Adversarial learning in credit card fraud detection

Conference Paper · April 2017

DOI: 10.1109/SIEDS.2017.7937699

6 authors, including:

Aksheetha Sridhar
University of Virginia
1 PUBLICATION **71** CITATIONS
SEE PROFILE

Stephen Adams
Virginia Tech (Virginia Polytechnic Institute and State University)
93 PUBLICATIONS **1,441** CITATIONS
SEE PROFILE

Donald Brown
University of Virginia
370 PUBLICATIONS **7,926** CITATIONS
SEE PROFILE

Peter Adam Beling
University of Virginia
241 PUBLICATIONS **2,719** CITATIONS
SEE PROFILE

Adversarial Learning in Credit Card Fraud Detection

Mary Frances Zeager, Aksheetha Sridhar, Nathan Fogal, Stephen Adams,
Donald E. Brown, and Peter A. Beling
University of Virginia, mcz4ug, as8zb, nf4mp, sca2c, deb, pb3a@virginia.edu

***Abstract* - Credit card fraud is an expensive problem for many financial institutions, costing billions of dollars to companies annually. Many adversaries still evade fraud detection systems because these systems often do not include information about the adversary's knowledge of the fraud detection mechanism. This project aims to include information about the "fraudster's" motivations and knowledge base into an adaptive fraud detection system. In this project, we use a game theoretical adversarial learning approach in order to model the fraudster's best strategy and pre-emptively adapt the fraud detection system to better classify these future fraudulent transactions. Using a logistic regression classifier as the fraud detection mechanism, we initially identify the best strategy for the adversary based on the number of fraudulent transactions that go undetected, and assume that the adversary uses this strategy for future transactions in order to improve our classifier. Prior research has used game theoretic models for adversarial learning in the domains of credit card fraud and email spam, but this project adds to the literature by extending these frameworks to a practical, real-world data set. Test results show that our adversarial framework produces an increasing AUC score on validation sets over several iterations in comparison to the static model usually employed by credit card companies.**

Index Terms - Game theory, Gaussian mixture model, Oversampling, Synthetic data

INTRODUCTION

In many domains, efficient classification systems are being built to fight fraud and other malicious activities. Yet there are adversaries who are able to evade a system's defenses, resulting in a security breach. Resulting breaches cost businesses billions of dollars every year as well as harm their reputation with customers. In the domain of credit card fraud detection, fraudsters are probing the classification system in order to generate fraudulent transactions that go undetected. For example, fraudsters can purchase thousands of credit card numbers and social security numbers as a means of testing and learning of the current fraud detection systems in use. The performance of the fraud detection system can progressively deteriorate and the amount of time and cost incurred in maintenance is significant. Currently, many fraud detection systems are deployed that are successful in flagging genuine fraudulent transactions, but there are few systems that actively incorporate an adversary's potential strategies when attempting to improve defenses. In order to deploy robust systems and create an adaptive model, knowledge of an adversary's most effective strategy is beneficial.

Previous work has utilized a game theoretic model, the adversarial classifier reverse engineering (ACRE) approach and Markov processes to model the interactions between the fraudster and the classifier [1][2][3]. This project adds to current research in this area by extending the game theoretic framework to a real-world data set in fraud detection, implementing the most effective adversarial strategy and retraining the classifier in multiple rounds of a game. To implement this approach, we employ a simple logistic regression model to classify charges as fraudulent or non-fraudulent, and then play a series of games to imitate the adversary's learning process and preemptively retrain the classifier. The contributions of this paper are the following:

- Introducing an adaptive fraud detection system that utilizes repeated games in the form of a feedforward model and incorporates the synthetic minority oversampling technique (SMOTE) to mitigate class imbalance.
- Utilizing Gaussian Mixture Models (GMMs) to segment the distribution space of continuous attributes as a means to find possible adversarial strategies.

The goal of this analysis is to develop an adaptive, continuously improving model that anticipates the adversary's best strategy and preemptively fights against it, improving its performance over other currently employed models in fraud detection. Test results reveal that modeling an adversary aware classifier is more effective than a static model as evident by the increasing area under curve (AUC) score on validation sets over several iterations of the adversarial framework.

RELATED WORKS

Increasing rates of credit card fraud have resulted in significant monetary losses for financial institutions [4]. As this is a recurring problem, fraud detection systems have evolved to utilize a wide number of common classification approaches such as logistic regression, support vector machines, random forest models to more advanced techniques such as artificial neural networks, genetic algorithms, hidden Markov models, and unsupervised methods [1][5][6][7].

978-1-5386-1848-6/17/$31.00 ©2017 IEEE

While certain models may be more successful at detecting fraud than others, all are equally susceptible to attack if not retrained.

Another challenge in fraud detection is the issue of class imbalance where the percentage of fraudulent transactions is typically less than 1%. To account for this, several techniques have been proposed which include oversampling of the minority class, undersampling of the majority class and implementation of cost-sensitive loss functions. In this paper, a version of oversampling known as SMOTE is used to balance the class ratio by generating synthetic instances of fraudulent transactions [8].

The concept of adversarial learning is a specialized area of machine learning that concentrates on the interactions between an adversary and an opponent. Modeling of adversarial scenarios has been previously implemented, each with varying assumptions on the amount of knowledge the adversary has about the classification system. Dalvi et al. investigated the method by which to find an adversary's optimal strategy in the domain of spam detection using a naive Bayes model [9]. This approach makes the assumption that the adversary has complete knowledge of the classifier, which in the fraud detection domain is unrealistic. In contrast, Lowd et al. assume that the adversary has incomplete knowledge of the classifier and introduce the adversarial classifier reverse engineering (ACRE) approach, which incorporates a cost function in order to find strategies that are high quality but low cost [2]. While this learning approach may find the optimal strategy, for an adversary to successfully commit fraud any low-cost tactic is sufficient in evading the classifier [10]. Liu et al. implement a game theoretic model based upon Stackelberg games and conclude that to reach equilibrium both adversary and classifier need to be playing the game with their best strategy [1]. While a working game theoretic model is implemented, subsequent strategies to retrain the classifier are not explored.

Using a two-tiered approach to build a dynamic fraud detection system, Vatsa et al. model an adversary's strategies as a two-state Markov process [3]. While in this game-theoretic model only transaction amount is used to segment the strategy space, other attributes could be incorporated to create more complex strategies. A method to segment multiple continuous attributes are GMMs. Kantarcıoğlu et al. utilize GMMs in a game theoretic model as an attribute selection method by choosing the subset of attributes that have the best performance at equilibrium [11].

Learning from skewed distributions is a more accurate representation of reality, especially in an adversarial environment. Dal Pozzolo et. al introduce the 'Propagate and Forget' approach which takes each new stream of data and includes fraudulent transactions from previous streams in order to reduce the effects of class imbalance [12][13]. The advantages of this approach is that the minority class is accumulated quickly and can prove to adapt. This approach along with SMOTE and GMMs for strategy generation are incorporated in our game-theoretic model to simulate a fraud detection system within an adversarial environment.

DATA

The data was provided by a major financial institution, which included 36 GB of data consisting of 86 million anonymized credit card transactions from a single year. Of the 69 total variables included in the provided data, we used the following for our analysis:

- Fraud indicator (target variable)
- Current balance of the account at the end of posting
- Merchant category code
- Total number of authorizations
- Authorization amount (transaction amount)
- Authorization outstanding amount (how much is left on the card's balance)
- Average daily authorization amount
- Plastic issue duration (in days)
- Point of sale entry method
- Recurring authorization (binary; 0/1)
- Approximate distance of customer's home from merchant

The variables above were chosen as attributes that an adversary could reasonably influence when trying to commit undetected fraud. In order to model reasonable strategies for the adversary, the variables must be controllable by the adversary. For example, one strategy an adversary may employ is to quickly charge a large amount of money on a card and then stop charging the card all together. So the variables relating to length of time the card has been issued and transaction amount could illustrate such a strategy in our adversarial learning algorithm.

METHODS

In order to test our hypothesis that adversaries are constantly learning about and adapting to the classifier, we utilized an algorithm that illustrates this process as a repeated game as shown in Figure I. In each round of this game, the adversary first has a choice between strategies (simplified in Figure I, to only two strategies), and then the classifier (the bank in this scenario) has the opportunity to either retrain the classifier or utilize the same classifier.

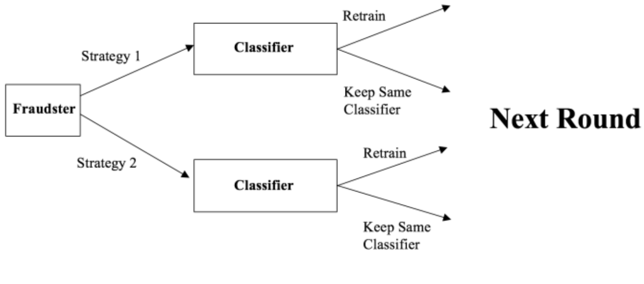

FIGURE I
ADVERSARIAL LEARNING GAME TREE

In a normal fraud detection algorithm, it is unknown what strategy adversaries are using more often (i.e. what branch they are on in the game tree), so this information is ignored when deciding whether or not to retrain the classifier. Our adversarial learning algorithm, illustrated as pseudocode in Algorithm 1, attempts to predict the best strategy for the adversary based on which strategy yields the adversary the highest false negative rate, so that more fraudulent charges go undetected by the classifier. This algorithm assumes that the adversary is constantly testing and learning more about the classifier, so adversaries will perform more of a certain transaction type if they believe it will go undetected.

These strategies can be simple, such as choosing the transaction amount, or more complex to include multiple variables. In Algorithm 1, we choose a GMM as an unsupervised way to create three distinct strategies that the adversary can choose from, as opposed to creating hard cutoffs for a variable in a particular strategy. This GMM used all of the variables listed in the data section, with the exception of the non-continuous variables (Fraud indicator, Point of sale entry method, Merchant category code, and the Recurring authorization indicator) that cannot be used in a GMM. The motive for creating three strategies came from segmentation of the transaction amount into low, medium and high bins, which was extended to incorporate the rest of the continuous attributes listed in the data section by applying a GMM to the entire variable space. We assign these strategies to each of the fraudulent transactions, and then choose the best strategy based on the highest false negative rate against the current model. We then select the subset of fraudulent transactions that were classified as having the best strategy.

We utilize an algorithm called SMOTE (Synthetic Minority Over-sampling Technique) in order to create synthetic data from our best strategy dataset. We add this synthetic data to our next round of the game in order to preemptively retrain the model to predict the next round of fraudulent transactions better. We add enough of the synthetic data such that each dataset has around 15% fraudulent transactions, which is an industry standard for oversampling. Although the original dataset only included about 0.1% fraudulent transactions, this oversampling of the fraudulent transactions also helps to improve the classifier's predictive capabilities [13].

For simplicity and transparency, we chose to use a logistic regression model with limited features (only the aforementioned features in the data section) as our model in this algorithm, although almost any classifier could be utilized in this approach.

The aforementioned adversarial learning algorithm (Algorithm 1) is described in the following pseudocode.

Algorithm 1 Adversarial Learning

Input: Transaction-level data
Output: ROC curve and AUC for each round of game

```
n ← number of rounds of the game
Split full dataset into n sequentially ordered sub-datasets
for i=1 to n:
    Split dataset$_i$ into training, testing, and validation sets
    Train logistic regression model using training set
    Test model using testing set to calculate confusion matrix
    Fit a Gaussian Mixture Model with 3 components
    Assign each transaction to the component (strategy) it most likely belongs to
    Best_strategy ← strategy that provides the adversary with the highest false negative rate
    Best_fraud ← subset of the fraudulent transactions that utilize Best_strategy
    Implement SMOTE on the Best_fraud data to create synthetic fraudulent data that exemplifies Best_strategy
    Append this synthetic data to dataset$_{i+1}$
    Test logistic regression model on validation set
    Return ROC and AUC
end for
```

In order to compare our adversarial learning model against another classifier, we use the extreme case in which the classifier stays the same, but the adversary keeps adapting. In this scenario, the aforementioned algorithm is kept relatively the same, with the only difference being that the model is never retrained. The oversampling and identification of the best strategy still exists in this case since this exemplifies the adversary learning more and more about the constant classifier. In order for adversarial learning to be shown to be effective, the adversarial learning model should outperform the model where the classifier remains the same.

Results

In our experimental approach, we chose to use ten rounds to illustrate the repeated game (n=10). In Figure II, we show four of the ten receiver operating characteristic (ROC) curves tested on the validation sets to illustrate how this adversarial learning classifier anticipates what the adversary will do next and improves as the game progresses. The ROC curves gradually shift upwards, with the AUC scores increasing from 0.78 in the first round to 0.84 in the last round of the game, supporting our hypothesis that the classifier's performance would improve as the rounds progressed in the game.

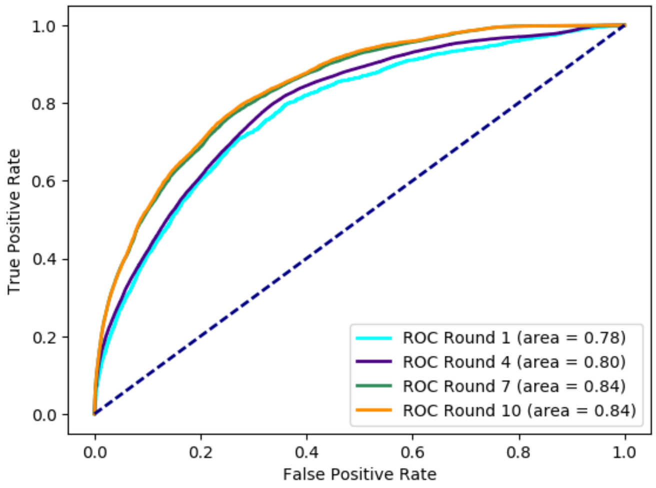

FIGURE II
ROC CURVE WITH ADVERSARIAL LEARNING

Figure III illustrates the extreme opposite case, where the adversary continues to learn about a constant classifier that is never retrained. We expected to see a sharp bending in the curves towards the 45 degree line as the rounds progressed, since the adversary would become extremely adept at avoiding detection as it learned more about the classifier. We did not see this drastic of a decrease, however, with the AUC scores decreasing from 0.78 in the first round to only 0.76 in the last round.

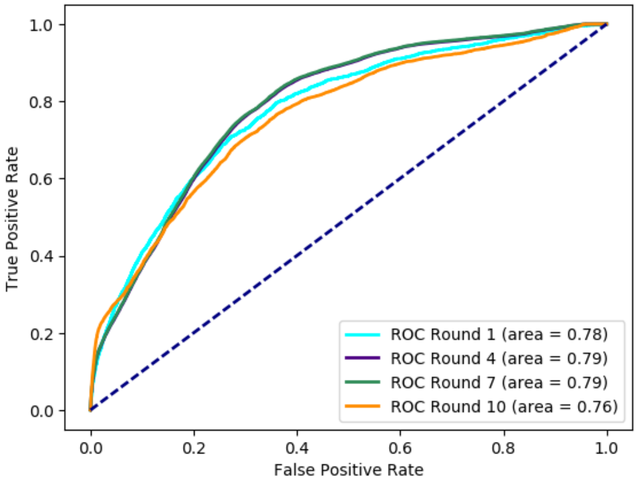

FIGURE III
ROC CURVE WITHOUT ADVERSARIAL LEARNING

Table I shows the performance of the models in all ten rounds of the game in terms of the AUC score. In all ten rounds, we find that the adversarial learning classifier always outperforms, or at least performs as well as, the non-adversarial learning classifier, which does support our hypothesis that anticipating the best move of the adversary does improve the classifier. We anticipated a much greater decrease in performance in the model where the adversary learns but the classifier remains static, but these results do illustrate the benefits of utilizing adversarial learning in a credit card fraud detection model over a more common, non-adversarial approach.

TABLE I
AUC VALUES FOR ALL TEN ROUNDS OF THE GAME

Round	AUC - Adversarial Learning	AUC - Classifier Stays The Same
Round 1	0.779	0.779
Round 2	0.810	0.805
Round 3	0.797	0.790
Round 4	0.797	0.790
Round 5	0.791	0.791
Round 6	0.799	0.790
Round 7	0.837	0.792
Round 8	0.839	0.790
Round 9	0.836	0.780
Round 10	0.841	0.762

CONCLUSION AND FURTHER RESEARCH

We conclude that modeling adversaries' possible strategies in order to preemptively retrain our model proved to outperform a static model in ability to detect fraudulent transactions. As rounds progressed, the separation between AUC scores of the adversarial learning model and the static fraud model increased. Although the differences in AUC may seem small, the slightest change in AUC could potentially result in a substantial reduction of costs due to fraud. By understanding the weaknesses of our own model, we were able to preemptively adjust our classifier to provide better defense mechanisms for detection against fraud.

The use of a GMM in determining a best strategy proved an effective way of finding optimal new transactions an adversary is likely to replicate. The use of SMOTE provided a useful tool in our ability to produce synthetic transactions of this best strategy. Overall, these two contributions provided tools able to mimic an adversary's learning and thought processes, giving the credit card company the ability to preemptively react to the changing transaction strategies.

In future research, there are many possible additions to our framework that would provide more information and realism in our models and could possibly improve our results. In order to differentiate the various possible fraud strategies, our GMM could be optimized to produce more regions of possible transaction types. In our SMOTE algorithm, we choose to introduce enough fraud for the next round to have 15% fraudulent transactions, though this number could also be optimized to the percentage of fraud that yields the most effective classifier.

To improve upon our current classification algorithm, we wish to include velocity variables in an effort to discover more revealing characteristics of the transactions. We also wish to apply this adversarial framework to differing classification algorithms to see how it fares against a static model. An element of randomness could also be introduced in an attempt to confuse the adversaries and disrupt their previous learned knowledge of the system.

To better understand possible business strategies developed from the analysis of costs of fraud, the best fraud strategy could have been chosen by the total monetary value of the missed transactions. Various other cost factors should be taken into account in model creation that were not. There exists a cost of retraining the classifier that was not addressed when retraining after every round of the game. The cost of retraining could be compared to the benefit of changing the classifier, revealing the ideal frequency of retraining. This could also be done systematically by retraining the classifier only when the AUC falls below a certain value.

ACKNOWLEDGEMENT

We would like to thank Will Franklin, Dan Teodorescu and Brian Sachtjen for their time, generosity and critical insights on this project.

REFERENCES

[1] W. Liu and S. Chawla. "A game theoretical model for adversarial learning." In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pp. 25-30, IEEE, 2009.

[2] D. Lowd and C. Meek. "Adversarial learning", In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641-647, ACM, 2005.

[3] V. Vatsa, S. Sural, and A. K. Majumdar, "A Rule-Based and Game-Theoretic Approach to On-Line Credit Card Fraud Detection," *Techniques and Applications for Advanced Information Privacy and Security: Emerging Organizational, Ethical, and Human Issues: Emerging Organizational, Ethical, and Human Issues,* 2009.

[4] A. Pascual, S. Miller, and K. Marchini. "Identity fraud: fraud hits an inflection point." *Javelin Strategy*, 2016, Web, <https://www.javelinstrategy.com/coverage-area/2016-identity-fraud-fraud-hits-inflection-point>.

[5] A. Shen, R. Tong, and Y Deng. "Application of classification models on credit card fraud detection." In *Service Systems and Service Management, 2007 International Conference on*, pp. 1-4. IEEE, 2007.

[6] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar. "Credit card fraud detection using hidden Markov model." *IEEE Transactions on dependable and secure computing*, pp. 37-48, 2008.

[7] R. J. Bolton., and D. J. Hand. "Unsupervised profiling methods for fraud detection." *Credit Scoring and Credit Control VII*, pp. 235-255, 2001.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, 16, pp. 321-357, 2002.

[9] N. Dalvi, P. Domingos, S. Sanghai and D. Varma, "Adversarial Classification", *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99-108, 2004.

[10] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar. "Adversarial machine learning." In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43-58. ACM, 2011.

[11] M. Kantarcıoğlu, B. Xi, and C. Clifton. "Classifier evaluation and attribute selection against active adversaries." *Data Mining and Knowledge Discovery* 22(1), pp. 291-335, 2011.

[12] A. Dal Pozzolo, O. Caelen, Y. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective", *Expert systems with applications*, 41(10), pp. 4915-4928, 2009.

[13] A. Dal Pozzolo. "Adaptive Machine Learning for Credit Card Fraud Detection", Thesis, Université Libre De Bruxelles, 2015, Web, <http://difusion.ulb.ac.be/vufind/Record/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/221654/Holdings>.

AUTHOR INFORMATION

Mary Frances Zeager, M.S. Student, Data Science Institute, University of Virginia

Aksheetha Sridhar, M.S. Student, Data Science Institute, University of Virginia

Nathan Fogal, M.S. Student, Data Science Institute, University of Virginia

Stephen Adams, Senior Scientist, Department of Systems and Information Engineering, University of Virginia

Donald E. Brown, Director, Data Science Institute and W.S. Calcott Professor, Department of Systems and Information Engineering, University of Virginia

Peter A. Beling, Associate Professor, Department of Systems and Information Engineering, University of Virginia