Pattern recognition of financial institutions' payment behavior

XXV Meeting of the Central Bank Researchers Network
CEMLA & Banco Central del Uruguay
October 28-30, 2020

Carlos León, Banco de la República & Tilburg University*

Paolo Barucca, University College London

Oscar Acero, Stratio (formerly at Banco de la República)

Gerardo Gage, Latin American Center for Monetary Studies

Fabio Ortega, Banco de la República

(*) Corresponding author, e-mail: cleonrin@banrep.gov.co / c.e.leonrincon@tilburguniversity.edu.

Disclaimer

Opinions and statements in this article are the sole responsibility of the authors and represent neither those of Banco de la República nor of its Board of Directors nor of the institutions the authors are affiliated to. We thank Serafín Martínez, Raúl Morales, and Deisy Zambrano for their contribution to the development of the methodology. We thank Clara Machado, Serafín Martínez and Raúl Morales for their comments and suggestions to this article. Any remaining errors are our own.

Link to the current version of the article:

https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9901/be_1130.pdf

Take home messages

- A supervised methodology to represent the payment behavior of financial institutions starting from a database of transactions in the Colombian large-value payment system.
- A feedforward artificial neural network to represent the payment patterns through 113 features corresponding to financial institutions' contribution to payments, funding habits, payments timing, payments concentration, centrality in the payments network, and systemic impact due to failure to pay.
- An out-of-sample classification error around three percent.
- The performance is robust to unsupervised feature selection.
- Network centrality and systemic impact features contribute to enhancing the performance of the methodology definitively.
- This is the first step towards the automated detection of individual financial institutions' anomalous behavior in payment systems—the failure of a good classifier as a warning sign.

Contents

Contents

Literature review

- Three strengths of ANNs for classification problems
  - They can deal with non-linear relationships between factors in the data (see Bishop, 1995; Han & Kamber, 2006; Fioramanti, 2008; Demyanyk & Hasan, 2009; Eletter, et al. 2010; Sarlin, 2014; Hagan, et al. 2014).
  - ANNs make no assumptions about the statistical distribution or properties of the data (see Zhang, et al., 1999; McNelis, 2005; Demyanyk & Hasan, 2009; Nazari & Alidadi, 2013; Sarlin, 2014).
  - Very effective classifiers, even better than the state-of-the-art models based on classical statistical methods (see Wu, 1997; Zhang, et al., 1999; McNelis, 2005; Han & Kamber, 2006).
- ANN for classification and anomaly detection in the financial domain:
  - Credit card fraud detection (see Aleskerov, et al., 1997; Ghosh & Reilly, 1994; Dorronsoro, et al., 1997).
  - Anti-money laundering (see Brause, et al., 1999).
  - To identify potential tax-evasion cases (see Wu, 1997). [...]

Literature review

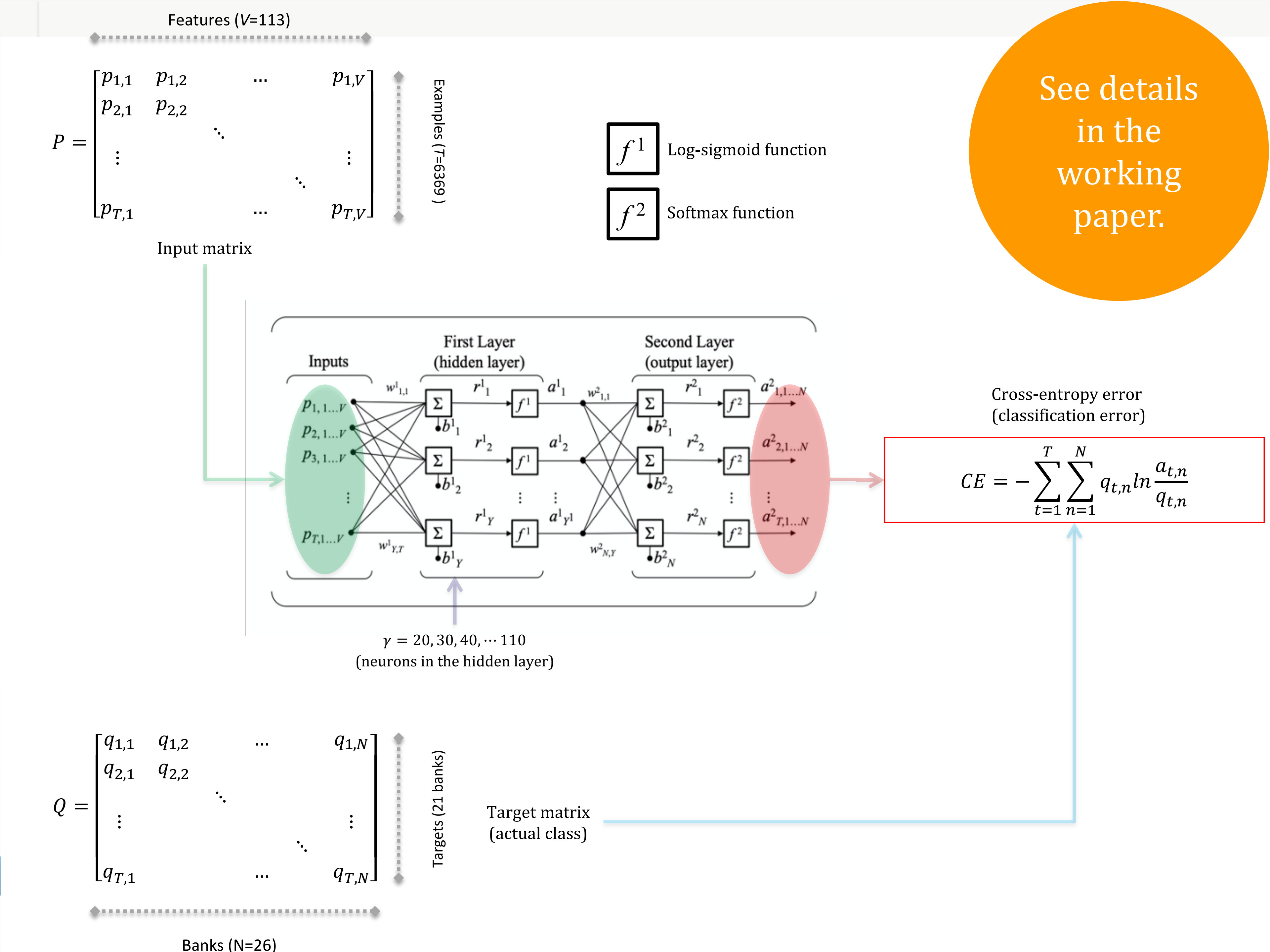
- ANN for classification and anomaly detection in the financial domain: [cont.]
  - Credit risk (see Angelini, et al., 2008; Eletter, et al., 2010; Nazari & Alidadi, 2013; Bekhet & Eletter, 2014; Tam & Kiang, 1990; Tam, 1991; Salchenberger, et al., 1992; Wilson & Sharda, 1994; Olmeda & Fernández, 1997; Zhang, et al., 1999; Atiya, 2001; Brédart, 2014).
  - Macro early-warning systems (see Fioramanti, 2008; Sarlin, 2014; Holopainen & Sarlin, 2016).
  - To classify banks as domestic or foreign (see Turkan, et al., 2011) and Islamic or conventional (see Khediri, et al., 2015).
  - To classify balance sheets into their corresponding bank (see León, et al., 2017).

- To detect anomalous payments networks (i.e. oversight of payment systems):
  - Dutch partition of TARGET2 payments networks (see Triepels, et al., 2017).
  - Canadian ACSS retail payment system networks (see Sabetti & Heijmans, 2020).

Contents

Methods

- The base case model:
  - A two-layer artificial neural network for pattern recognition on a set of 113 features that capture the behavior of 26 banking institutions participating in the Colombian large-value payment system during 2019 (total examples 6369).
  - Non-banking institutions excluded for tractability (i.e. banks are the most contributive). Results are robust to including non-banking (in Appendix).
- Feature selection (i.e. the inputs):
  - Based on payment systems literature (McAndrews & Rajan, 2000; Becher, et al., 2008; Bernal, et al., 2012; Diehl, 2013; Denbee, et al., 2014; Martínez & Cepeda, 2018), 103 features that capture behavior of financial institutions.
  - By type, those 103 traditional features aim at measuring i) contribution to payments, ii) funding habits, iii) payments timing, and iv) payments concentration.
  - Additionally, we use non-traditional features:
    - Nine features measure importance (i.e. centrality) in the payments network.
    - One feature measures the systemic footprint in case of failure (i.e. impact due to failure to make discretionary payments—simulation methods).



Features (V=113)
P =
Examples (T=6369)
Input matrix
Log-sigmoid function
Softmax function
See details in the working paper.
Inputs
First Layer (hidden layer)
Second Layer (output layer)
Cross-entropy error (classification error)
$CE = -\sum_{t=1}^{T}\sum_{n=1}^{N} q_{t,n} ln \frac{a_{t,n}}{q_{t,n}}$
$\gamma = 20, 30, 40, \cdots 110$
(neurons in the hidden layer)
Q =
Targets (21 banks)
Target matrix (actual class)
Banks (N=26)

Methods

See details in the working paper.

- Training: adjusting W and b to attain an input-output relatio
under the chosen transfer functions for a set of examples.
- How do we train? Backpropagation: W and b are modified in backwards direction, from the output layer.
- How do we avoid overfitting*? Early stopping with cross-validation: Halt the minimization process before the complexity of the solution inhibits its generalization capability.

The goal is not to memorize the training data, but to model the underlying generator of the data (Bishop, 1995)

(*) The ability to succeed at fitting in-sample but to fail at fitting out-of-sample (see Shmueli, 2010; Varian, 2014).

Methods

See details in the working paper.

Training dataset (70%, 4459)

Validation dataset (15%, 955)

Test dataset (15%, 955)

The training set is used to minimize the error between the prediction and the actual target value

The validation dataset is used simultaneously (as the neural network is trained) to check how the estimated parameters fit out-of-sample data. When validation error starts to increase (i.e. overfitting starts), the training stops.

The error obtained on the test dataset is used to check the future performance of the artificial neural network on out-of-sample data, i.e. its generalization capability.

Based on Hagan et al. (2014)

Methods

See details in the working paper.

- Dimensionality reduction on the set of features:
  - 113 features to classify 26 banks (or 111 financial institutions) may contain potentially redundant or noisy data.
  - Further reducing the number of features may contribute to test the robustness of the chosen features and the classification model.
  - Instead of subjectively discarding leading indicators, we implement principal component analysis (PCA) dimensionality reduction on the 113 selected features.
  - We build a projection of the 113 features with a variance target of ~90% (see Vishwanathan, et al., 2010, Sree & Venkata, 2014, Alpaydin, 2014, Ding & Tian, 2016, Mehta, et al., 2019).
  - We obtain a new input set of 26 features.

Methods

See details in the working paper.

- Other details:
  - A two-layer artificial neural network. Often a single hidden laye[...] necessary (see Zhang et al., 1999, Witten et al., 2011)—our results concu[...]
  - We measure the performance with the misclassification (i.e. classification error), which is the percentage of financial institutions that are incorrectly classified.
  - Besides misclassification, we report confusion matrices, i.e. square table that relates the target class (in rows) with the output class achieved by the model (in columns).
  - We try different number of neurons in the hidden layer, from 20 to 110 (in 10-neuron increments). Misclassification is low and stable after ~60 neurons.
  - As usual, to avoid issues related to the scale of features across different financial institutions and days, inputs are row normalized.
  - As results are dependent on initialization parameters (w & b) and the cross-validation partition, we run each configuration 100 times—independently.
  - We test the importance of non-traditional features (i.e. centrality in payments networks and systemic footprint by simulated failure to pay).

Contents

Results

- Base case model (113 features to classify 26 banks)
- Base case after excluding non-standard features
- Base case after feature selection by PCA
- All financial institutions (113 features to classify 111 financial institutions)

Results

- Base case model (113 features to classify 26 banks)
- Base case after excluding non-standard features
- Base case after feature selection by PCA
- All financial institutions (113 features to classify 111 financial institutions)

Results

Base case model (113 features to classify 26 banks)

Set	Number of neurons in the hidden layer									
	20	30	40	50	60	70	80	**90**	100	110
Training	1.87 (4.60)	0.99 (0.52)	0.84 (0.41)	0.84 (0.38)	0.80 (0.37)	0.80 (0.82)	0.77 (0.36)	**0.74** **(0.34)**	0.81 (0.46)	0.86 (0.85)
Validation	4.94 (4.47)	3.46 (0.69)	3.17 (0.63)	3.02 (0.60)	2.90 (0.62)	2.94 (0.95)	2.80 (0.55)	**2.64** **(0.58)**	2.70 (0.65)	2.87 (1.07)
Test	5.20 (4.47)	3.65 (0.72)	3.37 (0.67)	3.25 (0.69)	3.08 (0.58)	2.96 (0.88)	2.88 (0.50)	**2.80** **(0.50)**	2.89 (0.62)	3.07 (1.18)

Table 1. Mean classification error for different choices of the number of neurons in the hidden layer. Calculated on 100 independent training processes; standard deviation is reported in parenthesis. The lowest mean classification error in the test set is in bold.

Results

Base case model (113 features to classify 26 banks)

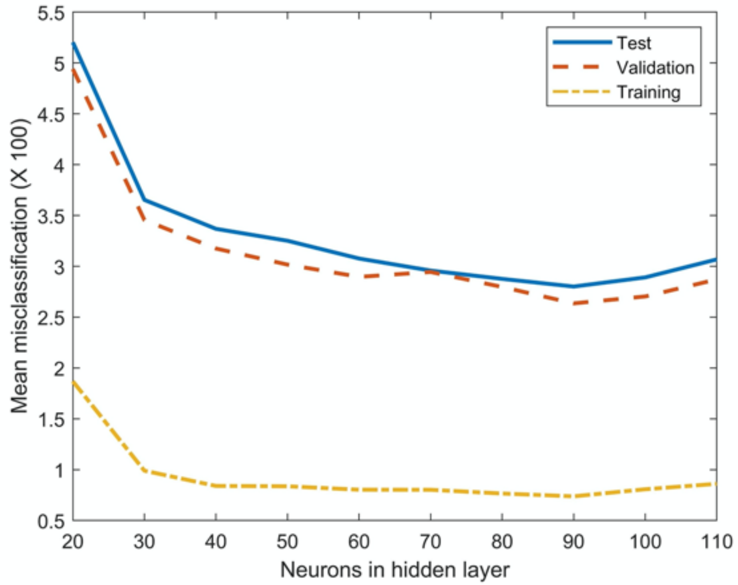


Figure 2. Mean classification error for different choices of the number of neurons in the hidden layer. Calculated on 100 independent training processes.

Results

Base case model (113 features to classify 26 banks)

Set	Number of neurons in the hidden layer									
	20	30	40	50	60	**70**	80	90	100	110
Test	2.83	2.09	1.68	1.88	1.68	**1.47**	1.68	1.78	1.78	1.57

Table 2. Lowest classification error for different choices of the number of neurons in the hidden layer. The overall lowest classification error is in bold.

Results

Base case model (113 features to classify 26 banks)

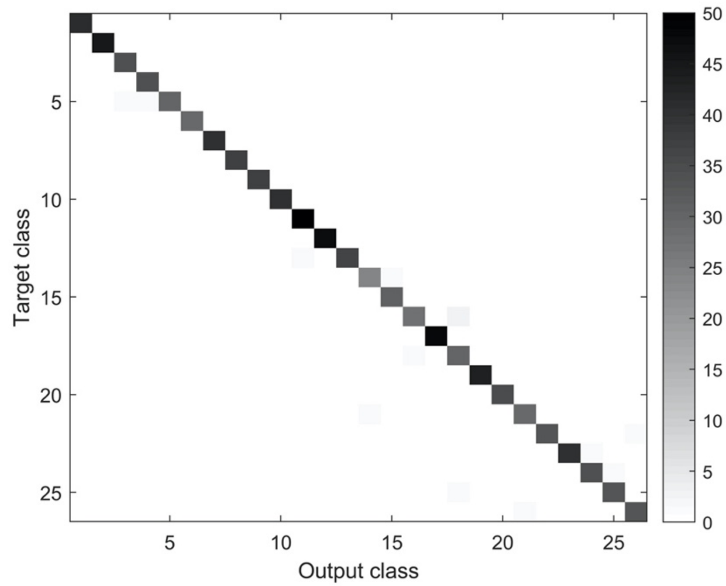


Figure 4. Confusion matrix of lowest classification error. The lowest classification error was achieved in a run with 70 neurons.

Results

- Base case model (113 features to classify 26 banks)
- Base case after excluding non-standard features
- Base case after feature selection by PCA
- All financial institutions (113 features to classify 111 financial institutions)

Results

Base case after excluding non-standard features

Set	Number of neurons in the hidden layer									
	20	30	40	50	60	70	**80**	90	100	110
Training	2.11 (0.93)	1.49 (0.65)	1.21 (0.51)	1.30 (0.98)	1.10 (0.42)	1.03 (0.43)	**1.09** **(0.41)**	1.07 (0.40)	1.08 (0.46)	1.13 (0.73)
Validation	5.76 (1.16)	4.70 (0.79)	4.10 (0.60)	3.94 (1.15)	3.74 (0.66)	3.60 (0.68)	**3.67** **(0.64)**	3.57 (0.72)	3.53 (0.83)	3.60 (0.89)
Test	6.00 (1.25)	4.65 (0.72)	4.19 (0.77)	4.15 (1.18)	3.81 (0.61)	3.84 (0.67)	**3.61** **(0.57)**	3.71 (0.64)	3.66 (0.72)	3.79 (0.92)

Table 3. Mean classification error for different choices of the number of neurons in the hidden layer, excluding network and simulation-based features. Calculated on 100 independent training processes; standard deviation is reported in parenthesis. The lowest mean classification error in the test set is in bold.

- The gain in classification performance from including network centrality and simulation-based features is about 22.44% in the lowest mean classification test error (2.80% vs. 3.61%).
- But, if we use non-standard features alone, the performance is poor (i.e. ~43% error).

Results

- Base case model (113 features to classify 26 banks)
- Base case after excluding non-standard features
- Base case after feature selection by PCA
- All financial institutions (113 features to classify 111 financial institutions)

Results

Base case after feature selection by PCA (26 features instead of 113)

Set	Number of neurons in the hidden layer									
	20	30	40	50	60	70	80	**90**	100	110
Training	4.30 (0.88)	3.73 (0.83)	3.72 (0.67)	3.62 (0.69)	3.39 (0.71)	3.36 (0.68)	3.34 (0.61)	**3.38 (0.61)**	3.32 (0.70)	3.31 (0.68)
Validation	7.05 (0.99)	6.61 (0.88)	6.32 (0.96)	6.16 (0.85)	6.03 (0.75)	5.99 (0.75)	5.80 (0.78)	**5.94 (0.74)**	5.70 (0.79)	5.79 (0.78)
Test	7.52 (0.88)	6.85 (0.76)	6.45 (0.81)	6.25 (0.76)	6.22 (0.84)	6.13 (0.73)	6.05 (0.75)	**6.19 (0.80)**	5.99 (0.85)	6.03 (0.83)

Table 4. Mean classification error for different choices of the number of neurons in the hidden layer, after feature selection. Calculated on 100 independent training processes; standard deviation is reported in parenthesis. The lowest mean classification error in the test set is in bold.

- The lowest mean misclassification error in the test set is achieved when using 90 neurons, 6.19%. This is about 2.2 times the lowest mean misclassification in the base case scenario.
- Running the base case scenario lasts about ~1.5hours (i.e. 1000 runs), whereas running the lower dimension feature matrix attained with PCA feature selection procedure lasts ~0.4 hours.

Results

- Base case model (113 features to classify 26 banks)
- Base case after excluding non-standard features
- Base case after feature selection by PCA
- All financial institutions (113 features to classify 111 financial institutions)

Results

All financial institutions (113 features to classify 111 financial institutions)

Set	Number of neurons in the hidden layer									
	20	30	40	50	60	70	80	90	100	**110**
Training	10.51 (2.55)	9.15 (0.62)	8.73 (0.55)	8.56 (0.54)	8.49 (0.48)	8.45 (0.56)	8.39 (0.44)	8.35 (0.46)	8.35 (0.48)	**8.31 (0.42)**
Validation	13.76 (2.39)	12.25 (0.66)	11.79 (0.64)	11.52 (0.53)	11.37 (0.64)	11.21 (0.57)	11.23 (0.50)	11.06 (0.54)	11.01 (0.54)	**11.00 (0.51)**
Test	13.80 (2.29)	12.38 (0.72)	11.83 (0.73)	11.55 (0.55)	11.46 (0.64)	11.43 (0.60)	11.18 (0.61)	11.21 (0.56)	11.22 (0.44)	**11.09 (0.57)**

Table A1. Mean classification error for different choices of the number of neurons in the hidden layer, including all financial institutions. Calculated on 100 independent training processes; standard deviation is reported in parenthesis. The lowest mean classification error in the test set is in bold.

Results

All financial institutions (113 features to classify 111 financial institutions)

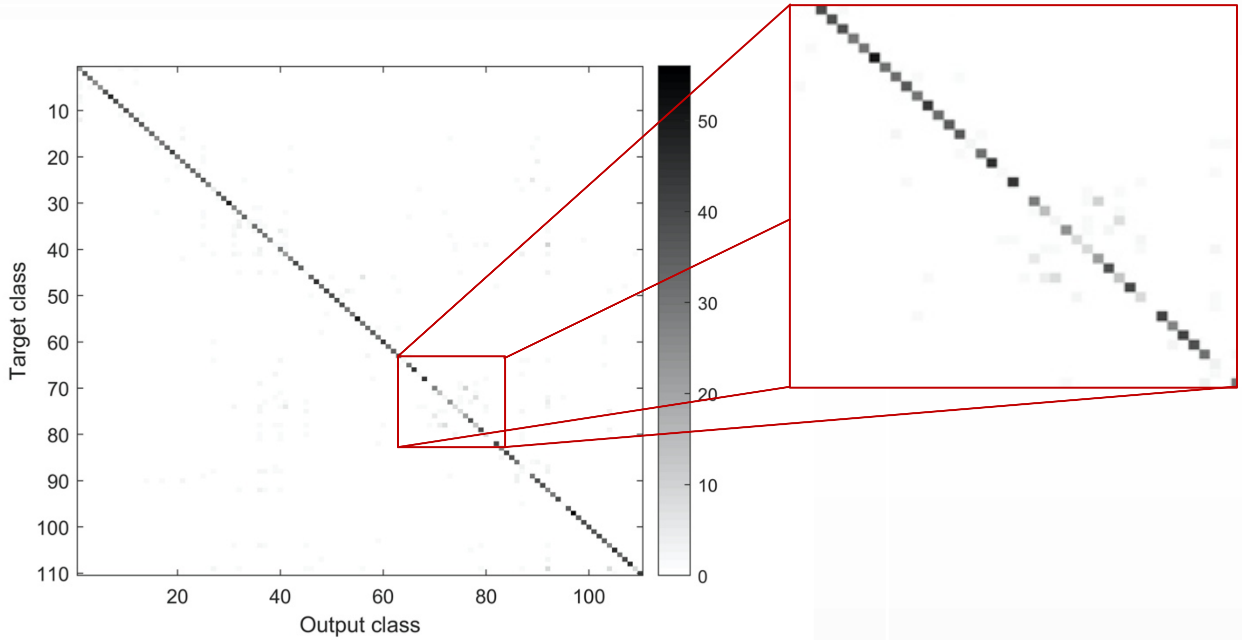


Figure A3. Confusion matrix of lowest classification error, including all financial institutions. The lowest classification error was achieved in a run with 80 neurons.

Contents

Conclusions

About the model…

- We achieve high-performance out-of-sample classification, with ~3% error.
- Stable performance after ~60 neurons.
- Robustness in the form of good (yet lower) performance when implementing a PCA feature selection procedure.
- Additionally, we test that network centrality and systemic impact features contribute to enhancing the performance of the methodology definitively.
- Including non-banking institutions increases classification error. And errors are clustered in non-banking institutions. But classification performance is still good.

Conclusions

As a monitoring tool for anomaly detection…

- A sizable change in the ability of the model to classify a financial institution is a signal of a change in its behavior within the payment system.
- Variations in individual or joint classification performance may be used as warning signals of behavioral changes that should be further studied.

But first, some challenges are to be addressed…

- Deciding on the neural network's training frequency.
- Deciding on a threshold to determine what a sizable change in individual classification performance is.

Conclusions

Promising results set the path for a new research project…

- As in most ANN models, the importance of the features is concealed.
- Other machine learning methods could shed some light on the features' importance and interactions.
- Random forest models would enable us to further understand how features drive the classification process.

Pattern recognition of financial institutions' payment behavior