Neural Network for Self-Checkout Fraud Detection

Haline Oliveira Velloso[1], **Maurício Acconcia Dias**[1]

[1]FHO-SIVA - Autonomous Vehicles and Intelligent Systems Group
Computer Engineering – Engineering Department
University Center of Herminio Ometto Foundation (FHO)
Araras – SP – Brazil

halinevelloso@alunos.fho.edu.br, macdias@fho.edu.br

Abstract. *People who use self-checkouts are not always honest and end up committing fraud with the products being purchased. These frauds are difficult to check without the procedure being embarrassing for the people involved. Considering this scenario, the main goal of this work is to design an Artificial Neural Network that is able to receive information about a determinate self-checkout procedure and indicate if there is fraud or not. Database used in this work for training and checking the ANN performance was provided by Data Mining Cup 2019. The final network architecture was able to correctly classify 90% of provided test data showing that this solution can be considered in real situations to help finding frauds on self-checkouts.*

Resumo. *Pessoas que utilizam o serviço de self-checkout nem sempre são honestas ao realizarem as compras e comentem fraudes. Estas fradues normalmente são difíceis de serem identificadas sem colocar as pessoas envolvidas em situações difíceis. Considerando este problema, o objetivo principal deste trabalho de pesquisa é o desenvolvimento de uma Rede Neural Artificial capaz de identificar a presença ou não de fraudes mediante a apresentação dos dados da compra. O banco de dados utilizado nesta pesquisa foi disponibilizado para a Data Mining Cup de 2019. A arquitetura final da RNA desenvolvida foi capaz de classificar corretamente 90% dos casos de teste mostrando que a solução obtida neste trabalho pode ser utilzada no auxílio da verificação de fraudes em self-checkouts.*

1. Introduction

Data mining has grown exponentially in the last decade and it's a consequence of the large amount of information that can be obtained through the analysis of this data. The main goal of this area is being able to obtain information related to associations, sequences, classification, clusters, and prognoses [Bhatia 2019]. Companies are investing in data collection and storage techniques since they realized how valuable data is for their future.

When companies start to collect data from every part of their sectors, the resulting data volume becomes exceptionally large and this phenomenon is known as Big Data [Buyya et al. 2016]. Consequently, the area that deals with the treatment of these data has shown a significant prominence in the scientific and corporate environment that is called Data Science and has been growing close to exponential since 2012 [Corea 2019]. In this context there are several initiatives to promote the development of Data Mining area

and professionals, known as data scientists. One of these initiatives is the Data Mining Cup[DMC 2019].

DMC is an annual competition since 2002 that university students from all over the world form teams and try to solve the problem proposed by the organization of the competition using data mining algorithms and techniques. DMC 2019 brought the self-checkout problem to be solved. A big database was provided by a large non-identified company composed by hundreds of self-checkout data together with its fraud/non-fraud classification. Since this database was available and trustworthy it was chosen.

This work's main goal was to design an Artificial Neural Network (ANN) that can receive the same information provided in DMC's database and correctly classify each one of the cases. Results showed that a relatively simple Multi-Layer Perceptron network was able to solve the problem with a hit rate of 90%. These results showed that this network can be used to help stores to identify self-checkout frauds and decrease their losses.

2. Tools and Methods

In recent years, there has been an advance in self-checkout technology in the retail sector, mainly in supermarkets and according to the consultancy Global Markets Insights it is estimated that the self-checkout market will exceed 4 billion dollars by 2024 [Wadhwani and Gankar 2019]. Retailers are able to relocate employees who would otherwise be working at the checkout for other tasks that cannot be automated but are tasks that improve the customer experience.

Database from DMC 2019 is organized with 8 inputs and one output. The inputs were: trust level of the sample, total scan time in seconds, the total purchase price, line item voids, scans without registration, quantity modifications, scanned line items per second, value per second. The output was 0 for no-fraud and 1 for fraud. The training data had about 2000 samples of real cases. These 2000 cases were extremely unbalanced (Table 1) so some modifications were proposed. It was created a new version of the database to be used for tests considering all the fraud cases and a correspondent number of non-frauds by selecting the same number of frauds and non-frauds cases (104) without no selection criteria. Some of the input data needed to be modified to numerical values. N-Fold Cross-Validation method was used to train the network with $N = 10$.

Table 1. Initial state of database from DMC 2019.

	# of Cases	# of Frauds
Training Set	1.878	104
Validation Set	498.120	23.728

The main features of the database showed that it was not necessary to design a deep neural network to solve this problem, primarily because of the small number of inputs within a small range and one binary output. In this work FANN library[Nissed 2003] was used to implement the ANN. As a recommendation for using FANN the database was normalized with input values between '0' and '1' C language considering floating point numbers precision. Network architectures started with one layer and were up to five

layers. The number of neurons in each layer started with five and were increased by five until a good result with a considerably fast convergence time was achieved.

3. Results

Initially some tests were executed to verify network's behavior according to the database configuration. Multi-Layer Perceptron (MPL) networks with 8 inputs and 1 output were designed with variations in hidden layers' amount and number of neurons. Several configurations were tested and the classification error for the balanced databases was near 40%, showing that some criteria to choose data have to be used or the raw database itself. The best networks in this case were composed by three hidden layers with 150 - 170 in the first hidden layer and 80 neurons on the second and third hidden layers.

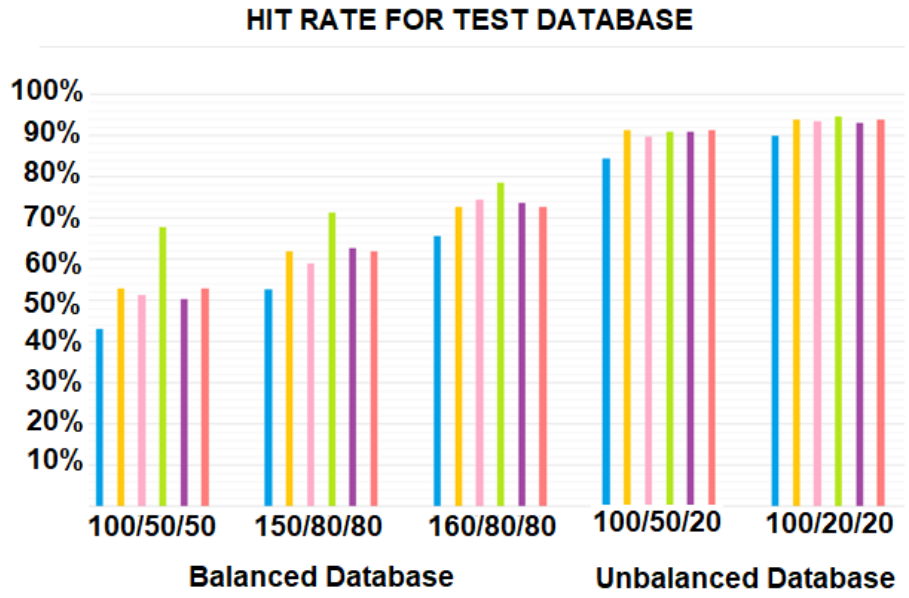

Figure 1. Designed ANNs results. Bars represent the six training/validation procedures executed for each designed network. The worst results on the left used a balanced database while good results on the right used an unbalanced database.

After these non-acceptable results, the balanced dataset was discarded and only the complete dataset was used to train and test the networks. The chosen training algorithm was the resilient backpropagation (Rprop) that has a faster convergence time because it only uses the sign of the derivative not its value [Haykin 2009]. Other parameters were learning rate of 50%, delta of 0.2 and the activation function for hidden layers was sigmoid symmetric. Best networks considering the unbalanced database were:

- 100/20/20 neurons on hidden layers achieved a Mean Square Error of 0.0005
- 100/50/20 neurons on hidden layers achieved a Mean Square Error of 0.000612
- 100/50/50 neurons on hidden layers achieved a Mean Square Error of 0.0009

The chart presented in Figure 1 shows each one of the five architectures tested considering both databases used in this work. It is possible to see that the unbalanced database achieved a significantly better result with smaller second and third hidden layers.

Presented results for a designed network are the mean of six executions for each scenario (each one represented by colored bars in Figure 1).

Bigger networks achieved an error around 0.0009 and smaller networks were not able to achieve errors below 0.5. Training time for ten entire databases was about 10 minutes considering a i7-7700HQ processor and 8GB of RAM. No parallel execution was added to FANN for these tests.

Results showed that the networks were not small but with size considerably smaller than deep neural networks. If the network architecture was being tested it was possible to notice that this database had a complex and unexpected behavior. After these training steps the network was executed using the test data from the DMA 2019. Results showed that the network able to achieve a 98% hit rate, that means it only classified 2% of the training dataset wrong. Also when only the frauds correct classification is considered, the network achieved a 90% hit rate. This result could place a team in the initial classification positions of the DMC, but this work was not able to compare achieved results to other team's results because they use to be published in data mining conferences only the following year that, in this case, is 2020.

4. Conclusion

This work presented a Artificial Neural Network to solve the self-checkout fraud detection problem. The database was obtained in Data Mining Cup 2019 website and the results showed that it is possible to train an ANN to solve this problem considering chosen parameters. Some possibilities for future works are the integration of the system in a real self-checkout to verify its accuracy, exploring different ANNs to solve the problem and compare the results and to execute the training of the same network with data from local commercial establishments.

References

Bhatia, P. (2019). *Data Mining and Data Warehousing Principles and Practical Techniques*. Cambridge University Press, 1st edition.

Buyya, R., Calheiros, R. N., and Dastjerdi, A. V. (2016). *Big Data*. Morgan Kaufmann, 1st edition.

Corea, F. (2019). *An Introduction to Data*. Springer International Publishing, 1st edition.

DMC (2019). *Data Mining Cup: international student competition*. https://www.data-mining-cup.com/.

Haykin, S. S. (2009). *Neural networks and learning machines*. Pearson Education, 3rd edition.

Nissed, S. (2003). *Implementation of a Fast Artificial Neural Network Library (FANN)*. Department of Computer Science University of Copenhagen (DIKU).

Wadhwani, P. and Gankar, S. (2019). *Self-Checkout Systems Market Size to exceed $4bn by 2024*. Global Market Insights, https://www.gminsights.com/pressrelease/self-checkout-system-market edition.