DETECTION OF FRAUDULENT ACTIVITIES IN ONLINE COMMUNITIES

By

Mostafizur Rahman & Mujahidul Islam

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE
AT
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
DHAKA 1000
2015

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical and Electronic Engineering for acceptance a thesis entitled "**Detection of Fraudulent Activities in Online Communities**" by **Mostafizur Rahman & Mujahidul Islam** in partial fulfillment of the requirements for the degree of **Bachelor Of Science**.

Dated: 2015

Supervisor: _____
Dr. S. M. Farhad

Readers: _____

_____

To all the people who works for open sources projects out of no personal benefit

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Online communities have become very popular in the recent times. They have become incredibly helpful in sharing knowledge, ideas, skills or even experiences nowadays. People ask questions on different topics and others who have knowledge on that topic reply to those questions. Thus a knowledge sharing culture has already grown and it's getting a proper shape slowly. A number of QA sites and we based discussion forum are being widely used by millions of users across the world. So it's important to make sure that the knowledge base is not corrupted. In this paper we will introduce major fradulant activities which are frequently found in online communities and also propose different approaches to classify those fradulant activities properly.

1.1 Motiovation

There is no doubt that the online communities are playing a vital role in learning new things and sharing knowledge. But there is no guaranty that the answer someone has given is absolutely correct. There must be strong logic and references to the answer to be a good one. Thus, awarding some points and other rewards like badges, medals and other cool stuffs to the answerer has become a popular mean to encourage to

provide more quality answers. This has somewhat established a trend that the more points and other rewards a person has, the more he knows.

One of the most widely used

Chapter 2

Related Works

This section examines major and commonly used techniques and algorithms.

2.1 Overview

Figure [**?**] shows that many existing fraud detection systems typically operate by adding fraudulent claims/applications/ transactions/accounts/sequences (A) to black lists to match for likely frauds in the new instances (E). Some use hard-coded rules which each transaction should meet such as matching addresses and phone numbers, and price and amount limits (Sherman, 2002).

Below outlines the complex nature of data used for fraud detection in general (Fawcett, 2003; 1997):

- Volume of both fraud and legal classes will fluctuate independently of each other; therefore class distributions (proportion of illegitimate examples to legitimate examples) will change over time.

- Multiple styles of fraud can happen at around the same time. Each style can have a regular, occasional, seasonal, or onceoff temporal characteristic.

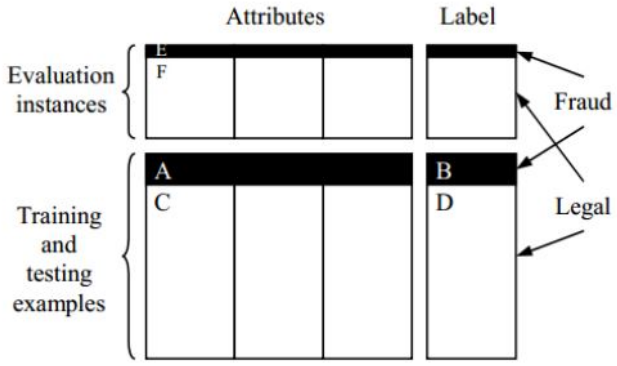

Figure 2.1: Structured diagram of the possible data for analysis. Data mining approaches can utilise training/testing data with labels, only legal examples, and no labels to predict/describe the evaluation data.

- Legal characteristics/behaviour can change over time.

1) Labelled training data (A + B + C + D) can be processed by single supervised algorithms. A better suggestion is to employ hybrids such as multiple supervised algorithms , or both supervised and unsupervised algorithms to output suspicion scores, rules and/or visual anomalies on evaluation data.

2) All known legal claims/applications/transactions/accounts/ sequences (C) should be used processed by semi-supervised algorithms to detect significant anomalies from consistent normal behaviour.

3) Combining training data (the class labels are not required here) with evaluation data (A + C + E + F). These should be processed by single or multiple unsupervised algorithms to output suspicion scores, rules and/or visual anomalies on evaluation data.

2.2 Supervised Approaches on Labelled Data (A + B + C + D)

Predictive supervised algorithms examine all previous labelled transactions to mathematically determine how a standard fraudulent transaction looks like by assigning a risk score (Sherman, 2002). Neural networks are popular and support vector machines (SVMs) have been applied. Ghosh and Reilly (1994) used a three-layer, feed-forward Radial Basis Function (RBF) neural network with only two training passes needed to produce a fraud score in every two hours for new credit card transactions. Barse et al (2003) used a multi-layer neural network with exponential trace memory to handle temporal dependencies in synthetic Video-on-Demand log data. Syeda et al (2002) propose fuzzy neural networks on parallel machines to speed up rule production for customer-specific credit card fraud detection.

The neural network and Bayesian network comparison study (Maes et al, 2002) uses the STAGE algorithm for Bayesian networks and backpropagation algorithm for neural networks in credit transactional fraud detection. Comparative results show that Bayesian networks were more accurate and much faster to train, but Bayesian networks are slower when applied to new instances.

Ezawa and Norton (1996) developed Bayesian network models in four stages with two parameters. They argue that regression, nearest-neighbour, and neural networks are too slow and decision

Statistical modelling such as regression has been extensively utilised. Foster and Stine (2004) use least squares regression and stepwise selection of predictors to show that standard statistical methods are competitive. Their version of fully automatic stepwise regression has three useful modifications: firstly, organises calculations

to accommodate interactions; secondly,exploits modern decision-theoretic criteria to choose predictors; thirdly, conservatively estimate p-values to handle sparse data and a binary response before calibrating regression predictions. If cost of false negative is much higher than a false positive, their regression model obtained significantly lesser misclassification costs than C4.5 for telecommunications bankruptcy prediction

2.3 Hybrid Approaches with Labelled Data

Popular supervised algorithms such as neural networks, Bayesian networks, and decision trees have been combined or applied in a sequential fashion to improve results. Chan et al (1999) utilises naive Bayes, C4.5, CART, and RIPPER as base classifiers and stacking to combine them. They also examine bridging incompatible data sets from different companies and the pruning of base classifiers. The results indicate high cost savings and 7 better efficiency on credit card transactions. Phua et al (2004) proposes backpropagation neural networks, naive Bayes, and C4.5 as base classifiers on data partitions derived from minority oversampling with replacement. Its originality lies in the use of a single meta-classifier (stacking) to choose the best base classifiers, and then combine these base classifiers predictions (bagging) to produce the best cost savings on automobile insurance claims.

Also, He et al (1999) propose genetic algorithms to determine optimal weights of the attributes, followed by k-nearest neighbour algorithm to classify the general practitioner data. They claim significantly better results than without feature weights and when compared to CBR.

2.4 Unsupervised Approaches with Unlabelled Data (A + C + E + F)

Link analysis and graph mining are hot research topics in antiterrorism, law enforcement, and other security areas, but these techniques seem to be relatively under-rated in fraud detection research. A white paper (NetMap, 2004) describes how the emergent group algorithm is used to form groups of tightly connected data and how it led to the capture of an actual elusive fraudster by visually analysing twelve months worth of insurance claims. There is a brief application description of a visual telecommunications fraud detection system (Cox, 1997) which flexibly encodes data using colour, position, size and other visual characteristics with multiple different views and levels. The intuition is to combine human detection with machine computation.

Cortes et al (2001) examines temporal evolution of large dynamic graphs for telecommunications fraud detection. Each graph is made up of subgraphs called Communities Of Interest (COI). To overcome instability of using just the current graph, and storage and weightage problems of using all graphs at all time steps; the authors used the exponential weighted average approach to update subgraphs daily. By linking mobile phone accounts using call quantity and durations to form COIs, the authors confirm two distinctive characteristics of fraudsters. First, fraudulent phone accounts are linked - fraudsters call each other or the same phone numbers. Second, fraudulent call behaviour from flagged frauds are reflected in some new phone accounts - fraudsters retaliate with application fraud/identity crime after being detected. Cortes et al (2003) states their contribution to dynamic graph research in the areas of scale, speed, dynamic updating, condensed representation of the graph, and measure direct interaction between nodes.

Some forms of unsupervised neural networks have been applied. Dorronsoro et al (1997) creates a non-linear discriminant analysis algorithm which do not need labels. It minimises the ratio of the determinants of the within and between class variances of weight projections. There is no history on each credit card accounts past transactions, so all transactions have to be segregated into different geographical locations. The authors explained that the installed detection system has low false positive rates, high cost savings, and high computational efficiency. Burge and ShaweTaylor (2001) use a recurrent neural network to form short-term and long-term statistical account behaviour profiles. Hellinger distance is used to compare the two probability distributions and give a suspicion score on telecommunications toll tickets.

In addition to cluster analysis , unsupervised approaches such as outlier detection, spike detection, and other forms of scoring have been applied. Yamanishi et al (2004) demonstrated the unsupervised SmartSifter algorithm which can handle both categorical and continuous variables, and detect statistical outliers using Hellinger distance, on medical insurance data.

Bolton and Hand (2001) recommend Peer Group Analysis to monitor inter-account behaviour over time. It compares the cumulative mean weekly amount between a target account and other similar accounts (peer group) at subsequent time points. The distance metric/suspicion score is a t-statistic which determines the standardised distance from the centroid of the peer group. The time window to calculate peer group is thirteen weeks and future time window is four weeks on credit card accounts.

Bolton and Hand (2001) also suggest Break Point Analysis to monitor intraaccount behaviour over time. It detects rapid spending or sharp increases in weekly spending within a single account. Accounts are ranked by the t-test. The fixed-length moving

transaction window contains twenty-four transactions: first twenty for training and next four for evaluation on credit card accounts. Brockett et al (2002) recommends Principal Component Analysis of RIDIT scores for rank-ordered categorical attributes on automobile insurance data.

Hollmen and Tresp (1998) present an experimental real-time fraud detection system based on a Hidden Markov Model (HMM).

Appendices

Appendix A

Further Study

A.1 Perfusion Index