ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Finding Originally Mislabels with MD-ELM

Anton Akusok[1], David Veganzones[2], Yoan Miche[1],
Eric Severin[4] and Amaury Lendasse[1,2,3,5]

1- Dept. of Information and Computer Science,
Aalto University School of Science, FI-00076, Finland

2- Dept. of Computer Science & Artificial Intelligence, Univ. del Pais Vasco,
Donostia/San Sebastian, Spain

3- IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

4- University of Lille 1, IAE, 104 avenue du peuple Belge,
59043 Lille, France

5- Arcada Univ. of Applied Science, Helsinki, Finland

Abstract. This paper presents a methodology which aims at detecting mislabeled samples, with a practical example in the field of bankruptcy prediction. Mislabeled samples are found in many classification problems and can bias the training of the desired classifier. This paper proposes a new method based on Extreme Learning Machine (ELM) which allows for identification of the most probable mislabeled samples. Two datasets are used in order to validate and test the proposed methodology: a toy example (XOR problem) and a real dataset from corporate finance (bankruptcy prediction).

1 Introduction

There is nowadays a large choice in Machine Learning techniques that can be used for the problem of classification. The focus of this paper is on the reliability of the data, and more precisely in that of the output. For various possible reasons, mislabeled data often exist in real life data sets, e.g. due to a human error or technical difficulties in the acquisition process. Subsequently, any machine learning method using this mislabeled data is building a model based on it, and perpetrates the learned mislabels from the training to the test. This problem becomes even more important with a limited amount of training samples. In this situation, there might not be enough correctly labeled samples in order to compensate for the mislabeled ones. Most general approaches for detecting outliers and mislabeled samples use a majority vote or consensus of results of several different classifiers [1, 2, 3]. In general, finding the mislabeled samples is not a trivial task to solve, as the model used to identify these mislabels still needs to be trained on that same data, with mislabels omitted. This paper proposes a new methodology which focuses on the case of binary classification (i.e. where each sample has only two possible labels, and thus only one possible re-labeling). The general assumption of the proposed methodology is that by correctly re-labeling a number of mislabeled samples, the error in generalization will decrease. and thus enable to identify the originally mislabeled samples. The Extreme Learning Machine (ELM) model for classification is a key part of the method, as it is a nonlinear universal approximator [4], yet a Leave-One-Out (LOO) Mean Square Error (MSE) for a given set of mislabeled samples is calculated with a direct and exact formula. The Mislabeled samples Detection ELM (MD-ELM) method works as follows. An ELM is trained using the initial labels, and a baseline LOO error is calculated. Then labels of different random sets of samples are flipped, and the LOO error is re-calculated using a PRESS [5, 6] error formula for each of the flips. If the result error is lower than the baseline one, the selected samples are marked as possibly mislabeled. After several iterations, it is possible to build an histogram depicting the frequency for each sample to have been identified as mislabeled. Samples with a high-enough frequency are deemed to be actual mislabels, then. The method can use several different ELMs to explore multiple views on the classification problem and avoid overfitting. The method is presented formally in the following section 2. The Experimental results section 3 presents an application of the MD-ELM to two datasets: a benchmark and a real-world dataset in the field of Bankruptcy prediction. The summary, conclusions and further works are presented in the Conclusions section.

2 Methodology

The proposed methodology is based on the idea that the generalization error of a chosen model decreases (estimated by its LOO error in this case) if some of the mislabeled samples of the training set have their labels changed to the actual correct class, *i.e.* the class they should have had. In fact, mislabeled samples can be misclassified themselves and/or influence the classification of other correctly labeled samples.

Denote by $\mathbf{X} = \{\mathbf{x}_i\}_{1 \leq i \leq N}$ the dataset of input samples, $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y} = \{y_i\}_{1 \leq i \leq N}$ the output labels, with $y_i \in C$ with C the set of all possible classes. In the following, the special case of binary classification, $C = \{-1, 1\}$ (which implies that there is only one possible change of label for each sample), is of concern. The Machine Learning method considered is the Extreme Learning Machine (ELM) [4] for its low computational cost: as the details of the proposed methodology highlight, there is a need for a large number of iterations using several models, and therefore, a very fast model is needed. The details of the methodology are given in subsection 2.2.

2.1 Extreme Learning Machine and LOO Error

The Extreme Learning Machine (ELM) [4] is a single-layer feedforward neural network with fixed random weights and biases. The hidden layer output is gathered to the matrix $\mathbf{H}$ (see [4] for details), and the output layer matrix β is computed as $\beta = \mathbf{H}^{\dagger}\mathbf{y}$, where $\mathbf{y}$ is the vector of outputs and $\mathbf{H}^{\dagger}$ stands for the Moore-Penrose pseudo-inverse of $\mathbf{H}$ [7]. Due to the linearity between the hidden neurons and the outputs, there exists a direct and exact formula for finding the PRESS (Prediction Sum of Squares) Leave-One-Out [6, 5, 8] error of the ELM:

$$\mathrm{MSE}_{\mathrm{PRESS}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \mathbf{h}_i\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}}{1-\mathbf{h}_i\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{h}_i^T}\right)^2, \tag{1}$$

with $\mathbf{H} = \{\mathbf{h}_i\}_{1\leq i\leq N}$, $\mathbf{h}_i \in \mathbb{R}^n$ (n being the number of neurons in the ELM model). This means that each observation is "predicted" using the other $N-1$ observations and the residuals are finally squared and summed up. A fast matrix-oriented algorithm for the evaluation of $\mathrm{MSE}_{\mathrm{PRESS}}$ can be found in [9]. The $\mathrm{MSE}_{\mathrm{PRESS}}$ formula is well suited for finding mislabeled samples by using the Leave-One-Out MSE, because the products $\mathbf{h}_i\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ are calculated only once, and used repeatedly in Eq. 1 for calculating $\mathrm{MSE}_{\mathrm{PRESS}}$ for different permutations of labels within $\mathbf{y}$. Indeed, permutations of values in $\mathbf{y}$ mean that only the last part of the matrix product $\mathbf{h}_i\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$ need be recomputed.

2.2 Finding mislabeled samples with ELM

The methodology proposed in this paper is named MD-ELM, for Mislabeled-Detecting-ELM, for which the overall algorithm is presented in algorithm 1. It uses several ELMs to circumvent possible effects of random initialization.

Algorithm 1 Algorithm of the proposed MD-ELM.

```
Inputs $\mathbf{X} \in \mathbb{R}^{N\times d}$ and outputs $\mathbf{y} \in C^N$
Set $K$ the maximum amount of mislabelled samples
Set $\mathbf{q} \in \mathbb{N}^N$ a vector holding the mislabel frequencies for each sample
Set $L$ the desired number of improvements found by each ELM
for ELM in a set of ELMs do
    Calculate $\mathbf{h}_i\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ (fixed for that ELM)
    Calculate the baseline $\underline{\mathrm{MSE}}_{\mathrm{PRESS}}$ using 1 and the initial $\mathbf{y}$
    Initialize number of improvements $l = 0$
    while $l < L$ do
        Select random $k^*$ s.t. $1 \leq k^* \leq K$
        Get $\mathbf{y}^*$ by switching randomly a set $s^*$ of $k^*$ labels in $\mathbf{y}$
        Get $\mathrm{MSE}^*_{\mathrm{PRESS}}$ using $\mathbf{y}^*$ instead of $\mathbf{y}$
        if $\mathrm{MSE}^*_{\mathrm{PRESS}} < \underline{\mathrm{MSE}}_{\mathrm{PRESS}}$ then
            Update frequencies in $\mathbf{q}$ as $q_i = q_i + 1, \forall i \in s^*$
            $l = l + 1$
        end if
    end while
end for
Select mislabelled samples by a threshold on the frequencies $\mathbf{q}$
```

In plain text, at the end of algorithm 1, the vector $\mathbf{q} \in \mathbb{N}^N$ holds for each sample the number of times that a flip of the label of this sample has resulted in an improvement of the Leave-One-Out MSE. Which means that the samples with the highest counts in $\mathbf{q}$ are the ones for which a flip of the output label is the most beneficial. Of the main ideas of the methodology is thus to use the computational speed of the calculation of the MSE for the Leave-One-Out case of the ELM to perform random flips of the class of the samples present in the training set. The frequency count of which flips resulted in better LOO MSE enables the identification of statistically likely to be mislabeled samples. The next section presents two applications of the method: one for the artificial XOR toy problem to show the correctness and feasibility of the methodology, and the other to a real-world financial problem of bankruptcy prediction.

3 Experimental results

3.1 XOR problem

XOR problem is a well-known benchmark classification problem, which cannot be solved linearly. The given problem is to predict the sign of a product of two scalars, both taken uniformly from the interval $[-1, 1]$. 500 samples are drawn in total for $\mathbf{X}$, out of which 50 output labels were switched on purpose in the dataset — and have therefore the wrong class. The obtained frequencies for each sample and originally mislabeled samples are shown on Figure 1.

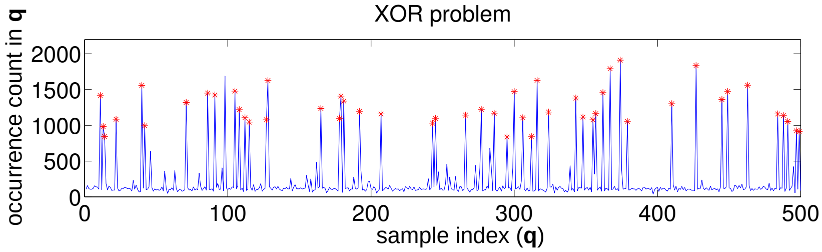

Fig. 1: XOR sample occurrence count in $\mathbf{q}$. True mislabeled samples are identified by red stars (the 50 labels that were flipped on purpose).

The method succeeds in finding 49 mislabeled samples out of 50. Mislabeled sample #295 is the only one which was not selected, and sample #98 is the correct one selected as a mislabeled. But these samples have input values of $(-0.02, 0.47)$ and $(-0.06, -0.01)$ respectively, so they both lie on a borderline or center of the XOR, which makes their classification difficult.

3.2 Bankruptcy prediction

The ability to predict bankruptcy of a firm is crucial for an investor or a creditor (bank) who wishes to ensure that he will be reimbursed later on. This experiment adopts binary classification to label the firms. An healthy company means that it is able to reimburse its debt and it has continuity and future. However, a bankrupted company is one that is unable to meet its financial obligations. In other words, it cannot pay back its debtors and begin a liquidation process that stands for sale or cessation of the company. The data set was built by du Jardin [10]. and includes 500 firms from year 2002. In the data set, the proportion of healthy and bankrupted firms is 50 : 50 and the firms are all from the trade sector. Before archiving the frequencies of each sample, variable selection was applied [11] with 7 variables selected for the training. The histogram of the frequencies **q** as for the previous experiment is shown on Figure 2.

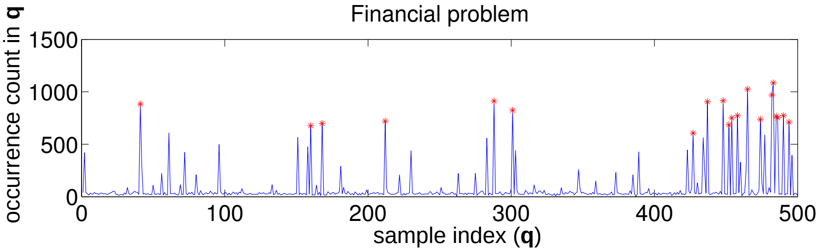

Fig. 2: Bankruptcy prediction occurrence count and expert-identified mislabeled samples.

The method shows 20 samples with high frequency that may be mislabeled. Those samples have been analyzed by two independent financial experts. For the selected samples #41, #212, #437 and #448, both experts consider that the samples are surely mislabeled in the first place. For the most of samples, one out of two experts considers that they are mislabeled. This is the case for the selected samples:·#160, #168, #301, #427, #458, #465, #474, #482, #483, #485, #490 and #494. For #288, #452, #454 and #486, the experts do not consider these selected samples as mislabeled. Taking into account the experts classification, the proposed method seems to be successful in 16 mislabeled samples out of 20. These selected mislabel samples will be investigated in detail in the future by other financial experts using more information about the selected companies. It can be considered that the proposed methodology is successful since only 20 companies have to been analyzed furthermore instead of the initial 500 companies, to identify mislabels.

4 Conclusions

A new method for finding originally mislabeled samples in a dataset is proposed for the problem of binary classification. It utilizes ELM as an extremely fast and nonlinear model for which the MSE of the Leave-One-Out can be computed at almost no cost, and then tries different combinations of re-labeling to find the mislabeled data points. The power of the method comes from testing combinations of re-labeled points instead of points separately, which captures correlations in data. Despite factorial amount of possible combinations of samples, the method shows good results in reasonable computation time, due to an extremely fast LOO error calculation with ELM. The method was tested on the XOR artificial benchmark problem, where the task was to predict the sign of a product of two numbers chosen randomly from interval $[-1, 1]$. The method successfully found 49 out of 50 mislabeled samples. For real dataset in the field of bankruptcy, the results are very promising and are helping greatly to decrease the work of financial experts that have to analyze the given datasets. In the further work the method will be applied to more diverse range of datasets to find the areas with maximal benefits provided by correctly re-labeled samples. An extension of the MD-ELM to Big Data with hundreds of thousands of samples will be explored, as such datasets, being labeled by humans, often suffer greatly from mislabeling.

References

[1] C. Brodley and M. Friedl. Identifying mislabeled training data. *arXiv preprint arXiv:1106.0219*, 2011.

[2] D. Guan, W. Yuan, Y.K.Lee, and S. Lee. Identifying mislabeled training data with the aid of unlabeled data. *Applied Intelligence*, 35(3):345–358, 2011.

[3] A. Guillen, L.J. Herrera, G. Rubio, H. Pomares, A. Lendasse, and I. Rojas. New method for instance or prototype selection using mutual information in time series prediction. *Neurocomputing*, 73(10-12):2030–2038, 2010.

[4] G.B. Huang, L. Chen, and C.K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. on Neural Networks*, 17(4):879–892, 2006.

[5] R.H. Myers. *Classical and modern regression with applications*, volume 2. Duxbury Press Belmont, CA, 1990.

[6] G. Bontempi, M. Birattari, and H. Bersini. Recursive lazy learning for modeling and control. In *Machine Learning: ECML-98*, pages 292–303. Springer, 1998.

[7] C. Rao and S.K. Mitra. *Generalized inverse of matrices and its applications*, volume 7. Wiley New York, 1971.

[8] D. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

[9] Y. Miche, M. van Heeswijk, P. Bas, O. Simula, and A. Lendasse. Trop-elm: A double-regularized ELM using LARS and tikhonov regularization. *Neurocomputing*, 74(16):2413 – 2421, 2011.

[10] P. du Jardin. *Prévision de la défaillance et réseaux de neurones: l'apport des méthodes numériques de sélection de variables*. Université de Nice-Sophia-Antipolis, 2007.

[11] L. Kainulainen, Y. Miche, E. Eirola, Q. Yu, B. Frénay, E. Séverin, and A. Lendasse. Ensembles of local linear models for bankruptcy analysis and prediction. *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, 4(2), November 2011.