JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AUTOENCODER TRANSACTION EVALUATION

[1] Mr Elaiyaraja P

[2]Monisha P ,[3] Pooja ,[4]Shamanth C ,[5]Vikram K N,

[1]Associate Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology ,VTU, Bengaluru, India
[2,3,4,5] B.E Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

Abstract : Imbalanced information classification issue has continuously been a well known theme within the field of machine learning investigate. In arrange to adjust the tests between larger part and minority course. Oversampling calculation is utilized to synthesize unused minority course tests, but it seem bring in commotion. Indicating to the noise issues, this paper proposed a denoising autoencoder neural organize (DAE) calculation which cannot as it were oversample minority course test through misclassification taken a toll, but it can denoise and classify the examined dataset. Through tests, compared with the denoising autoencoder neural arrange (DAE) with oversampling handle and conventional completely associated neural systems, the comes about appeared the proposed calculation progresses the classification exactness of minority lesson of imbalanced datasets.

IndexTerms **- Imbalanced data; Oversampling; Denoising autoencoder neural network; Classification**

I. INTRODUCTION:

Credit card extortion could be a growing risk with distant coming to results within the back industry, organizations and government. Extortion can be characterized as criminal deception with aim of obtaining monetary pick up. As credit card got to be the foremost popular method of instalment for both online and offline exchange, the extortion rate moreover quickens. The most reasons for extortion is due to the need of security, which includes the utilize of stolen credit card to urge cash from bank through genuine get to. This comes about in tall trouble of avoiding credit card extortion.

So how to do extortion discovery is exceptionally noteworthy. A part of investigates have been proposed to the location of such credit card extortion, which account for lion's share of credit card fakes. Identifying utilizing conventional strategy is infeasible since of the enormous information. Be that as it may, budgetary teach have focused their consideration to later computational techniques to handle credit card extortion issue.

Classification issue is one of the key inquire about subjects within the field of machine learning. Right now accessible classification strategies can as it were accomplish best execution on adjusted datasets. Be that as it may, there are a expansive number of imbalanced datasets in commonsense application. For the extortion issue, the minority lesson, which is the abnormal transaction, is more imperative [1]. For occurrence, when minority lesson accounts for less than 1 percent of the overall dataset, the in general exactness comes to more than 99% indeed in spite of the fact that all the minority course has been misclassified.

Minority course testing could be a common strategy to handle with the imbalanced information classification issue. The most reason of oversampling is to extend the number of minority lesson tests so that the unique classification data can get superior maintenance. In this manner, within the areas where there's higher request for the classification exactness, oversampling algorithm is chosen in common.

This paper looks for to actualize credit card extortion discovery utilizing denoising autoencoder and oversampling. For imbalanced information, we chosen utilize over strategy to attain legitimate show.

II. RELATED WORKS

Information mining procedure is one striking strategies utilized in understanding extortion location issue. This is often the method of distinguishing those exchanges that are have a place to fakes or not, which is based on the behaviors and propensities of cardholder, numerous methods have been connected to this region, manufactured neural arrange [2], hereditary calculation, back vector machine, visit thing set mining, choice tree, relocating feathered creatures optimization calculation, Naïve Bayes. A comparative examination of calculated relapse and Naïve Bayes is carried out in [3]. The execution of Bayesian and neural organize [4] is assessed on credit card extortion information. Choice tree, neural systems and calculated relapse are tried for their appropriateness in extortion discoveries [5].

In a workshop work, [6] proposes two progressed data mining approaches, bolster vector machines and arbitrary timberlands, at the side calculated relapse, as portion of an endeavor to superior identify credit card extortion whereas neural arrange and calculated relapse is connected on credit card extortion detection problem [7]. A number of challenges are related with credit card discovery, to be specific false behavior profile is energetic, that's false exchanges tend to see like true blue ones; credit card exchange datasets are rarely available and highly imbalanced (or skewed); ideal highlight (factors) determination for the models; reasonable metric to assess execution of procedures on skewed credit card extortion information. Credit card extortion location execution is enormously influenced by sort of inspecting approach utilized, choice of factors and discovery technique(s) utilized.

III. BACKGROUND

3.1 Autoencoder

A. Conventional Autoencoder Neural Network (AE)

Autoencoder is an counterfeit neural network utilized for unsupervised learning. The point of autoencoder is to memorize representations to recreates highlights for a set of information, regularly for the purpose of dimensionality lessening. The best shape of an autoencoder could be a feedforward, non-recurrent neural organize which is comparative to the multilayer perceptron [8]. As the figure 1 appeared, it has 2 parts: one is encoder and the other is decoder which are comprise of by an input layer, one or more covered up layers and an yield layer. The noteworthy contrast between autoencoder and multiplayer perceptron is that the yield layer of autoencoder has the same number of neurons as the input layer. The reason is to reproduce its claim inputs rather than anticipating the target esteem from the given inputs.

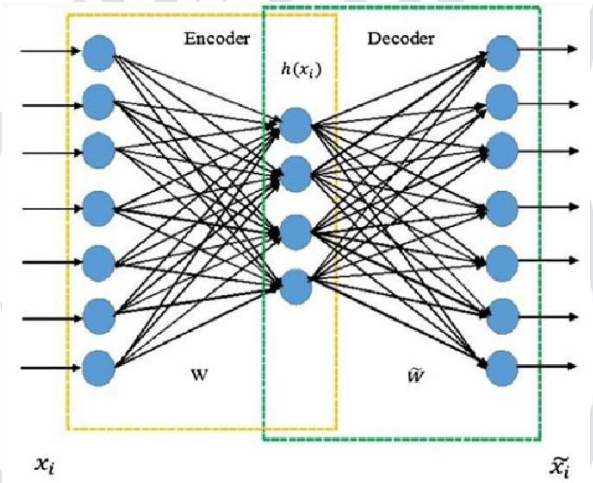


Fig. 1 architecture of autoencoder neural network

In autoencoder, the network structure has associations between layers, but has no association interior each layer, xi is input test, $\hat{x}i$ is yield include.

The preparing of autoencoder neural organize is to optimize remaking mistake utilizing the given tests. The taken a toll work of autoencoder neural organize characterized within the extend is (1)

$$J_{A,E} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{2} \ \|\hat{x}_i - x_i\|^2 \right)$$

(1) where m represents number of input samples.

B. Denoising Autoencoder Neural Network (DAE)

For human, when individuals see an question, in case there's a little portion of the question is blocked, they can still recognize it. But how the autoencoder does for the “contaminated” information? There's a variety of conventional autoencoder named denoising autoencoder which may make autoencoder neural arrange learn how to expel the noise and reproduce undisturbed input as much as conceivable [9].

As appeared in figure 2, the initial information is x, and $\tilde{x}$ is the information debased with noise. Through the total prepare of denoising autoencoder, the yield is $\hat{x}$. The misfortune function tries to play down the distinction between the yield and the initial information so that the autoencoder has the capacity of killing the impact of commotion and extricating highlights from the adulterated information. Subsequently, the highlights created from the learning of input adulterated with noise are more vigorous, which made strides the information generalization capacity of autoencoder neural arrange show to input information.

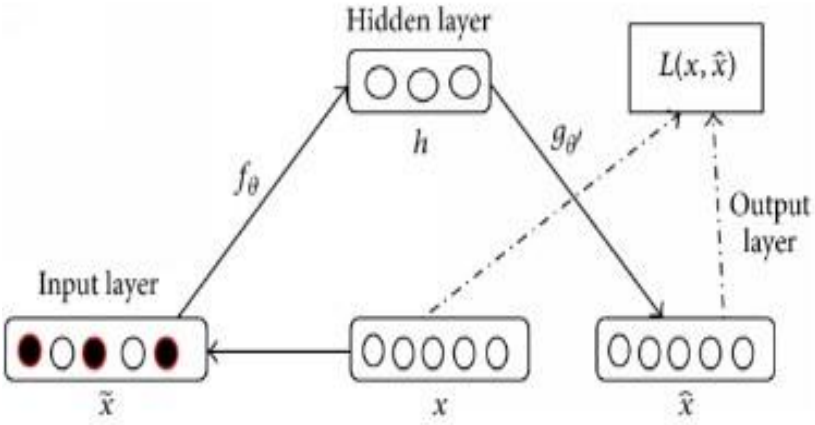


Fig. 2 Denoising autoencoder neural network

The commonly utilized clamors are Gaussian noise, and Salt and pepper noise. And the fetched work of denoising autoencoder neural arrange is characterized agreeing to (2)

$$J_{DA,E} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{2} \ \|\hat{x_i} - x_i\|^2 \right)$$

(2) where $\hat{x} = f(\sum(w\tilde{x} + b))$, w represents weights and b represents bias.

3.2 Oversampling

Imbalanced dataset could be a common issue confronted in machine learning, since most conventional machine learning classification show can't handle imbalanced dataset. Tall misclassification fetched frequently happened on minority course, since classification demonstrate will attempt to classify all the information sample to the lion's share course.

Oversampling may be a strategy utilized to bargain with imbalanced dataset, its subject to form particular course test so the lesson dissemination of the initial dataset can be adjusted. The good thing about utilizing oversampling is appeared in figure 3.

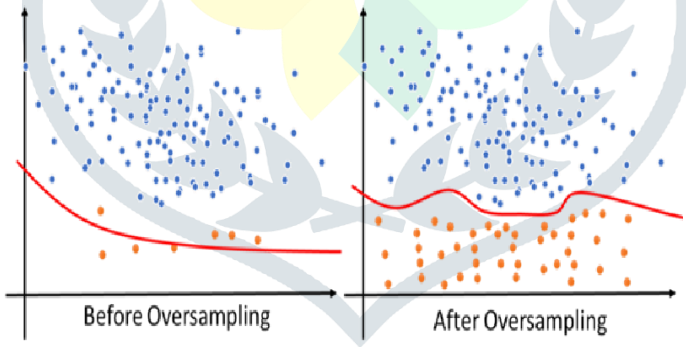


Fig. 3 Benefit of using oversampling

SMOTE (Synthetic Minority Oversampling Technique) is one of the foremost well known oversampling technique. In arrange to make a manufactured information point, first we have to be discover a k-nearest-neighbors cluster within the highlight space, at that point arbitrarily discover a point inside this cluster, at last utilizing weighted normal to “forge” the modern information point.

3.3 Classification fully connected model

Deep fully connected neural network is frequently utilized in classification problem, with SoftMax cross entropy as the loss function, deep learning classification show can accomplish exceptionally tall accuracy.

The SoftMax function is regularly utilized within the final layer of a neural network-based classifier, it first calculates the exponential esteem of each output, then normalize all the output and let the sum of the output equal to 1. SoftMax work is regularly utilized for probability distribution transformation, since the output of SoftMax function is inside range 0 to 1 that include up to 1, appeared within the equation 3,

$$\mathrm{P}(y_i|x_i; \mathrm{W}) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$$

(3)

Entropy may be a measure for data substance and may well be characterized as the unpredictability of an event. So, the greater the probability is, the littler the unpredictability is, which implies the data contents is additionally exceptionally small. In case an event occurs definitely with the probability of 100%, at that point the unpredictability and information content are 0. cross-entropy loss function takes advantages of highlight of entropy equation, cross-entropy loss function can degree the goodness of a classification demonstrate, which is appeared in equation 4,

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y_i = j\} log \frac{e^{\theta_j^T x_i}}{\sum_{i=1}^{k} e^{\theta_j^T x_i}}$$

(4)

Cross-entropy can be utilized in multi-classification problems with the combination of SoftMax (don't consider regularization). Compared with quadratic loss function, cross-entropy loss function gives way better training execution on neural systems.

3.4 Model evaluation metric

Accuracy isn't adequate to assess a classification demonstrate, particularly for imbalanced dataset. For illustration, an imbalanced dataset with 99.9% of ordinary information and 0.1% of anomalous information, in case the classification names all the test as ordinary lesson, the show can still accomplish 99.9curacy. In any case, for irregularity detection, the detection rate of anomaly class is exceptionally important. Confusion matrix is regularly utilized in this circumstance.

Classification	Actual Positive Sample	Actual Negative Sample
predict as positive	TP	FP
predict as negative	FN	TN

Table 1. Confusion matrix for two-class problem

Recall (Detection rate) is the proportion between the number of accurately identified irregularities and the entire number of irregularities, it assesses how much of the inconsistencies can be identified in this classification model

IV. METHODOLOGY

The credit card extortion exchange dataset we are utilizing is downloaded from Kaggle, with completely 28315 exchange detail and 0.5% of them are labeled as extortion, the dataset is appeared within the fig 4. The subject is to construct a classification show for irregularity location. Dataset contains as it were numerical input after doing PCA change. Highlights V1, V2, ... V28 are the foremost components, the as it were highlights which have not been changed with PCA are 'Time' and 'Amount'. Include 'Class' is the reaction variable and it takes esteem 1 in case of extortion and something else.

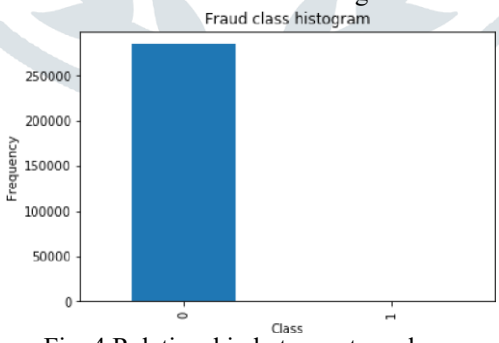


Fig. 4 Relationship between two classes

The thought is exceptionally straight forward. To begin with, utilize oversampling to convert imbalanced dataset to adjusted dataset. Then use denoised autoencoder to induce denoised dataset. At long last utilizing profound completely associated neural organize show for last classification.

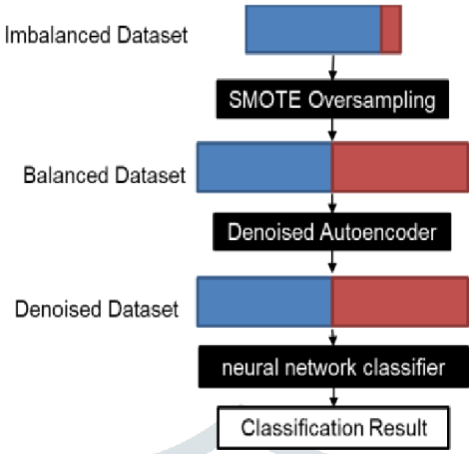


Fig. 5 Flowchart of the process

4.1 Data Preprocessing

For dataset preprocessing, drop "TIME" data, and normalized the "AMOUNT" part. Other features are obtained by PCA, don't have to be do normalization. Then select the test sample, which account for 20% of the total sample.

4.2 Oversampling

Our group as it were perform oversampling on the training dataset. Some time recently oversampling, there are total 22652 transaction records in training dataset, with 22538 samples in normal class and 114 samples in abnormal class. After oversampling, the training dataset contains 22538 samples in normal class and 22538 samples in abnormal class.

4.3 Denoising autoencoder

Our group outlined a 7 layers autoencoder for dataset denoising prepare. After we got balanced training dataset from oversampling, we include Gaussian noise to the training dataset, then feed the training dataset into this denoised autoencoder. After training this denoised autoencoder model, this autoencoder has the capability to denoise the testing dataset in the prediction process.

Dataset with noise (29)
Fully-Connected-Layer (22)
Fully-Connected-Layer (15)
Fully-Connected-Layer (10)
Fully-Connected-Layer (15)
Fully-Connected-Layer (22)
Fully-Connected-Layer (29)
Square Loss Function

Table 2. Model design for denoised autoencoder

4.4 Classifier

Our group designed a 6 layers autoencoder for dataset denoise process. After we got denoised training dataset from denoised autoencoder, we feed the training dataset into this deep fully connected neural network classifier. In the end, we are using SoftMax with cross-entropy as the loss function for final classification.

Denoised Dataset (29)
Fully-Connected-Layer (22)
Fully-Connected-Layer (15)
Fully-Connected-Layer (10)
Fully-Connected-Layer (5)
Fully-Connected-Layer (2)
SoftMax Cross Entropy Loss Function

Table 3. Model design for classifier

V. EVALUATION AND RESULTS

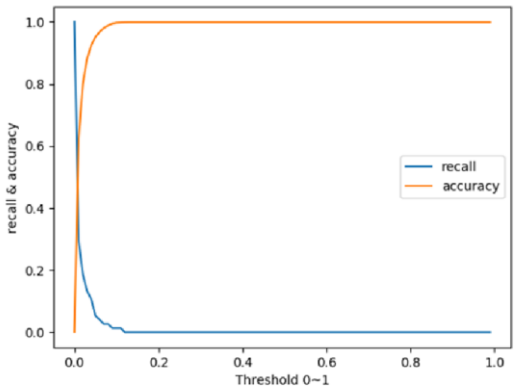
This area to begin with examines the usage points of interest, at that point presents evaluation results comparing the oversampling model with show without oversampling.

5.1 Implementation details

Our gather utilizing built-in work from "sklearn" bundle for dataset normalization, and built-in work "SMOTE" from "imblearn" bundle for oversampling. In expansion, we actualize the denoised autoencoder demonstrate and profound completely associated neural network classifier with "TensorFlow". We choose "TensorFlow" since its able of GPU acceleration. All models are trained on GTX 1060 discrete GPU w/6GB GDDR5 design memory. It took 10 minutes for each model to converge.

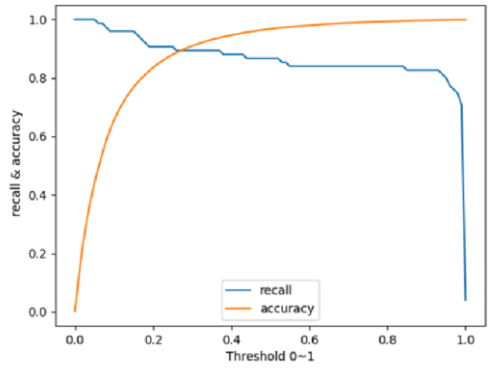
5.2 Results

After the training process, we perform assessment process utilizing another isolated evaluation dataset. the accuracy rate and recall rate are connected to assess the accuracy of each model. The results are appeared within the fig 6 and fig 7.



Model 1: Without oversampling and autoencoder

Fig. 6 Result for model 1



Model 2: With oversampling and autoencoder

Fig. 7 Result for model 2

For model 1 without the utilization of oversampling and autoencoder, the review rate is very low, since the model classifies all the test as ordinary, which implies most fraud transaction isn't detected. For model 2 with oversampling and autoencoder, the recall rate is acceptable, which means most fraud transaction can be recognized. Some evaluation result of model 2 is appeared in Table 4.

Threshold	Recall Rate	Accuracy
0.2	90.66%	83.56%
0.3	89.33%	90.93%
0.4	88%	94.58%
0.5	86.66%	96.73%
0.6	84%	97.93%

Table 4. Model 2 Evaluation Result

VI. CONCLUSION

In machine learning area, lopsidedness information classification gets expanding consideration as huge information gotten to be prevalent. On account of the downsides of conventional strategy, oversampling calculation and autoencoder can be utilized. This ponder combined stacked denoising autoencoder neural arrange with oversampling to construct the demonstrate, which can accomplish minority lesson inspecting on the premise of misclassification fetched, and denoise and classify the tested datasets. The proposed calculation increments classification exactness of minority course compared to the previous strategies, we are able accomplish diverse exactness by controlling the limit. In this consider, when edge break even with to 0.6, we will achieve the leading execution, which is 97.93%. In any case, the dimensionality lessening of high-dimensional information still got to be advance investigated.

VII. REFERENCES

[1] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," Expert Systems with Applications,vol. 40, pp. 5916-5923, 2013.

[2] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Vol. 6, No. 3, pp. 311 – 322

[3] Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2, 841848.

[4] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In Proceedings of the 1st international naiso congress on neuro fuzzy technologies (pp. 261-270).

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

[6] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613.

[7] Sahin, Y. and Duman, E., (2011). Detecting credit card fraud by ANN and logistic regression. In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 315-319). IEEE.

[8] Autoencoder for Words, Liou, C.-Y., Cheng, C.-W., Liou, J.-W., and Liou, D.-R., Neurocomputing, Volume 139, 84–96 (2014), doi:10.1016/j.neucom.2013.09.055

[9] M. Koziarski and M. Woźniak, "CCR: A combined cleaning and resampling algorithm for imbalanced data classification", International Journal of Applied Mathematics and Computer Science, vol. 27, no. 4, 2017.