Minor Project Synopsis

A Combined approach of SMOT data balancing with data mining algorithms for detecting credit card fruad

Delhi Technological University

(Formerly Delhi College of Engineering)

Shahbad Daulatpur,

Delhi-110042

Submitted to –	Submitted by –
Ms Indu Singh	Ajit Singh Kushwaha (2K14/CO/006)
Assistant Professor	Vaibhav Kashyap (2K14/CO/135)
CSE Deptt, DTU	Vivek (2K14/CO/144)
	Bablu Singh (2K14/CO/020)

Approval Certificate

THIS IS CERTIFYING THAT I APPROVED THE PROPOSAL OF MINOR PROJECT ON TOPIC “**A COMBINED APPROACH OF SMOT DATA BALANCING WITH DATA MINING ALGORITHMS FOR DETECTING CREDIT CARD FRAUD**”.

Project Mentor

Ms. Indu Singh
Assistant Professor
CSE Deptt. DTU

Abstract

Credit card fraud is a serious and growing problem. While predictive models for credit card fraud detection are in active use in practice, reported studies on the use of data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. This paper evaluates two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study is based on real-life data of transactions from an international credit card operation.

Billions of dollars are lost annually due to credit card fraud. The 10th annual online fraud report by CyberSource shows that although the percentage loss of revenues has been a steady 1.4% of online payments for the last three years (2006 to 2008), the actual amount has gone up due to growth in online sales. The estimated loss due to online fraud is $4 billion for 2008, an increase of 11% on the 2007 loss of $3.6 billion. With the growth in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud, and 70% of U.S. consumers are noted to be significantly concerned about identity fraud. Additionally, credit card fraud has broader ramifications, as such fraud helps fund organized crime, international narcotics trafficking, and even terrorist financing. Over the years, along with the evolution of fraud detection methods, perpetrators of fraud have also been evolving their fraud practices to avoid detection. Therefore, credit card fraud detection methods need constant innovation. In this study, we evaluate two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. The study is based on real-life data of transactions from an international credit card operation.

Objective

The objective of this study is to examine the performance of two advanced data mining techniques, random forests and support vector machines, together with the well-known logistic regression, for credit card fraud identification. We also want to compare the effect of extent of data undersampling on the performance of these techniques. This section describes the data used for training and testing the models and performance measures used.

Introduction

Credit card fraud is a wide-ranging term for theft and fraud committed using or involving a payment card, such as a credit card or debit card, as a fraudulent source of funds in a transaction. The purpose may be to obtain goods without paying, or to obtain unauthorized funds from an account. Credit card fraud is also an adjunct to identity theft. According to the United States Federal Trade Commission, while the rate of identity theft had been holding steady during the mid 2000s, it increased by 21 percent in 2008. However, credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints for the sixth year in a row.

Although incidence of credit card fraud is limited to about 0.1% of all card transactions, this has resulted in huge financial losses as the fraudulent transactions have been large value transactions. In 1999, out of 12 billion transactions made annually, approximately 10 million—or one out of every 1200 transactions—turned out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts were fraudulent. Even with tremendous volume and value increase in credit card transactions since then, these proportions have stayed the same or have decreased due to sophisticated fraud detection and prevention systems. Today's fraud detection systems are designed to prevent one twelfth of one percent of all transactions processed which still translates into billions of dollars in losses.

Card fraud begins either with the theft of the physical card or with the compromise of data associated with the account, including the card account number or other information that would routinely and necessarily be available to a merchant during a legitimate transaction. The compromise can occur by many common routes and can usually be conducted without tipping off the card holder, the merchant, or the issuer at least until the account is ultimately used for fraud. A simple example is that of a store clerk copying sales receipts for later use. The rapid growth of credit card use on the Internet has made database security lapses particularly costly; in some cases, millions of accounts have been compromised.

Stolen cards can be reported quickly by cardholders, but a compromised account can be hoarded by a thief for weeks or months before any fraudulent use, making it difficult to identify the source of the compromise. The cardholder may not discover fraudulent use until receiving a billing statement, which may be delivered infrequently. Cardholders can mitigate this fraud risk by checking their account frequently to ensure constant awareness in case there are any suspicious, unknown transactions or activities.

With the growth in credit card transactions, as a share of the payment system, there has also been an increase in credit card fraud, and 70% of U.S. consumers are noted to be significantly concerned about identity fraud. Additionally, credit card fraud has broader ramifications, as such fraud helps fund organized crime, international narcotics trafficking, and even terrorist financing. In this study, we evaluate two advanced data mining approaches, support vector machines and random forests, together with the well-known logistic regression, as part of an attempt to better detect (and thus control and prosecute) credit card fraud. Statistical fraud detection methods have been divided into two broad categories: supervised and unsupervised.

In supervised fraud detection methods, models are estimated based on the samples of fraudulent and legitimate transactions, to classify new transactions as fraudulent or legitimate. In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both these fraud detection methods predict the probability of fraud in any given transaction. In supervised fraud detection methods, models are estimated based on the samples of fraudulent and legitimate transactions, to classify new transactions as fraudulent or legitimate. In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both these fraud detection methods predict the probability of fraud in any given transaction.

Credit card fraud is essentially of two types: application and behavioural fraud. Application fraud is where fraudsters obtaining new cards from issuing companies using false information or other people's information. Behavioural fraud can be of four types: mail theft, stolen/lost card, counterfeit card and ‘card holder not present’ fraud. Mail theft fraud occurs when fraudsters intercept credit cards in mail before they reach cardholders or pilfer personal information from bank and credit card statements. Stolen/lost card fraud happens when fraudsters get hold of credit cards through theft of purse/wallet or gain access to lost cards. However, with the increase in usage of online transactions, there has been a significant rise in counterfeit card and ‘card holder not present’ fraud. In both of these two types of fraud, credit card details are obtained without the knowledge of card holders and then either counterfeit cards are made or the information is used to conduct ‘card holder not present’ transactions, i.e. through mail, phone, or the Internet.

Data-mining techniques

As stated above, we investigated the performance of three techniques in predicting fraud: Logistic Regression (LR), Support Vector Machines (SVM), and RandomForest (RF). In the paragraphs below, we briefly describe the three techniques employed in this study.

Logistic regression

Qualitative response models are appropriate when dependent variable is categorical. In this study, our dependent variable fraud is binary, and logistic regression is a widely used technique in such problems. Binary choice models have been used in studying fraud. For example, used binary choice models in the case of insurance frauds to predict the likelihood of a claim being fraudulent. In case of insurance fraud, investigators use the estimated probabilities to flag individuals that are more likely to submit a fraudulent claim.

Support vector machines

Support vector machines (SVMs) are statistical learning techniques that have been found to be very successful in a variety of classification tasks. Several unique features of these algorithms make them especially suitable for binary classification problems like fraud detection. SVMs are linear classifiers that work in a high-dimensional feature space that is a non-linear mapping of the input space of the problem at hand. An advantage of working in a high-dimensional feature space is that, in many problems the non-linear classification task in the original input space becomes a linear classification task in the high-dimensional feature space. SVMs work in the high dimensional feature space without incorporating any additional computational complexity. The simplicity of a linear classifier and the capability to work in a feature-rich space make SVMs attractive for fraud detection tasks where highly unbalanced nature of the data (fraud and non-fraud cases) make extraction of meaningful features critical to the detection of fraudulent transactions is difficult to achieve. Applications of SVMs include bioinformatics, machine vision, text categorization, and time series analysis. The strength of SVMs comes from two important properties they possess — kernel representation and margin optimization. In SVMs, mapping to a high-dimensional feature space and learning the classification task in that space without any additional computational complexity are achieved by the use of a kernel function.

Random forests

The popularity of decision tree models in data mining arises from their ease of use, flexibility in terms of handling various data attribute types, and interpretability. Single tree models, however, can be unstable and overly sensitive to specific training data. Ensemble methods seek to address this problem by developing a set of models and aggregating their predictions in determining the class label for a data point. A random forest model is an ensemble of classification (or regression) trees. Ensembles perform well when individual members are dissimilar, and random forests obtain variation among individual trees using two sources for randomness: first, each tree is built on separate bootstrapped samples of the training data; secondly, only a randomly selected subset of data attributes is considered at each node in building the individual trees. Random forests thus combine the concepts of bagging, where individual models in an ensemble are developed through sampling with replacement from the training data, and the random subspace method, where each tree in an ensemble is built from a random subset of attributes.

Given a training data set of N cases described by B attributes, each tree in the ensemble is developed as follows:

- Obtain a bootstrap sample of N cases
- At each node, randomly select a subset of bbB attributes.
- Determine the best split at the node from this reduced set of b
- attributes
- Grow the full tree without pruning

k - Nearest Neighbours

In pattern recognition, the ***k*-nearest neighbours algorithm** (***k*-NN**) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbour.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

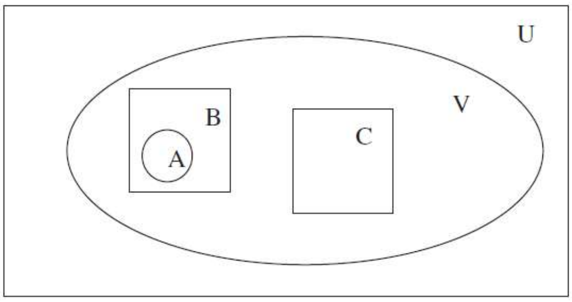
Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor.[2]

The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Data

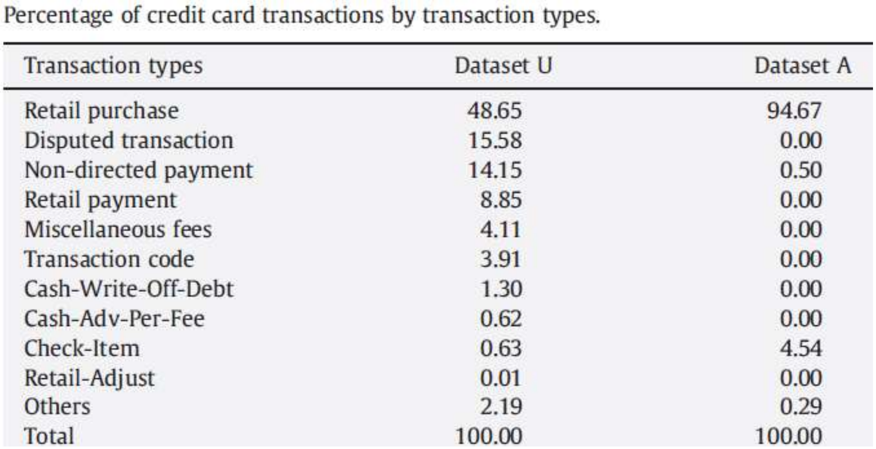
This section describes the real-life data on credit card transactions and how it is used in our study. It also describes the primary attributes in the data and the derived attributes created.

In this study we use the dataset which was obtained from an international credit card operation. The previous study used Artificial Neural Networks (ANN) tuned by Genetic Algorithms (GAs) to detect fraud. This dataset has 13 months, from January 2006 to January 2007, of about 50 million (49,858,600 transactions) credit card transactions on about one million (1,167,757 credit cards) credit cards from a single country. For the purpose of this study, we call this dataset of all transactions, dataset U. A much smaller subset of this large dataset is dataset A, which has 2420 known fraudulent transactions with 506 credit cards.



Dataset U: All transactions	49,858,600 transactions
Dataset A: Fraud Dataset	2,420 transactions
Dataset B: All transactions with Fraudulent Credit Cards	37,280 transactions
Dataset V:All transactions with transaction types where *fraud* occurred	31,671,185 transactions
Dataset C: Random sample of transactions from dataset V-B	340,589 transactions

Fig. 1. Dataset description.



Percentage of credit card transactions by transaction types.

Transaction types	Dataset U	Dataset A
Retail purchase	48.65	94.67
Disputed transaction	15.58	0.00
Non-directed payment	14.15	0.50
Retail payment	8.85	0.00
Miscellaneous fees	4.11	0.00
Transaction code	3.91	0.00
Cash-Write-Off-Debt	1.30	0.00
Cash-Adv-Per-Fee	0.62	0.00
Check-Item	0.63	4.54
Retail-Adjust	0.01	0.00
Others	2.19	0.29
Total	100.00	100.00

References

IDE and Interface used –

- R – 3.4.0 ; https://www.r-project.org/about.html ; https://cran.r-project.org/
- RStudio – 1.0.143 ; https://www.rstudio.com/

R Packages used –

- caret - https://cran.r-project.org/web/packages/caret/index.html
- ggplot2 - https://cran.r-project.org/web/packages/ggplot2/index.html
- DMwR - https://cran.r-project.org/web/packages/DMwR/index.html\
- ROSE - https://cran.r-project.org/web/packages/ROSE/index.html
- e1071 - https://cran.r-project.org/web/packages/e1071/index.html
- randomForest - https://cran.r-project.org/web/packages/randomForest/
- plotly - https://cran.r-project.org/web/packages/plotly/
- stats - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- cluster - https://cran.r-project.org/web/packages/cluster/
- caTools - https://cran.r-project.org/web/packages/caTools/index.html
- data.table - https://cran.r-project.org/web/packages/data.table/index.html

Idea and Concepts -

- S. Kotsiantis, D. Kanellopoulos, P. Pintelas (2006). Handling imbalanced datasets: A review. International Transactions on Computer Science and Engineering.
- G.H. John, P. Langley (1995). Estimating continuous distributions in Bayesian classifiers. in: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, (1995); 338 — 345.
- I3I R.J. Bolton, D.J. Hand (2001). Unsupervised profiling methods for fraud detection. In Conference on credit scoring and credit control, Edinburgh.
- D. Kibler. D.W. Aha, M. Albeit (1989). Instance-based prediction of real-valued attributes. Computational Intelligent
- P.K. Chan, W. Fan, A.L. Prodromidis. S.J. Stolfo (1999). Distributed Data Mining in Credit Card Fraud Detection. IEEE Intelligent Systems. pp 67—74.
- C. Cortes, V. Vapnik (1995). Support vector networks. Machine Learning. 20:273—297.
- T.M. Cover, P.E. Hart (1967). Nearest neighbor pattern classification. IEEE Trans. Information Theory, l3(1):21—27.
- G. Potamilis (2013). Design and Implementation of a Fraud Detection Expert System using Ontology-Based Techniques. A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences.
- E. David (2012). Bayesian inference-the future of online fraud protection. Computer Fraud & Security, 8-11.

- S. Ghosh, D.L. Reilly. (1994). Credit Card Fraud Detection with a Neural- Network. In Proceedings of the International Conference on System Science, pages 621-630.
- Jha. Sanjeev, G. Montserrat, J.C. Westland (2012). Employing transaction aggregation strategy to detect credit card fraud. Expert system with application, 39: 12650-12657.
- L. Breiman - Random forests. Machine Learning. (2001). Vol (45); 5—32.
- J. Piotr., AM. Niall. J.D. Hand, C. Whitrow, J. David (2008). Off the peg and bespoke classifiers for fraud detection. Computationa Statistics and Data Analysis, 52
- L. Qibei & J. Chunhua (2011). Research on Credit Card Fraud Detection Model Based on Class Weighted Support Vector Machine. Journal of Convergence Information Technology, 6(1). 62-68.
- S. Maes, K. Tuyls, B.Vanschoenwinkel, B.Manderick (1993). Credit card fraud detection using Bayesian and neural networks. In Proceedings for the First International NAISO Congress on Nettro Fuzzy Technologies, pages 261-270.
- E.W.T. Ngai, H.Yong., Y.H.Wong, Y.Chen, X. Sun (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50:559—569
- M. Zareapoor, Seeja, K.R. & M. Alam, Afshar (2012). Analyzing Credit Card Fraud Detection Techniques Based On Certain Design Criteria. International Journal of Computer Application. 52(3):35-42.
- E.Aleskerov,B.Freisleben, B.Rao, CARDWATCH: a neural network based database mining system for credit card fraud detection, in computational intelligence for financial engineering, Proceedings of the IEEE/IAFE, IEEE, Piscataway, NJ, 1998, pp. 220–226.
- M. Artis, M. Ayuso, M. Guillen, Detection of automobile insurance fraud with discrete choice models and misclassified claims, The Journal of Risk and Insurance 69 (3) (2002) 325–340.
- R.J. Bolton, D.J. Hand, Unsupervised profiling methods for fraud detection, Conference on Credit Scoring and Credit Control, Edinburgh, 2001.
- R.J. Bolton, D.J. Hand, Statistical fraud detection: a review, Statistical Science 17 (3) (2002) 235–249.
- R. Brause, T. Langsdorf, M. Hepp, Neural data mining for credit card fraud detection, Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, 1999, pp. 103–106.
- L. Breiman, Random forest, Machine Learning 45 (2001) 5–32. [7] C.R. Bollinger, M.H. David, Modeling discrete choice with response error: food stamp participation, Journal of the American Statistical Association 92 (1997) 827–835.
- Capital One Identity theft guide for victims, retrieved January 10,2009, from http: //www.capitalone.com/fraud/ID Theft Package V012172004We.pdf? \linkid=WWW_Z_Z_Z_FRD_D1_01_T_FIDTP.
- R. Caruana, N. Karampatziakis, A. Yessenalina, An Empirical Evaluation of Supervised Learning in High Dimensions, in: Proceedings of the 25th international Conference on Machine Learning Helsinki, Finland, July, 2008.

- R. Caruana, A. Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, June, 2006.
- S.B.Caudill, M.Ayuso, M.Guillen, Fraud detection using a multinomial logit model with missing information, The Journal of Risk and Insurance 72 (4) (2005) 539–550.
- P.K. Chan, W. Fan, A.L. Prodromidis, S.J. Stolfo, Distributed Data Mining in Credit Card Fraud Detection, Data Mining, (November/December), 1999, pp. 67–74.
- N.V. Chawla, N. Japkowicz, A. Kotcz, Editorial: special issue on learning from imbalanced data sets, ACM SIGKDD Explorations Newsletter 6 (1) (2004).
- R.C. Chen, T.S. Chen, C.C. Lin, A new binary support vector system for increasing detection rate of credit card fraud, International JournalofPatternRecognition20 (2) (2006) 227–239.
- C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Technical Report 666, University of California at Berkeley, Statistics Department, 2004.
- N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
- Cyber Source. Online fraud report: online payment, fraud trends, merchant practices, and bench marks, retrieved January 8,2009, from http://www.cybersource.com.
- T.G. Dietterich, Ensemble learning, in: M.A. Arbib (Ed.), The Handbook of Brain Theoryand Neural Networks, Second edition, The MIT Press, Cambridge,MA,2002.
- J.R.Dorronsoro, F.Ginel, C.Sanchez, C.Santa Cruz, Neural fraud detection in credit card operations, IEEE Transactions on Neural Networks 8 (1997) 827–834.
- C.Everett, Credit Card Fraud Funds Terrorism, Computer Fraudand Security, May, 1, 2009.
- FairIssac. Falcon Fraud Manager, retrieved January 8, 2009, http://www.fairisaac. com/ficx/Products/dmapps/Falcon-Fraud-Manager.html.
- S. Ghosh, D.L. Reilly, Credit card fraud detection with a neural-network, in: J.F. Nunamaker, R.H. Sprague (Eds.), Proceedings of the 27th Annual Hawaii International Conference on System Science, Vol 3, Information Systems: DSS/ Knowledge-based Systems, Los Alamitos, CA, USA, 1994.
- J.A. Hausman, J. Abrevaya, F.M. Scott-Morton, Misclassification of a dependent variable in a discrete-response setting, Journal of Econometrics87(1998)239–269.
- D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd Ed, Wiley Interscience, 2000.
- J.V. Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th International Conference on Machine Learning, Corvallis, Oregon, June, 2007.
- Y. Jin, R.M. Rejesus, B.B. Little, Binary choice models for rare events data: a crop insurance fraud application, Applied Economics 37 (7) (2005) 841–848.