Expert Systems with Applications 38 (2011) 13158–13169

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Detecting fraud in online games of chance and lotteries

I.T. Christou [a,b,*], M. Bakopoulos [a,d], T. Dimitriou [a,b], E. Amolochitis [a,d], S. Tsekeridou [a], C. Dimitriadis [c]

[a] Athens Information Technology, 19Km. Markopoulou Ave., P.O. Box 68, Paiania 19002, Greece
[b] Information Networking Institute, Carnegie-Mellon University, Pittsburgh, PA, USA
[c] Intralot S.A., 64 Kifissias Ave., Athens 15125, Greece
[d] Center for TeleInFrastructure (CTiF), Aalborg University (AAU), 9220 Aalborg East, Denmark

ARTICLE INFO

Keywords:
Fraud detection
Data cubes
Money laundering detection
Unsupervised learning
Cluster ensembles
Outlier detector fusion

ABSTRACT

Fraud detection has been an important topic of research in the data mining community for the past two decades. Supervised, semi-supervised, and unsupervised approaches to fraud detection have been proposed for the telecommunications, credit, insurance and health-care industries. We describe a novel hybrid system for detecting fraud in the highly growing lotteries and online games of chance sector. While the objectives of fraudsters in this sector are not unique, money laundering and insider attack scenarios are much more prevalent in lotteries than in the previously studied sectors. The lack of labeled data for supervised classifier design, user anonymity, and the size of the data-sets are the other key factors differentiating the problem from previous studies, and are the key drivers behind the design and implementation decisions for the system described. The system employs *online* algorithms that optimally aggregate statistical information from raw data and applies a number of pre-specified checks against known fraud scenarios as well as novel clustering-based algorithms for outlier detection which are then fused together to produce alerts with high detection rates at acceptable false alarm levels.

1. Introduction

Cyber-crime has been a constant threat for profit as well as non-profit organizations since the beginning of the commercial internet in the mid-nineties. As a response, several approaches to counter-measure cyber-criminal attacks have been proposed from various communities: more powerful authentication and other techniques from the computer and communications security communities aimed at denying falsifying or stealing identities (client or server alike) while intrusion detection and fraud detection techniques aimed at detecting changes from "normal profile behavior", or equivalently, at detecting "anomalous behavior". For anomaly or outlier detection, a wide range of traditional techniques from the machine learning and data mining literature have been proposed, including supervised learning approaches where a set of training data is provided that contains labeled "normal" and "fraudulent" transactions (Koutsoutos, Christou, & Efremidis, 2007; Sherman, 2002) as well as unsupervised approaches (Breunig, Kriegel, Ng, & Sander, 2000; Yamanishi, Takeuchi, & Williams, 2004). So far, the major industries that have taken steps to prevent or at least detect fraud include the insurance, credit card & banking, telecommunications (Fawcett & Provost, 1997), and health-care sectors (Phua, Alahakoon, & Lee, 2004; Phua, Lee, Smith, & Gayler, 2005). However, (online or not) government-run as well as privately-held lotteries and book-makers are also facing very significant threats to their operations by fraudsters attempting to appropriate a portion of the value of the gambles taking place there. In fact, given the very large revenues generated by lotteries and betting companies, detecting and preventing fraud in this sector is a major issue that to our knowledge has not been studied enough.

In the games of chance sector, the most prevalent problem is that of money laundering. While not detrimental to the organization organizing the game of chance or lottery in immediate financial terms (the organization does not lose money), its long-term effects are loss of reputation and the public's belief that the organization is run for the benefit of criminals; such effects can have significant long-term negative effect on the image and profitability of the organization. Therefore, organizations running games of chance should take every possible step to show they are battling money laundering schemes as best as they can.

The second most prevalent problem that an organization running lotteries and betting games faces is that of insider attacks. Authorized agents running terminals for public participation to a lottery have been known to attempt to scam the organization's systems for their own profit; some of the various scenarios they use will be described in Section 3. Finally, large data sizes are very common in the fraud-detection domain, mainly because it is

* Corresponding author at: Athens Information Technology, 19Km. Markopoulou Ave., P.O. Box 68, Paiania 19002, Greece. Tel.: +30 210 668 2725.
E-mail addresses: ichr@ait.edu.gr (I.T. Christou), mbak@ait.edu.gr (M. Bakopoulos), tdim@ait.edu.gr (T. Dimitriou), emam@ait.edu.gr (E. Amolochitis), sots@ait.edu.gr (S. Tsekeridou), dimitriadis@intralot.com (C. Dimitriadis).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.124

through the sheer data size that the perpetrator can hope to hide their criminal intents. In the case of lotteries and betting games however, the volume of data can reach up to 14,000,000 transactions in a single day, and issues such as I/O, space requirements and linear-time complexity of the algorithms employed become major aspects for the design of the system.

1.1. Our contribution

Our contribution starts with the development of highly efficient data structures for keeping sufficient and necessary statistics for performing a combination of statistical tests and cluster analysis in order to detect highly unlikely events in lotteries and betting games, both in the individual user level as well as at the aggregator (agent) level, which make possible the detection of anomalies in the behavior of individual users as well as groups of anonymous users or agents of the lotteries.

Based on an analysis in Section 3 of the major attack scenarios that organizations running lotteries and games of chance face, we have designed and implemented a special-purpose data structure, namely a data-cube inspired from OLAP databases that is used to hold the statistical information necessary to detect deviations in very large data sets. Unlike standard data-cube structures however which usually hold only a single quantity in each cell, we design the cube so as to keep a number of data in each cell, including frequency statistics of various important quantities for fraud detection. Transaction aggregation for the purposes of detecting fraud has been proposed before by Whitrow, Hand, Juszcak, Weston, and Adams (2009), and also very recently by Krivko (2010), but both studies only consider aggregating credit card transactions along a single dimension, namely time, whereas our data-cube aggregates data simultaneously along multiple dimensions, including of course the time dimension. Another major difference between our study and that of Whitrow et al. (2009) and Krivko (2010) is that they deal with supervised learning and design optimized classifiers having a labeled set of transactions to provide a ground-truth model for what constitutes normal or fraudulent activity.

We present novel clustering-based outlier-detection techniques that partition the data-cube into a *large* number of clusters using structural entropy of Sum of Square Errors (SSE) as the clustering criterion and detect smaller-than-expected clusters that are candidate outlier clusters. New techniques for cluster ensemble coordination employed, allow the data-cube to be efficiently partitioned among a large number of clusters in a near-optimal way, avoiding the central limit catastrophe (Baum, 1986) that is known to plague most other clustering algorithms. Experimental results verify that our new techniques produce much better detection rates for a given false-alarm rate than standard techniques such as Local Outlying Factor (Breunig et al., 2000) as discussed in Section 7.

Another major characteristic of our system not found in others' works is that it works with batch – anonymous – data coming from agents terminals, as well as with individual – identified – user transactions, in which case, individual user profiles are built and monitored in near real-time to detect large deviations from the "expected" behavior. We discuss in detail both modes of system operation.

The rest of this paper is organized as follows: in Section 2 we present a detailed review of related work that has been carried out during the past two decades. In Section 3, we described the major attack scenarios an organization running lotteries and games of chance faces and an analysis of these threats, and in Section 4 we describe the data-cube data-structure designed to handle very large volumes of data. Then, in Section 5, we present the suite of algorithms that we have implemented to detect anomalies in the transactions processed by the system. We show the user interaction with the system and the visualization of abnormalities detected in Section 6. In Section 7 we present the results of running the system on both benchmark data-sets as well as on real-world data-sets provided by an international organization running lotteries and games of chance. Then, we present our conclusions and a list of future directions this research will take in the near future.

2. Related work

Detecting anomalies in data is a topic that has intrigued researchers as early as 1777 (Beckman and Cook (1983) trace the topic's origins in comments made by Bernoulli at that time). Books and monographs covering the subject of outlier detection in statistical data include (Barnett & Lewis, 1994; Hawkins, 1980).

An early approach to detecting outliers in uni-variate data assumed the data come from a known underlying distribution. Probability-based tests can be developed to determine the probability $p(x)$ of any data point, and consequently, to consider any data point with very low probability, an outlier. Detecting outliers in multivariate data is the subject of the studies in (Davies & Gather, 1993; Rocke & Woodruff, 1996). Other statistical techniques based on the χ^2 test are discussed in Ye and Chen (2000).

Other unsupervised approaches to anomaly-detection include clustering-based methods and distance-based methods. The easiest distance-based method (that has a quadratic complexity in the number of data-points though) is to consider for each point, the distance of its kth nearest neighbor to be its degree of "outlier-ness", or alternatively, the average distance to its first k nearest neighbors (Tan, Steinbach, & Kumar, 2006). When the distributions of the data-points are known, the Mahalanobis distance provides a more accurate measure of distance between the data-points than the standard Euclidean distance.

In clustering-based techniques, the idea is that outliers will either form single-point clusters, or will be far from the center of the cluster to which they belong. Such techniques ought to be used with high values of k – the number of clusters sought – which however entails the danger of "central limit catastrophe" (Baum, 1986), which refers to the fact that local improvement based techniques for clustering such as K-Means usually get trapped in very low-quality local minima when the number of clusters sought is high compared to the number of points in the data set. In density-based techniques, a density-aware measure of similarity between objects is used to avoid declaring outliers points that naturally form low-density clusters. The Local Outlying Factor technique (Breunig et al., 2000) is a powerful technique for assigning degrees of "outlier-ness" to points in a data-set that is robust in the presence of clusters of widely differing densities.

When there exist labeled transaction records that can serve as training records for learning a model of fraud and/or normal behavior, supervised classification methods can be used to build appropriate classifiers trained on the available training set that predict whether a new transaction is fraudulent or normal. Since the fraudulent transactions are typically far less than the normal transactions, special techniques must be used to deal with this inherent skew of the data (Phua et al., 2004). In the credit card industry in particular, aggregating transactions over a short period of time for each account (using a period length between one and three days) has been tested with good success by Whitrow et al. (2009). However, they aggregate information only along a single dimension, and they do it so as to build a classifier via standard supervised learning techniques. Their classifier directly attempts to optimize a score other than the Area-Under-the-Curve (AUC) metric or the True-Positive over False Positives ratio as the authors explicitly take into account the fact that a type I (False Negative)