Euclidean distance among the data samples which belong to the cluster. In this paper, the density is generalized by using a general inner matrix norm. The proposed method is a continuation of previously published work in the area of evolving systems, such as in [7], where evolving Takagi–Sugeno fuzzy system for streaming data (eTS+) is presented, or in [15,16] where cloud-based identification is presented and basic problems of such identification are given. The evolving methodology was used in [44] for LRF data mapping and in [36] for process monitoring.

The presented method Evolving Cauchy Possibilistic Clustering was implemented as a basis for large-scale monitoring of cyber-attacks. For this purpose a classifier is built which is capable of distinguishing between a normal network connection and attack. In this paper, the data set generated and managed by MIT Lincoln Labs for the purpose of 1998 DARPA Intrusion Detection Evaluation Program was used [45]. The intrusion detection systems play a fundamental role in the prevention of security breaches. Usually, the intrusion detection systems (IDSs) are rule-based. This limits the detection of novel intrusions. Moreover, encoding rules is time-consuming and highly depends on the knowledge of known intrusions [46]. Therefore, several approaches were proposed over the past decade, that are based on machine learning and data mining principles [47]. In [46] a random forest approach is proposed to build the classifier. The classifier is tested on a 10 percent KDD data set and achieves accuracy around 98 percent. In [48] a classifier is designed based on the principal component analysis. The reported attack detection rate of this classifier was 98 percent while