Active Learning for Credit Card Fraud Detection

Fan Wang
Master of Applied Science
Department of Electrical and Computer Engineering
Email: fan.wang@ryerson.ca

Abstract—**Fraud detection in credit card transactions has unique challenges compared to traditional supervised learning. The dataset is highly imbalanced and changes with time. It also needs to be actively labeled as most transactions are unlabeled. This paper seeks to solve these problems and compares the accuracy of various strategies, such as high risk querying, exploratory active learning, and stochastic semi-supervised learning. The traditional exploitation/exploration tradeoff in machine learning is also studied in the context of credit card fraud detection, and it is found that a small exploration budget is ideal. Simulations are run with generated datasets as real-world credit card data are classified information. High Risk Querying is found to be sufficient for a small dataset, and other strategies offer little if any improvements over it. With a large dataset, such as practical applications for credit card fraud detection, it is expected that combining stochastic semi-supervised learning with exploratory active learning will yield best accuracy.**

I. Introduction

Applying machine learning techniques to credit card fraud detection has three unique and major problems: the number of fraudulent transactions are much less than the number of genuine transactions (big imbalance between genuine and fraudulent transactions), constant behavior changes of customers and fraudsters (methods of committing fraud constantly evolving), and the need to actively label data by human investigators (large amounts of unlabeled data that are costly to label).

Active learning seeks to address the above mentioned problems. The method is described by the iteration of three steps:

1) Selection: Select a number of data points with the highest risk predicted by an existing classifier.
2) Querying: Selected data points are labeled by human investigators.
3) Training: The labeled data points are used to train/update the existing classifier.

Selection and querying are the active learning steps which differs from traditional supervised learning, in which the classifier is trained using all of the data in the dataset minus the testing set, and where all the data are already labeled. The selection and querying step only selects data points with the highest estimated probability of fraud, and is actively labeled by human investigators.

It is expected that credit card user and fraudster's behaviors will change over time, and thus the data base needs to be constantly updated with the newest and arguably most relevant data. This active learning approach is termed streaming active learning, as opposed to pool-based active learning, in which the selection and query steps selects from the same set of data. In pool-based active learning, the classifier's accuracy will improve over time when tested with a testing data set taken from the same data pool, but this isn't a guarantee in streaming active learning. Streaming active learning may not be accurate for newer data sets if there is a pattern change within the newer data sets. For this reason, a weighted sum of two classifiers, one trained on old data sets, and one trained on the most recent data sets is sometimes used for non-stationary data.

$$P_C(+|x) = \omega^A P_{Di}(+|x) + (1 - \omega^A) P_{Fi}(+|x) \tag{1}$$

Where $D_i$is a classifier trained with past data using a Random Forest, and Fi is trained on recently alerted data points.

II. The Fraud Detection Classifier

The classifier is created by selecting a number of already labeled transactions and implemented using the Random Decision Forest method.

Active learning strategy is implemented to cheaply and efficiently train the classifier with massive amounts of unlabeled and highly imbalanced data.

Every day a number of alerts are raised based on the highest probability of fraud obtained from an initial classifier. These data points are labeled by real investigators, and the labeled data is used to retrain the classifier. A realistic number of daily alerts raised is set to 100, this is based on the high cost of human investigators, who can only affordably label 100 data points a day.

Any supervised learning method, whether regression or classification can be used to train the fraud detection classifier, including but not limited to, k-nearest neighbor, linear regression, logistic regression, and deep neural networks.

Decision forest method was chosen as the focus of this paper is on active learning techniques, and not conventional supervised learning. It could be argued that a deep neural network can potentially produce a better result.

TABLE I
SUMMARY OF ACTIVE LEARNING STRATEGIES AND THEIR ABBREVIATIONS

Id	Strategy	Type
HRQ	Highest Risk Querying	Base
R	Random Querying	Exploratory Active Learning (EAL)
U	Uncertainty Querying	
M	Mix of Random and Uncertainty Querying	
SR	SSSL on Random Points	Stochastic Semi-Supervised Learning (SSSL)
SU	SSSL on Uncertain Points	
SM	SSSL on Random/Uncertain Points	
SE	SSSL on Most Likely to be Genuine Points	
SR-U	SSSL on Uncertain Points + Random Sampling	SSSL and EAL
SR-R	SSSL on Random Points + Random Sampling	
SR-M	SSSL on Random/Uncertain Points + Random Sampling	

FIGURE 1
HIGH RISK QUERYING PSEUDO-CODE

```
Code                                          Comments
Initialize k                                  Number of Daily Alert
Initialize q                                  Exploration Budget
Initialize m                                  SSSL Budget
Initialize D                                  Training Set
 C <- Train(D)
Scores <- { P(x), x ∈ k}
Sel <- points with highest risk scores
If (q > 0) then
  Sel <- sel + q explorative points
End If
Labeled <- investigators labeling Sel
If (m>0) then
   SSSLset <- m points labeled as genuine
  Labeledset <- Labeledset + SSSLset
End If
D <- D + Labeledset
```

A. *High Risk Querying*

Each day a number of data points with the highest estimation of a posterior probability $P_c = (+|x)$ is selected, where x is a vector that includes features of the transaction (such as transaction amount, terminal, time of day and etc.), and + is a fraudulent transaction.

The data points are then labeled by human investigators and the classifier is re-trained using the combination of both old and new data. This method labels data points of interest and addresses the issue of a very large imbalanced and unlabeled data set.

A pseudo-code for highest risky querying is seen in figure 1. In the pseudo-code, an initial daily number of alerts, exploration budget, a SSSL budget, and an initial training set have to be given. Realistically, a daily alert of 100, an exploration budget of 5, and a SSSL budget of 1000 is found to be the most effective.

The daily alerts are chosen based on the highest probability of risk given by the classifier, these data will be added to a data set that are going to be labeled by human investigators. A q number of exploratory data chosen with an exploratory strategy will also be added to the data set to be labeled. Human investigators then label the data set. A stochastic semi-supervised learning scheme will then choose m data points from the unlabeled dataset and automatically label them as genuine. The combination of the human investigator labeled data and SSSL labeled data is then added to the training set and the classifier is retrained.

Specific weights can be given to newly labeled data to combat the non-stationary and changing behavior of data in credit card fraud detection.

B. *Exploratory Active Learning*

Exploratory active learning strategy modifies the high risk querying strategy in order to train the classifier in case of fraudster behavior change. A certain number of q data is chosen along with the highest risk k data points to be labeled by human investigators. Two methods are proposed to select the q data points:

1) *Random selection, in which query points are selected randomly in the data set.*
2) *Selection by maximum uncertainty, in which query points with $P_c = (+|x) \approx 0.5$ are selected.*

Random selection selects points randomly, due to the highly imbalanced data set, this selection criteria expects to select mostly genuine data points. Selection by maximum uncertainty selects points in which the classifier has the most amount of uncertainty ($P_c = (+|x) \approx 0.5$), this selection criteria also expects to select mostly genuine data points, but should in theory select more fraudulent data points than random selection.

A combination of the two techniques can be applied, in which some data are selected using random selection and some using selection by maximum uncertainty.

C. *Stochastic Semi-Supervised Learning*

The stochastic semi-supervised learning strategy infers labels in case of highly imbalanced data sets with a large number of unlabeled points. The method relies on the fact that the probability of randomly selecting a fraudulent transaction is extremely small due to the imbalance of the data set.

SSSL will decrease labeling cost as the number of data selected by SSSL method is automatically labeled.

The strategy has one additional step to the above mentioned method and consists of four steps:

1) Selection: Select a number of data points with highest risk querying and exploratory active learning, each with a different weighting.
2) Querying: Selected data points are labeled by human investigators.
3) Majority Assumption: a number of transactions are labeled as genuine by majority assumption.
4) Training: The labeled data points are used to train/update the existing classifier.

Four methods for choosing the majority data points are proposed:

1) Uncertainty: select an amount of data with the most uncertainty and label them as genuine (probability of selecting a fraudulent transaction is 0.35%).
2) Random Attribution: randomly select an amount of data and label them as genuine (probability of selecting a fraudulent transaction is 0.13%):
3) Mix of Randomness and Uncertainty: select some data with the Random Attribution method and some data with the Uncertainty method and label them as genuine.
4) Low Predicted Risk: select data with the highest probability of being genuine and label them as genuine.

Since the probability of selecting a fraudulent data is extremely low, it makes sense to label these data as genuine. It is expected that low predicted risk selection will give the best accuracy to the classifier, as low predicted risk will select the least amount of fraudulent transactions to label as genuine.

Random attribution selects data randomly and labels them as genuine, and it has a probability of selecting a fraudulent transaction of 0.13%. Uncertainty criteria selects the data points that the classifier is the most unsure about, and it has a probability of selecting a fraudulent transaction of 0.35%.

It would appear choosing random attribution over uncertainty would yield better results as the chances of mislabeling data is smaller, however, as the dataset is streamlining and non-stationary, a mix of randomness and uncertainty will produce a better accuracy in the long run and be less prone to bias.

TABLE II
STRATEGY ACCURACY SCORES

Symbol	ROC Accuracy	PRC Accuracy
HRQ	0.870361	0.647306
R	0.866259	0.645293
U	0.868814	0.643669
M	0.868508	0.645564
SR	0.826645	0.626771
SU	0.827927	0.632565
SM	0.875416	0.634952
SE	0.829602	0.632142
SR-U	0.835443	0.627991
SR-R	0.822399	0.62159
SR-M	0.831149	0.631798

FIGURE 2
AREA UNDER RECEIVER OPERATOR CHARACTERISTIC

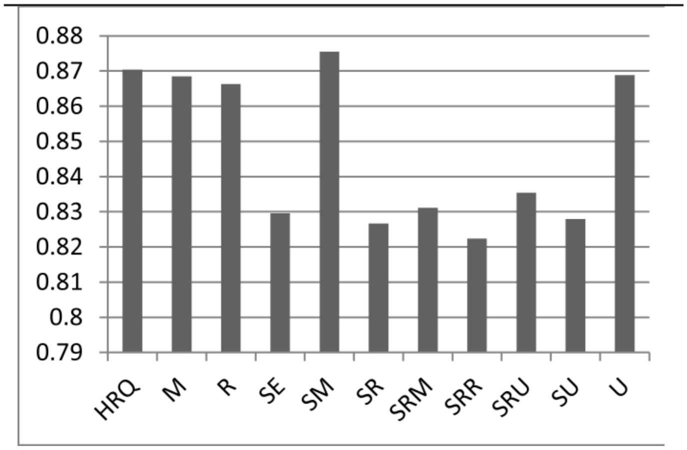
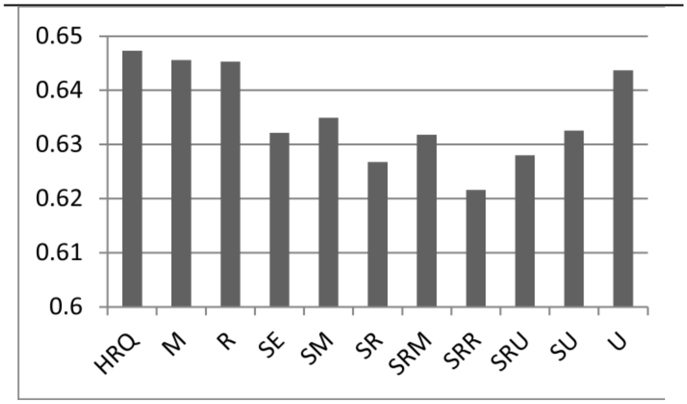


FIGURE 3
AREA UNDER PRECISION-RECALL



III. SIMULATION

Transaction data of 500 sample transactions per day for 30 days are generated, each sample includes five features. Due to the imbalance of the data set, conditions on features will be set so that only 0.13% of transactions will be fraudulent. The classifier is trained with 95 daily alerts that are to be labeled by investigators (which are already labeled in for the sake of the experiment), 5 exploration budget, and 300 SSSL budget. The remaining 100 samples are used as a testing set. The classifier is trained using the conventional Random Forest algorithm. To compare the accuracy of various active learning techniques, Area under the Precision-Recall (PRC) and Area under the Receiver Operator Characteristics (ROC) are used as the measuring metrics. Area under the Precision-Recall is likely the better accuracy indicator as the class is highly imbalanced. The various active learning techniques' and their abbreviations are outlined in Table 1.

Receiver operator characteristic plots the true positive rate against the false positive rate. Since the class is highly imbalanced, the true positive rate will significantly outweigh the false positive rate, as the prior probability for genuine transactions is much higher.

Precision recall plots the precision against the recall. Precision is the relevant data point amongst all picked data points; in this case, it is the actual fraudulent data amongst classifier chosen fraudulent data. Recall is the correctly picked fraudulent data by the classifier amongst all fraudulent data. This serves as a better indicator of accuracy due the data set being so highly imbalanced, and area under ROC will tend to overestimate the accuracy of the classifier.

The simulation is coded in R, and data is generated by random and put into excel format.

The accuracy results are averaged over the 30 days. The small data set generated has a higher variance by nature and can cause SSSL techniques to label more fraudulent transactions as genuine. SSSL should in theory produce worse results as some fraudulent data will be labeled as genuine regardless of the variance or size of the data set. SSSL is meant to label more data faster than simply relying on human investigators, and this serves as a mean to provide sufficient data to train the classifier initially.

A. Results

The results of the simulation favor high risk querying, in terms of both ROC and PRC, and exploratory active learning has the second highest accuracy. The experiment is meant to be performed with 12 million credit card transactions, but due to the inaccessibility of credit card transaction data, only 15000 transactions data was generated and used for this simulation, therefore many data could be incorrectly labeled as genuine due to the high variance nature of a small dataset. The accuracy measurements are shown in Table 2.

It is expected that a combination of SSSL and EAL would outperform the baseline HRQ, but due to the small size of the training set, SSSL is not very useful, as there is not a huge amount of unlabeled data, which means human investigators can realistically label them all, and produce a better accuracy.

The accuracy of ROC measurement of various strategies are shown in Figure 2, and accuracy of PRC measurement of various strategies are shown in Figure 3.

It is seen that generally, exploratory active learning outperforms stochastic semi-supervised learning or the combination of the two. The reason why exploratory active learning yields better results with smaller data set is likely due to the fact that SSSL will incorrectly label some fraudulent data as genuine. Since the exploratory budget is low, it will not have a big impact on the accuracy of the classifier, but will improve the accuracy slightly over some period of time.

SSSL strategy will get the classifier trained in a faster amount of time, and improve it at a much faster pace with minimum effect on accuracy. If tested on actual large imbalanced data set, the SSSL and EAL strategy will produce a better accuracy much faster than high risk querying.

SSSL strategies produce better accuracy on the first few days of training, as shown in Table 3. This is expected as there are less labeled data initially, as labeling a portion of them as genuine will produce more labeled data to train the classifier.

TABLE III
EARLY DAYS TRAINING ACCURACY

Day	HRQ Accuracy	PRC Accuracy
1	0.612926	0.631931
2	0.656288	0.679719
3	0.749503	0.759436
4	0.862752	0.894574
5	0.945377	0.958778
6	0.836296	0.856779
7	0.829821	0.833454

IV. CONCLUSION

This investigation simulated a proposed strategy of detecting fraudulent credit card transactions. Exploratory active learning, stochastic semi-supervised learning and a combination of both are coded and simulated in R programming language.

High risk querying is the baseline strategy to train a classifier to recognize fraudulent credit card transaction. It selects data points with the highest estimated risk and lets human investigators label them. The classifier can have bias when trained using high risk queried data. To combat this, exploratory active learning is proposed. Exploratory active learning picks data points that the classifier wouldn't normally label as high risk and gives it to human investigators to label as well. Data points chosen with exploratory active learning can be random, most uncertain, or a combination of both.

Labeling data with human investigators can be very expensive, so a stochastic semi-supervised learning method is proposed. SSSL takes advantage of highly imbalanced datasets, in the sense that there is a very low chance of choosing a fraudulent data point by random, uncertainty, low risk, or a mixture of any of the three. SSSL strategy labels data chosen with a given scheme as genuine, and it can label much more data points per day than a human investigator, as long as it doesn't affect the classifier accuracy.

The results show that Highest Risk Querying performs best when given a small data set with high variance, but according to the paper, combining stochastic semi-supervised learning with exploratory active learning will yield better results when bigger data set is used, as the variance in the data set is smaller.

In the future, combination of EAL and SSSL can be further analyzed to find the theoretical optimal budget for each of them, and a general solution can be found which can solve all classification problems with a large, unbalanced, and unlabeled dataset.

REFERENCES

[1] F. Carcillo, Y. A. L. Borgne, O. Caelen and G. Bontempi, "An Assessment of Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection," *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Tokyo, 2017, pp. 631-639.