A FRAUD DETECTION ALGORITHM FOR ONLINE BANKING

Delio Panaro [1], Eva Riccomagno[2] and Fabrizio Malfanti[3]

[1] iBe Think Solve, Execute (e-mail: d.panaro@be-tse.it)

[2] Department of Mathematics, University of Genoa
(e-mail: riccomag@dima.unige.it)

[3] Intelligrate s.r.l. (e-mail: fabrizio.malfanti@intelligrate.it)

ABSTRACT: A general two layer statistical classifier for sensitive, highly skewed, massive data sets is developed by combining simple classifiers with an Adaboost meta-classifier. The algorithm has been inspired by the necessity of analysing a data set of about fifteen million real world online banking transactions, spanning from the year 2011 to 2013 with the aim of distinguishing frauds from legitimate operations. Each data point is recorded as server log track and there are about 300 million single entries. Results show that the algorithm is particularly effective in detecting anomalies, achieving high true positive rates and reasonably low false positive rates. The algorithm has also been validated on the NSL-KDD data set in order to compare his performances versus classical classification algorithms and on shufflings of the online banking data set.

KEYWORDS: fraud detection, statistical classifier, SVM, Adaboost

1 Introduction

This work proposes a two layers algorithm for supervised binary classification. The algorithm is specifically designed for anomaly detection and, in particular, fraud detection.

The algorithm is composed by a combination of machine learning algorithms that allows classification accuracy and a very short elaboration time. For a similar approach in other context see for example **?**. The real-time feature sets the algorithm in between fraud prevention and fraud detection. The algorithm has been designed to be able to manage a wide range of problems. It is particularly effective on unbalanced data sets, with asymmetrical cost functions and in which the detection of true positives proves to be hard.

The algorithm has been motivated by a project related to a fraud detection in online bank and has been tested on a data set of online bank transfers pro-

vided by the industrial partner of the project. Main characteristics of the data set are size and imbalance between licit and fraudulent operations.

2 Sample features and data preprocessing

The data set at our disposal is composed by $14,967,432$ bank server logs stored in a MySQL database. It records all operations involving money transfer from three major Italian banks and covers the period from January 1st, 2011 to May 31st, 2013. For each single log there are eighteen entries containing information such as IP address, username, its date and time of occurence and so on.

The share of operations labelled as frauds is about 1 in $25,000$ giving rise to a strong imbalance. More precisely, there are $14,966,796$ licit operations against 636 fraudulent operations.

3 Classifier architecture

The proposed classifier is based on two layers: the first layer is composed by simple classifiers, whereas the second layer is based on an Adaboost Meta Classifier (**?**).

The two layer architecture is motivated by the need to avoid over-fitting problems encountered with classical Adaboost algorithm. Furthermore a multi layer architecture allows the reduction of computational effort and improved performances. For similar issues on a different multi-layer architecture see **?**.

Regardless how has been constructed , a simple classifier is allocated to the first or second layer of the meta-classifier according to its accuracy defined as the number of correctly classified instances over the number of processed instances. Poorly performing classifiers are discarded, best performing classifiers are allocated to the first layer and the other ones to the second layer (see Fig.1).

Meta-algorithm classifying performances have been tested using two kinds of simple classifiers: Support Vector Machines (SVMs) and *Behavioural*.

4 Preliminary results

Preliminary results show that the algorithm is particularly effective in detecting anomalies, achieving high true positive rates and reasonably low false positive rates.

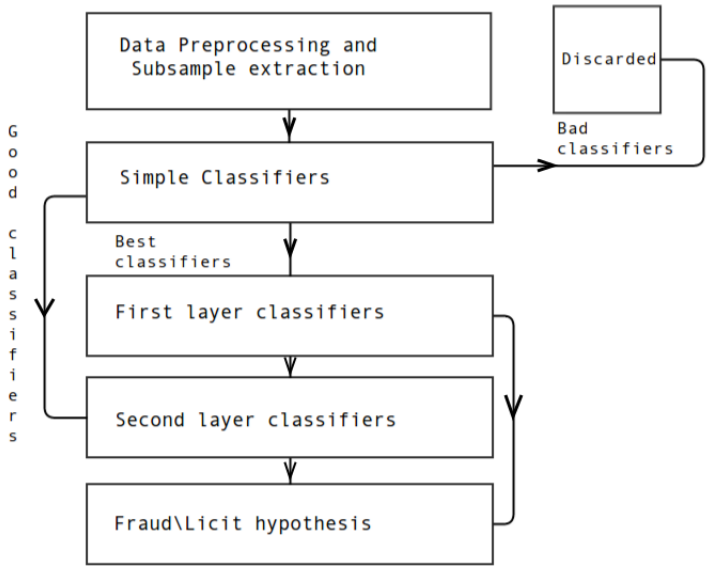

Figure 1. *Classifier working scheme*

The algorithm has also been validated on the NSL-KDD data set in order to compare his performances versus classical classification algorithms and on a *synthetic* dataset built shuffling rows of the original online banking data set.

References

BOLTON, RICHARD J, & HAND, DAVID J. 2002. Statistical fraud detection: A review. *Statistical Science*, 235–249.

CHAN, PHILIP K, FAN, WEI, PRODROMIDIS, ANDREAS L, & STOLFO, SALVATORE J. 1999. Distributed data mining in credit card fraud detection. *Intelligent Systems and their Applications*, **14**(6), 67–74.

CORTES, CORINNA, & VAPNIK, VLADIMIR. 1995. Support-vector networks. *Machine Learning*, **20**(3), 273–297.

FAN, WEI, STOLFO, SALVATORE J, & ZHANG, JUNXIN. 1999. The application of AdaBoost for distributed, scalable and on-line learning. *Pages 362–366 of: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

FREUND, YOAV, & SCHAPIRE, ROBERT E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. *Pages 23–37 of: Computational learning theory*. Springer.

FREUND, YOAV, SCHAPIRE, ROBERT, & ABE, N. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, **14**(1612), 771–780.

LI, XUCHUN, WANG, LEI, & SUNG, ERIC. 2005. A study of AdaBoost with SVM based weak learners. *Pages 196–201 of: Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 1. IEEE.

PHUA, CLIFTON, LEE, VINCENT, SMITH, KATE, & GAYLER, ROSS. 2010. A comprehensive survey of data mining-based fraud detection research. *Pages 50–53 of: Intelligent Computation Technology and Automation (ICICTA)*, vol. 1. IEEE.

ROJAS, RAÚL. 2009. AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep*.

VIOLA, PAUL, & JONES, MICHAEL J. 2004. Robust real-time face detection. *International journal of computer vision*, **57**(2), 137–154.

WU, XINDONG, & KUMAR, VIPIN EDS. 2009. *Top 10 algorithms in data mining*. Data Mining and Knowledge Discovery. Boca Raton: Chapman & Hall/CRC.

ZHU, JI, ROSSET, SAHARON, ZOU, HUI, & HASTIE, TREVOR. 2006. Multi-class adaboost. *Ann Arbor*, **1001**(48109), 1612.