Neural Network Approach Applied for Classification in Business and Trade Statistics

Jana Juriová[1]

Abstract This research is oriented on analysing data in business and trade statistics using one of the soft computing methods – neural networks. Their advantage is that they can deal efficiently with huge databases and in the case of classification are especially suitable when the borders of classes are not exactly defined. By means of the proposed classification approach two classification problems are solved. A multilayer perceptron neural network model is used for the classification in both cases. The first topic relates to data for small enterprises and presents the simplified version of classification approach for data analysis in the field of statistics. The second topic relates to the classification of Intrastat data. The proposed neural network identifies the most similar class for each statistical unit and this enables the imputation of missing values. This research suggests also the possibility to use neural networks for further improvement of Intrastat data, specifically for forecasting delayed Intrastat data, as in the time Intrastat data are published, cca 10% of them have to be estimated.

1 Introduction

Neural networks (NNs) belong among soft-computing tools that can be used for effective analysis of large databases. Their big advantage is the ability to generalize from abstract and this function can be in general used e.g. for recognizing patterns in the presence of noise or for classification of data when the borders of classes are not exactly defined. The advantages of this technique can be taken also by statistical institutes that have been collecting and storing vast amount of data (survey data or administrative data). Gaining new information from collected data can reduce further potential response burden. However, performance of the NNs is mostly dependent on the success of training process (Kulluk et al., 2012). The process of training a NN is generally interested in adjusting the individual weights between each of the individual neurons. At the beginning of the learning process a dataset, which is named as a training set, is presented to the inputs to determine the correct outputs. When the learning process is finished, a testing dataset is used to evaluate the generalization capability of the classifier.

2 Feed-forward neural networks

Feed-forward NNs, which are also known as Multi-Layer Perceptrons (MLP), are one of the most popular and most widely used NNs models in many practical applications due to their high capability to forecasting (e.g. in Farzan, 2006) and classification. In feed-forward neural networks the signal is being spread from input neurons (neurons whose inputs are signals from environment) through hidden neurons (neurons that are connected by its inputs and also outputs with other neurons) towards output neurons (neurons whose outputs lead into environment). Thus, in this network the information moves in only one direction, forward and there are no cycles or loops in this network. More on perceptrons can be found in (Rojas, 1996).

3 Classification in business statistics

A three-layer feed forward neural network is constructed and trained for classification of data into two groups. The input data are individual data surveyed for small enterprises. In this example two indicators are used: the registered number of employees and the registered number of employees - women. The reason for two indicators is the possibility to depict results in two-dimensional space. For the construction and training of the neural network individual data from two branches (according to the classification NACE II) are used, and then two classification groups are created. The first group of statistical units is of the Warehousing and storage - NACE 52 (Figure 1) and the second group of the Motion picture, video and television - NACE 59 (Figure 2). This way each group or class is characterized by a certain structure of employees: number of all employees and number of employed women.

Figure 1: *Structure of employees*

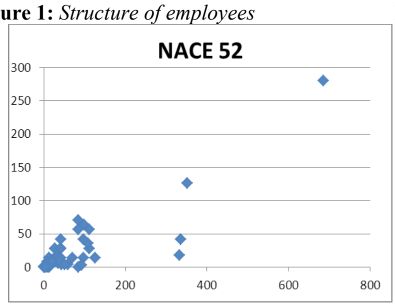
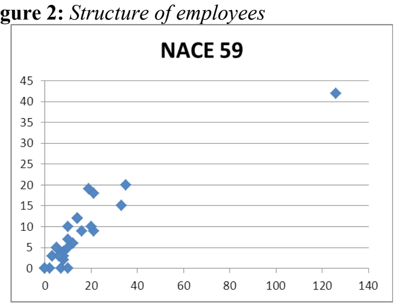


Figure 2: *Structure of employees*



Firstly a three-layer neural network is created with two inputs - data from two-dimensional space, two outputs - two classification groups and ten hidden neurons. As the output function of neurons the logistic function was used. After the construction of neural network its training for classification into two groups can be conducted. The trained network can be used for classification of original data; this step stands for verification of the model. The borders of groups created by neural network are not exact as the neural network computes probabilities, with which a certain unit belongs to each class. On the Figure 3 the function (green line) is depicted that divides the two-dimensional space into 2 classes (red and blue points). As we can see the borders are not really exact. However, our testing data are not so large data sets, but they are data with outliers (for the reason of simplicity). The probability of inclusion into a certain class is relatively high (it oscillates around 80 per cent).

Once the network is trained new data can be run through it and classified. The network will classify new data based on the previous data it was trained with. If an exact match cannot be found, it will match with the closest pattern in memory. In our case e.g. data from other branch can be chosen and we can analyze, to which of two classes it belongs. The result is the information on similarity of employee's structure of both groups. For example we have data from the branch of Transport - NACE 49 (Figure 4) and we would like to analyze, to which of the two branches mentioned above it is more similar. After applying the constructed neural network it is obvious that the branch of Transport is more similar by its employees' structure to the first class - the branch of Warehousing and storage. The probability is in average 91% for the first class.

Figure 3: Border*s of classes*

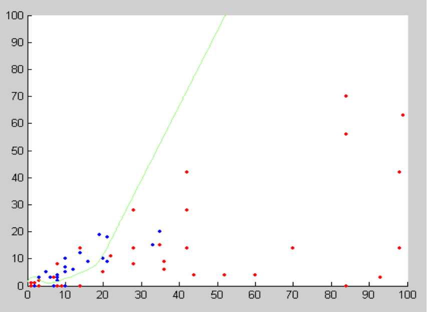
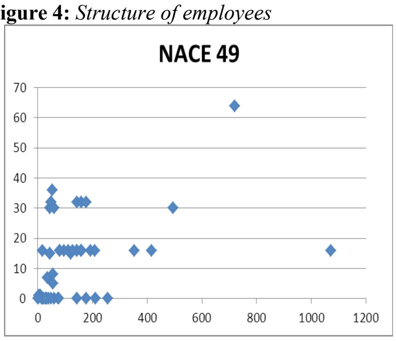


Figure 4: *Structure of employees*



The presented approach can be enlarged for more indicators and more classification groups and this is described in the following chapter.

4 Classification approach applied for imputing missing values in Intrastat system

The proposed classification approach is tested also on surveyed Intrastat data – data on foreign trade. Monthly data are available for every month of year 2009. The subject of analysis are data on dispatches from the full declaration (if a company reaches turnover from dispatches above the exemption threshold 1 700 000 EUR, it has to fill the full declaration). Data are classified into 8 groups according to the region of origin. The NN proposed in the previous chapter is used, defined with 9 inputs which represent different surveyed items for individual business reports (identification number of company, time period, number of items in report, serial number of item, code of item, invoiced sum, region, net weight, amount of goods).

The NN is trained for the classification into 8 regions using the groups with filled values of region. In this case the softmax function is used, because it is more suitable for classification into more than 2 classes (the softmax activation function is a biologically plausible approximation to the maximum operation). The trained network is used for the classification of the original data to verify the proposed classifier. The probabilities of including into particular regions are given in the Table 1.

Table 1: *Probabilities of inclusion into regions computed by NN*

Region	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***
probability (%)	25	13	14	14	18	7	4	5

The computed probabilities of inclusion into classes are very low and therefore the proposed NN is not suitable for imputing missing values. The reasons have to be further researched. But one of the factors very incomplete dataset could be in this case, as the Intrastat microdata are full of holes.

5 Conclusion

The proposed NN proved to be an appropriate tool for analysis of data from business statistics. However, in the case of huge incomplete datasets (Intrastat data system) this classification approach is not appropriately proposed yet. This research suggests also the possibility to use neural networks for further improvement of Intrastat data, specifically for forecasting delayed Intrastat data, as in the time Intrastat data are published, cca 10% of them have to be estimated.

The research reported herein was funded by the European Commission via the Seventh Framework Programme for Research (FP7/2007-2013) under Grant agreement n°244767.
This work was supported by the Slovak Research and Development Agency under the contract No. DO7RP-0024-10.

References

1. Aminian, F., Suarez, E.D., Aminian, M., Walz, D.T.: Forecasting Economic Data with Neural Networks. In: Computational Economics 28, pp.71-88. Springer (2006)
2. Kulluk, S., Ozbakir, L., Baykasoglu, Ad.: Training neural networks with harmony search algorithms for classification problems. In: Engineering Applications of Artificial Intelligence 25, pp. 11-19. Elsevier (2012)
3. Rojas R.: Neural Networks: A Systematic Introduction, Springer-Verlag, Berlin (1996)