Knowledge-Based Systems 70 (2014) 324–334

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Fraud detection using self-organizing map visualizing the user profiles

Dominik Olszewski *

Faculty of Electrical Engineering, Warsaw University of Technology, Poland

ARTICLE INFO

Article history:
Received 11 April 2014
Received in revised form 15 July 2014
Accepted 16 July 2014
Available online 24 July 2014

Keywords:
Fraud detection
Self-organizing map
Threshold classification
Classification threshold setting
Data visualization

ABSTRACT

We propose a fraud detection method based on the user accounts visualization and threshold-type detection. The visualization technique employed in our approach is the Self-Organizing Map (SOM). Since the SOM technique in its original form visualizes only the vectors, and the user accounts are represented in our work as the matrices storing a collection of records reflecting the user sequential activities, we propose a method of the matrices visualization on the SOM grid, which constitutes the main contribution of this paper. Furthermore, we propose a method of the detection threshold setting on the basis of the SOM U-matrix. The results of the conducted experimental study on real data in three different research fields confirm the advantages and effectiveness of the proposed approach.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The Self-Organizing Map (SOM) is an example of the artificial neural network architecture. It was by introduced by Kohonen [1] as a generalization and extension of the concepts proposed in [2]. This algorithm can be also interpreted as a visualization technique, because it may perform a projection from an input multidimensional space onto an output 2-dimensional space, in this way creating a map structure. The location of points in 2-dimensional grid aims to reflect the similarities between the corresponding objects in multidimensional space. Therefore, the SOM algorithm allows for visualization of relationships between objects in multidimensional space. An exhaustive and detailed description of the SOM method can be found in [3].

We employ the SOM method in the fraud detection framework. Fraud detection using a visualization technique is a particularly profitable approach, since it assures the two significant advantages. First of all, a graphical representation of an analyzed dataset is eligible for convenient analysis and interpretation (in case of images' dimensions eligible for displaying on a screen), even by a non-expert, who can formulate some conclusions or at least suspicions due to the analyzed data (for example, the person may notice certain fraudulent activity). Second of all, the efficient and effective visualization technique (like, for example, the SOM method) combined with a chosen classification algorithm (performing the actual detection) leads to satisfactory detection results.

In case of real-world complex knowledge-based systems collecting and managing information on multiple individuals (for example, users or customers), typically, data corresponding to a single individual is organized as a set of data records, where each of the records represents a single activity of the individual, for example, a single phone call or a single credit card payment. The knowledge organized in the system in this way cannot be directly and straightforwardly visualized using the SOM method, because the method in its traditional form uses vectors as input data, and it is not capable to handle input data represented as matrices. This constitutes the main problem addressed in this paper, and the paper's proposal contains a method providing the aforementioned SOM capability, i.e., matrices visualization. Furthermore, we propose using a threshold-type binary classification algorithm allowing for the fraudulent activity detection on the basis of the SOM visualization, which leads to achieving the final goal of our research.

1.1. Our proposal

In this paper, we propose a fraud detection method based on the SOM visualization and classification. Consequently, the proposed approach consists of the two main steps:

Step 1. SOM visualization of the multidimensional data of the user accounts.
Step 2. The actual fraud detection on the basis of the threshold-type binary classification algorithm.

In Step 1, the entire user accounts are visualized in 2-dimensional space of the SOM grid. The user accounts are numerically

* Address: Koszykowa 75 Street, 00-662 Warsaw, Poland. Tel.: +48 22 234 7618, +48 22 625 6278; fax: +48 22 625 6278.
E-mail address: dominik.olszewski@ee.pw.edu.pl

http://dx.doi.org/10.1016/j.knosys.2014.07.008
0950-7051/© 2014 Elsevier B.V. All rights reserved.

represented as data matrices storing a collection of records reflecting the user activities. In other words, the accounts are the data objects characterized by features possessing their own inner-dimensionality (consequently, creating the matrices). Since the standard SOM technique visualizes only single vectors, it is necessary to formulate a method of the user accounts visualization on the SOM lattice, which is the main proposal of this paper (described in detail in Section 5).

In Step 2, the threshold-type binary classification algorithm performs the final detection of fraudulent accounts. In our paper, we propose also a method of the classification threshold setting on the basis of the SOM U-matrix.

The advantage of the introduced method is that it is a general fraud detection approach, i.e., it is not oriented to certain particular field or application, and it can be easily adopted in every information system collecting the data deriving from users sequential activity. Furthermore, our approach is an unsupervised technique, thus, avoiding the problems associated with insufficient training data, which essentially affect the final detection results of supervised data mining methods (mentioned in Section 2).

The experimental study has been conducted on real data in three different research fields, i.e., in the field of fraud detection in telecommunications, in the field of the computer network intrusion detection, and in the field of the credit card fraud detection. The empirical research verifies and confirms the usefulness and effectiveness of the proposed approach, and it demonstrates the benefits associated with the preliminary data visualization, which transforms the input high-dimensional information into a 2-dimensional image – easy and convenient for analysis and interpretation, even by non-experts. Although our experiments have been carried out on the three chosen specific datasets, the proposed approach itself is more general, and it can be easily applied in case of various datasets containing the data reflecting the users' repeatedly evinced behavior.

1.2. The remainder of this paper

The rest of this paper is organized as follows: in Section 2, the related work is discussed as the background for the present paper's proposal; in Section 3, the SOM algorithm is described; in Section 4, the representation of the user data is described and explained; in Section 5, the main proposal of our paper, i.e., a method of the user accounts visualization on the SOM, is presented; in Section 6, the actual fraud detection is described; in Section 7, the experimental results are reported; while Section 8 summarizes the whole paper, provides some concluding remarks, and points out certain directions of the future research.

2. Related work

There is a number of fraud detection problems, including financial fraud detection [4–14], Internet fraud detection [15–22], telecommunications fraud detection [23–27], and various other areas of fraudulent activity, like those from the papers [28–35], to name but a few. A common major difficulty associated with all those fraud detection fields is that there is a large amount of data that needs to be analyzed, and simultaneously, there is only a small number of fraudulent samples, which could be used as the training data for the supervised methods. Consequently, this problem essentially inhibits and limits an application of the supervised techniques. The SOM approach proposed for fraud detection in this work is an unsupervised method, therefore, it is robust to the mentioned before difficulty, and consequently, it is especially useful in the fraud detection framework.

The general problem of fraud detection has been reviewed in [36,28,37–39].

An idea of using SOM (or its modified version) as the preliminary data analysis tool preceding the actual detection appears in the papers [7,5]. However, the difference between those papers' proposals and our research is that in [7,5], the SOM technique is not used for the visualization of user accounts purpose. In [7], it is employed along with the neural gas method as a clustering technique in order to identify clusters or groups of similar behavior in the universe of taxpayers. Afterwards, certain classification algorithms are applied to achieve the final goal of fraud detection. In the paper [5], in turn, a novel dual Growing Hierarchical SOM (GHSOM) approach is developed to discover the topological patterns of fraudulent financial reporting. Subsequently, on the basis of the topological patterns a classification rule detecting the financial frauds is presented. Compared with our research, both of the papers [7,5] expect a different result when applying SOM and a novel GHSOM, respectively. Their aim is not to exploit purely visual nature of the SOM output, and their study focuses on obtaining data clusters based on the SOM (or the novel GHSOM) grid, rather than utilizing the SOM visualization in its original form. Furthermore, the papers [7,5] use different classification algorithms in the actual detection phase, and they both consider fraud detection problem in a particular application field, i.e., in the field of financial fraud detection, whereas our work addresses the issue in a wider spectrum of application areas.

A fraud detection method based on the SOM technique is proposed also in the work [40]. However, the authors of [40] train SOM to store probabilistic models of the user profiles. The models are build using the negative log probability introduced by the authors. Frauds are detected on the basis of the clustering of the probabilistic models. The approach from [40] is oriented to a specific application field, i.e., to the mobile telecommunications network user profiling. Our method, on the other hand, is more general, and it can be applied in a variety of areas. Moreover, the approach proposed in the present paper does not require building any additional probabilistic models, and it uses a threshold-type classification algorithm instead of the models clustering.

Although there are many papers regarding the general idea of visualization in the fraud detection framework (for example, see [41,9,12]), the visualization meant as a projection from an input high-dimensional space onto an output 2-dimensional space appears relatively rarely in the literature, and the problem apparently has not yet gained the deserved attention.

In the paper [41], the main visualization is established using a spiral, on which the events are drawn according to their timestamp. Suspicious events are considered those which appear along the same radius or on close radii.

The papers [9,12] employ the graph theory in order to establish the desirable visualization. In [9], the volume and the complexity of the collected data, which are modeled as Financial Activity Networks (FANs), whose nodes represent persons or companies, and whose edges represent their connections according to a set of possible criteria. In [12], the trading networks are visualized using node-link diagrams to reveal social relational structures among traders. In the visualization of stock trading network, a node (circle) represents a particular trader and an edge (link) represents the trading relationship.

On the other hand, the works [18,11] consider the data visualization meant as a mapping between a multidimensional space and a 2-dimensional space, and therefore, these papers are of particular interest from the point of view of our research.

In the paper [18], a neural visualization of network traffic data for computer intrusion detection is proposed. The system introduced in [18] applies neural projection architectures to detect anomalous situations taking place in a computer network. By its

Download English Version:

https://daneshyari.com/en/article/405049

Download Persian Version:

https://daneshyari.com/article/405049

Daneshyari.com