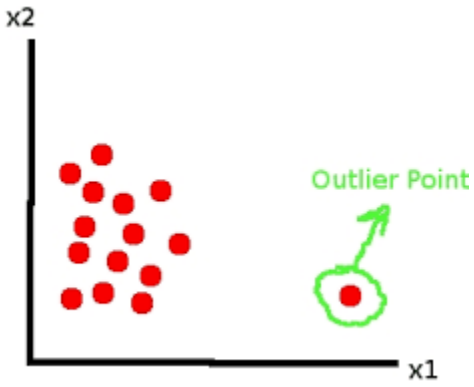
Comparison of effectiveness of 3-layer neural network to Density based anomaly detection

Ankur Yadav

ankur221b@gmail.com

www.ai-techsystems.com

Abstract - In data mining, **anomaly detection** (also **outlier detection**) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.



Keywords—Credit Card, Neural Network, Anomaly Detection

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 incidences of credit card fraud are limited to about 0.1% of all card transactions, they have 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.[3] 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.

In the decade to 2008, general credit card losses have been 7 basis points or lower (i.e. losses of $0.07 or less per $100 of transactions).

PROJECT PLAN

The purpose of this project to develop a best algorithm to find the outliers or frauds in case of credit cards. We will implement machine learning and deep learning algorithms and compare them and choose the best algorithm.

DATASET

The datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2 ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

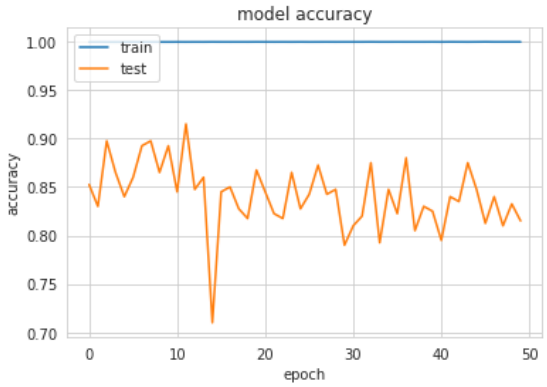
NEURAL NETWORK

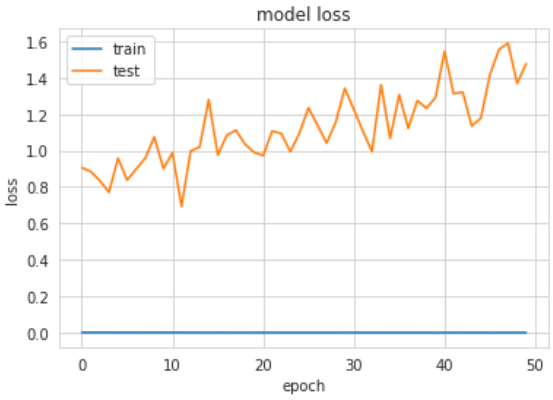
Neural Network are concept of deep learning. Keras is used to implement NN. In this case we are using 3 layers. First layer or Input layer consist of 512 neurons. The hidden layer or 2nd layer consist of 128 neurons. The output layer is the 3rd and final layer where we get the classified output. It consist of 2 neurons. The output either be 1 or 0 where 1 indicate fraud case and 0 indicate normal. Dropout regularization is added after first and second layer.

- Optimizer used : **adam**
- Activation function used : **ReLU and softmax**
- Loss function used : **categorical_crossentropy**
- Metric used : **accuracy**

➢ ***Result*** - Training accuracy: 99.95%

Testing accuracy: 91.25%





ANOMALY DETECTION ALGORITHM

Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers. It has many applications in business, from intrusion detection (identifying strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumor in an MRI scan), and from fraud detection in credit card transactions to fault detection in operating environments.

a) ***K-NEAREST NEIGHBOR*** : The distance to the *k*th nearest neighbor can also be seen as a local density estimate and thus is also a popular outlier score in anomaly detection. The larger the distance to the *k*-NN, the lower the local density, the more likely the query point is an outlier. To take into account the whole neighborhood of the query point, the average distance to the *k*-NN can be used.

- Libraries used for algorithm: **sklearn.neighbors**
- Hyperparameters used: n_neighbors**=1,** metric**='euclidean',** leaf_size**=20,** weight**='uniform',** n_jobs**=-1**

➢ ***Result*** **–** Accuracy: 88.25%

b) ISOLATION FOREST: In Isolation forest we partition randomly, unlike Decision trees where the partition is based on gain. Partitions are created by randomly selecting a feature and then randomly creating a split value between the maximum and the minimum value of the feature. We keep on creating the partitions until we isolate all the points (in most cases we also set a limit on number of partitions/heights of the tree).

- Libraries used: **sklearn.ensemble**
- Hyperparameters used: max_features=**1**, max_samples=**100**, n_jobs=**-1**, behavior=**'old'**

➢ ***Result*** **–** Accuracy: 92%

CONCLUSION

Many algorithms like k-nearest neighbor, isolation forest and neural networks are used. All the algorithms have been analyzed and compared on basis of accuracy they are giving on same data. Upon analyzing we conclude that 3-Layer Neural Network have been the best algorithms for the purpose of credit card fraud detection as it provides best accuracy. For better performance we can play with hyperparameters and provide more data.

ACKNOWLEDGEMENT

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud

project Please cite the following works: Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015 Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi) Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing Bertrand Lebichot, Yann-Aël

REFERENCES

[1] http://mlg.ulb.ac.be

[2] https://www.researchgate.net/project/Fraud-detection-5

[3] https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

[4]https://www.kaggle.com/mlg-ulb/creditcardfraud