Seminar Report
On

"USING SELF-ORGANIZING MAPS FOR FRAUD PREDICTION AT ONLINE AUCTION SITES"

Submitted by
MITHUNMOHAN KADAVILMADANAMOHANAN
B110765CS

Under the Course Coordinator
Mr. SREENU NAIK BHUKYA, Assistant Professor, CSED

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY CALICUT

CERTIFICATE

This is to certify that the seminar entitled "**USING SELF-ORGANIZING MAPS FOR FRAUD PREDICTION AT ONLINE AUCTION SITES**" submitted by **MITHUNMOHAN KADAVILMADANAMOHANAN** bearing roll number **B110765CS** to the Department of Computer Science and Engineering, National Institute of Technology Calicut towards partial fulfillment of the completion of course **SEMINAR** during the academic year 2014 - 2015 is a bonafide record of the work carried out by him.

Mr. SREENU NAIK BHUKYA
Assistant Professor, CSED
Course Coordinator

Date:

ABSTRACT

Online auction sites have to deal with an enormous amount of product listings, of which a fraction is fraudulent. Although small in proportion, fraudulent listings are costly for site operators, buyers and legitimate sellers. Fraud prediction in this scenario can benefit significantly from machine learning techniques, although interpretability of model predictions is a concern. In this work we extend an unsupervised learning technique – Self-Organizing Maps – to use labeled data for binary classification under a constraint on the proportion of false positives. The resulting technique was applied to the prediction of non-delivery fraud, achieving good results while being easier to interpret.

TABLE OF CONTENTS

1. INTRODUCTION

In order to keep their business growing, online auction sites like eBay need to protect buyers from unscrupulous sellers. Among the several types of fraudulent behavior that takes place in online auction sites, the most frequent one is non-delivery fraud in which fake sellers list nonexistent products for sale, receive payments and disappear, possibly reentering the market with a different identity. The challenge faced by site operators is to identify fraudsters before they strike, in order to avoid losses due to unpaid taxes, insurance, badmouthing etc. In other words, for a given product listing they need to predict whether or not it will end up being a fraud case, in order to prevent damage. Online auction sites usually have feedback systems, which makes fraud detection – identification of deceptive behavior after it has occurred – a much easier task. Usually such feedback systems are made with the help of supervised learning models of artificial neural networks, in which they can count on a relatively large number of identified fraud cases. However, these models face two important challenges when it comes to tackling this problem: the high class imbalance (many legitimate listings for each fraudulent one) and the difficult interpretation of many supervised learning models. Henceforth in this paper, an algorithm that combines an unsupervised learning model of artificial neural network, a clustering technique, called the Self-Organizing Map along with the supervised learning paradigm through the use of labeled data is proposed. The proposed algorithm tackles the problem of binary classification for highly imbalanced data, which is the case when it comes to fraud prediction. Using labeled data to automatically identify clusters of listings with high probability of fraud makes a Self-Organizing Map a tool useful for exploratory data analysis, which helps understanding the data, and machine learning, which is essential when it comes to classify thousands of new listings each day.

2. ARTIFICIAL NEURAL NETWORKS (ANN)

Artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems (in particular the brain) which is capable of machine learning as well as pattern recognition. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs. Designed around the brain-paradigm of Artificial Intelligence, neural networks attempt to model the biological brain. Neural networks are very different from most standard computer science concepts. In a typical program, data is stored in some structure such as frames, which are then stored within a centralized database, such as with an Expert System or a Natural Language Processor. In neural networks, however, information is distributed throughout the network. This mirrors the biological brain, which stores its information (memories) throughout its' synapses. Each node in a neural network is essentially its own autonomous entity. Each performs only a small computation in the grand-scheme of the problem. This architecture allows for parallel implementation. ANN's are fault tolerant in nature for small amount of bad input data.

Based on the way they learn, all artificial neural networks can be divided into two learning categories - supervised and unsupervised.

→In supervised learning, a desired output result for each input vector is required when the network is trained. An ANN of the supervised learning type, such as the multi-layer perceptron, uses the target result to guide the formation of the neural parameters. It is thus possible to make the neural network learn the behavior of the process under study.

→In unsupervised learning, the training of the network is entirely data-driven and no target results for the input data vectors are provided. An ANN of the unsupervised learning type, such as the self-organizing map, can be used for clustering the input data and find features inherent to the problem.

3. SELF-ORGANIZING MAPS (SOM)

The Self-Organizing Map is a feed-forward neural network based on unsupervised learning, developed by professor Kohonen, whose units are linear and topologically ordered in a two dimensional lattice of a given size. It is a model inspired in the several types of “maps” that exist in the brain of higher animals, linking for example the skin sensations of the different body portions to specific areas in the cortex. Being one of the most popular neural network models, it belongs to the category of competitive learning networks. SOM provides a topology preserving mapping from the high dimensional space to map units which is a two dimensional lattice. The property of topology preserving means that the mapping preserves the relative distance between the points i.e. points that are near each other in the input space are mapped to nearby map units in the SOM. The SOM can thus serve as a cluster analyzing tool of high-dimensional data. Also, the SOM has the capability to generalize which means that the network can recognize or characterize inputs it has never encountered before.

3.1 INPUT AND OUTPUT FOR TRAINING SOM

Input

Training data: vectors, X_1, X_2 … X_j …. X_p, each vector is of length n and the vector components are real numbers.

$X_1 = (x_{1,1}, x_{1,2} \ldots. x_{1,i} \ldots x_{1,n})$

…

$X_j = (x_{j,1}, x_{j,2} \ldots x_{j,i} \ldots x_{j,n})$

…

$X_p = (x_{p,1}, x_{p,2} \ldots x_{p,i} \ldots x_{p,n})$

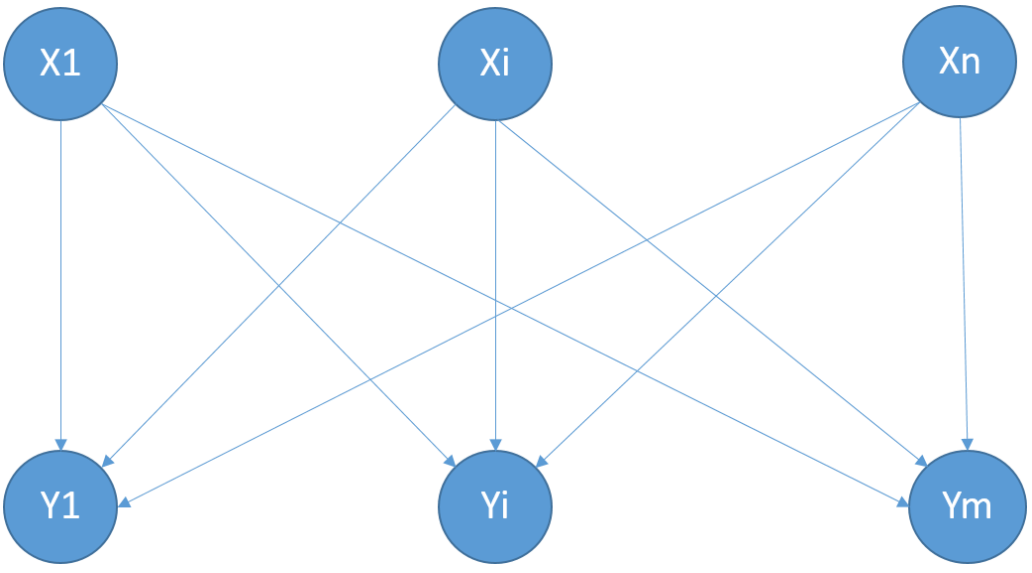
p Distinct input training vectors

Output

A vector, Y, of length *m*: (Y_1, Y_2 … Y_i … Y_m), sometimes m < n, sometimes m > n, sometimes m = n.

Each of the p vectors in the training data is classified as falling in one of m clusters or categories.

3.2 NETWORK ARCHITECTURE OF SOM



n Input units

m Output units

3.3 TRAINING ALGORITHM FOR SOM

→Select output layer network topology.

→Initialize current neighborhood distance, D(0), to a positive value.

→Initialize learning rate η(t) (This is usually some predefined function).

→Initialize weights from inputs to outputs to small random values.

→Let t = 1.

→While computational bounds are not exceeded do

1) Select an input sample, i_l

2) Compute the square of the Euclidean distance of i_l from weight vectors (w_j) associated with each output node

$$\sum_{k=0}^{n}(i_{l,k} - w_{j,k}(t))^2$$

3) Select output node *j** that has weight vector with minimum value (from step 2).

4) Update weights to all nodes within a topological distance given by D(t) from *j**, using the weight update rule:

$$w_j(t+1) = w_j(t) + \eta(t)(i_l - w_j(t))$$

5) Increment t

→Endwhile.

Learning rate generally decreases with time: $0 < \eta(t) <= \eta(t-1)$
D(t) also decreases with time: $0< D(t) < D(t-1)$

4. BINARY CLASSIFICATION USING SOM

The two-phase algorithm for binary classification of highly imbalanced datasets (in our case fraud and legitimate listings) combines clustering using Self-Organizing Maps with additional steps to label observations. The observations of the minority class is referred to as the positive observations, while the others will be referred as the negative observations. In our problem, the positive observations are the fraudulent listings and the negative observations are the legitimate listings.

To use SOM for supervised learning, the following general procedure is followed in the two phases of the binary classification algorithm:

→ Use training data to identify cluster centers (weights).This is the training of the SOM map. From now on we will use the term cluster to refer to a SOM's unit.

→ Cluster all training data using the calculated clusters' centers. This means finding for each training observation which is the closest cluster center.

→ Use training labels to label the clusters based on the distribution of training data in the clusters.

→ Cluster new data using calculated clusters' centers.

→ Label each new observation according to the label of its cluster.

4.1 PHASE 1: FILTERING OUT NEGATIVE OBSERVATIONS

The objective of this phase is to generate a labeling with the following characteristics:
(i) Almost all positive observations are (correctly) classified as such
(ii) A substantial part of the negative observations is (correctly) classified as negative.
The observations classified as positive need further processing with another classifier, which will have the advantage that this new set will be less imbalanced.

This phase follows the general procedure described in the previous section with the changes below:

→ In step 1: SOM clusters' centers are calculated using only negative observations.

→ In step 3: A cluster is labeled as positive if in the training data assigned to it at least one observation belongs to the positive class.

4.2 PHASE 2: CLASSIFICATION WITH CONSTRAINT ON FALSE POSITIVES

This phase gives the final label to the observations coming from the filtering phase, but with the added aspect of enforcement of the false positives rate. Fraud prediction and detection usually have a trade-off regarding true versus false positives and hence the existence of a constraint in the proportion of false positives – FPmax is assumed. In the context of fraud prediction, an online auction site might tolerate FPmax = 15%, while other one might tolerate FPmax = 25%.

To enforce this restriction, the general procedure presented in Section 4 is adapted in the following way:

→ In step 1: SOM clusters' centers are calculated using only negative observations, similarly to phase 1.

→ In step 3: a cluster is labeled as negative if in the training data assigned to it all observations belong to the negative class; otherwise, it is temporally labeled as undecided and is further processed using the following algorithm in order to give the final label.

The undecided clusters have mixed observations, so labeling one of them as positive necessarily increases both false positives and true positives (and vice versa), although these changes are usually different, since some clusters have proportionally more positive observations then others. Given that we have a discrete set of clusters to choose and each one has a discrete (positive and negative) set of observations, we need to solve an optimization problem: finding the set of clusters that, if labeled as positive, maximizes the number of true positives for the given maximum acceptable number of false positives. Posed this way, our problem can be reduced to the classical 0-1 knapsack problem: given a set of items, each one with a weight and a value, find the best subset of items in terms of total value for a given maximum weight, with the restriction that each item can appear at most once.
In our case:
• The individual items are the clusters of the SOM;
• The weight of each cluster is the number of negative observations assigned to it;
• The value of each cluster is the number of positive observations assigned to it;
• The maximum weight is the maximum acceptable number of false positives, which is the number of negative observations in the training set times FPmax minus the number of observations already labeled as negative in phase 1 and in step 3 of phase 2.

4.3 FINAL ALGORITHM

Although conceptually the two phases are run sequentially, in practice there is some interleaving, since it is not necessary to retrain the SOMs every time. The real sequence is the following:

→ Training (done once for one training set): steps 1–3 of filtering phase → steps 1-3 of classification phase.

→ Applying: steps 4–5 of filtering phase → steps 4–5 of classification phase.

The final algorithm has three parameters: besides FPmax, it also needs the sizes of the two SOMs. These two parameters can be selected through cross-validation with the training set, in order to find the pair that maximizes the true positives rate.

5. CONCLUSION

The main contribution of this paper is a binary classification algorithm for highly imbalanced datasets, which is the case in the problem of fraud prediction. The algorithm is based on Self-Organizing Maps (SOMs) and cluster labeling based on the 0-1 knapsack algorithm, a novel combination which gives results easy to interpret using the visualization properties of SOMs. While other works already used SOMs for classification in the realm of a similar problem (fraud detection), they relied on the fact that fraudulent observations were outliers, a premise that is not valid in fraud prediction. Besides being resistant to the class imbalance problem, the proposed algorithm also has the advantage of simplicity and easiness of interpretation, since it reduces the fraud prediction problem to labeling the appropriate SOM cells as positive, which is equivalent to choosing which regions of the feature space should be treated as suspicious.

As future work, the clusters can be ranked, in order to give the user the chance to concentrate his efforts in the most risky listings. Also this method can be tested with other imbalanced datasets to see if the same results hold.

6. REFERENCES

→ V. Almendra and D. Enachescu, "Using Self-Organizing Maps for fraud prediction at online auction sites," in Proceedings of the 15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2013), (Timisoara, Romania), IEEE Computer Society, 2013.

→ www.wikipedia.com

→ Mehotra, K., Mohan, C. K., & Ranka, S. (1997). Elements of Artificial Neural Networks. MIT Press. pp. 187-202

→ Fausett, L. (1994). Fundamentals of Neural Networks.
Prentice Hall. pp. 169-187