Ensemble methods with non-negative matrix factorization for non-payment prevention system

RYSZARD SZUPILUK[1,2], PIOTR WOJEWNIK[1,2], TOMASZ ZĄBKOWSKI[1,3]
[1]Polska Telefonia Cyfrowa Ltd., Al. Jerozolimskie 181, 02-222 Warsaw
[2]Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw
[3]Warsaw Agricultural University, Nowoursynowska 159, 02-787 Warsaw
POLAND

Abstract: A well designed and reliable prevention system for non-payment event is very important for the telecom company. Monitoring is especially needed in case the client exceeds the level of his standard payments what can lead to his financial problems. In this paper, we propose a system describing client's behavior and informing about possible problems in advance. In this approach we apply novel ensemble methods to integrate information from many models predicting the customer's behaviour. The ensemble methods base on non-negative matrix factorization which allows identifying the fundamental prediction components. The practical experiment with prevention system confirmed that proposed procedure.

Key-Words: Ensemble methods, non-negative matrix factorization, customer data modeling, prediction.

1 Introduction

One of the most important tasks in the company where the information systems are introduced is online monitoring of individual customer behaviour. In this paper we focus on the prevention system for non-payment event. A key issue is to build an appropriate model describing client's behaviour. Assuming that different methods can model the customer behaviour in a slightly different way it seems natural to integrate and to use the information generated by many models [7]. From analytical point of view the presented methodology can be treated as ensemble methods for prediction improvement. Usually solutions of ensemble methods propose the combination of a few models by mixing their results or parameters [1,8,16]. In this paper we propose an alternative concept based on the assumption that prediction results contain the latent destructive and constructive components common to all the model results [14,15]. The elimination of the destructive ones should improve the final results. To find the latent components we apply a multidimensional decompositions with non-negative matrix factorization (NMF) [6,11].

The method will be described in the framework of a flexible system for adapting and managing the dunning process, but can be applied as ensemble method to any regression problem.

2 Prediction improvement

We assume that we test many models eg. neural networks for prediction customer behaviour. Next, we assume that each prediction result includes two types of components: constructive associated with the target and destructive associated with the inaccurate learning data, individual properties of models, missing data, not precise parameter estimation, distribution assumptions etc. We collect particular model results together $\mathbf{x}_i \in \Re^{N\times 1}$, where N is number of observations, and treat them as multivariate variable $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$, $\mathbf{X} \in R^{m\times N}$. In similar way we assume that the set of basis components is represented by $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n]^T$, $\mathbf{S} \in R^{n\times N}$. The relation between observed prediction results and latent basis components is expressed by

$$\mathbf{X} = \mathbf{AS}, \qquad (1)$$

and means matrix X factorisation by basis components matrix S and mixing matrix A. Our aim is to find such mixing matrix A and unknown basis components set that matrix S (after rows reordering) can be described as

$$\mathbf{S} = \left[\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, ..., \hat{\mathbf{s}}_p, \bar{\mathbf{s}}_{p+1}, \bar{\mathbf{s}}_{p+2}, ..., \bar{\mathbf{s}}_n\right]^T, \qquad (2)$$

where $\hat{\mathbf{s}}_i \in R^{N\times 1}$ is i-th constructive component, $\bar{\mathbf{s}}_i \in R^{N\times 1}$ is i-th destructive component. After basic components are classified into destructive and constructive ones we can reject the destructive components (replace them with zero) to obtain only constructive basis components matrix $\hat{\mathbf{S}} = \left[\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, ..., \hat{\mathbf{s}}_p, \mathbf{0}_{p+1}, \mathbf{0}_{p+2}, ..., \mathbf{0}_n\right]^T$. Now we can mix the cleaned basis results back to obtain improved prediction results

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\mathbf{S}} \quad . \tag{3}$$

The replacement of destructive signal by zero in (2) is equivalent to putting zero in corresponding column of A. If we express the mixing matrix as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n]$ the purified results can be described as

$$\hat{\mathbf{X}} = \mathbf{A}\hat{\mathbf{S}} = \hat{\mathbf{A}}\mathbf{S} \tag{4}$$

where $\hat{\mathbf{A}} = \left[\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p, \mathbf{0}_{p+1}, \mathbf{0}_{p+2}, ..., \mathbf{0}_n\right]$.

The effectiveness of the method highly depends on the application of proper transformation providing searched basis components and next it is important to perform proper distinction $\hat{\mathbf{s}}_i$ from $\bar{\mathbf{s}}_i$. Among many possible transformations leading to basis signals we focus on blind signals separation (BSS) methods [5,9].

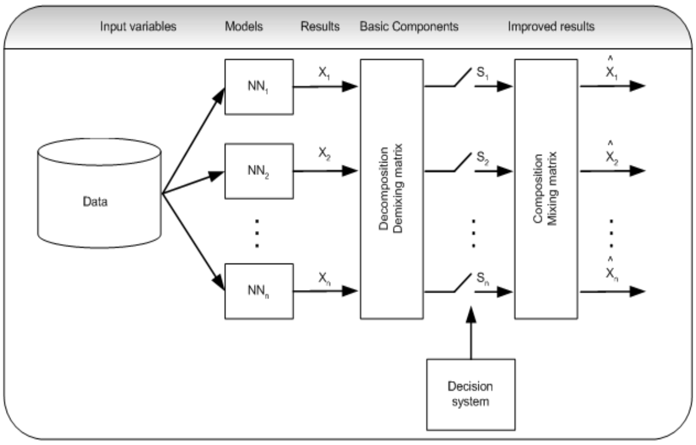

Fig. 1. System for integration of neural network results

3 The blind signal separation and data representation

Blind signals separation (BSS) methods aim at identification of the unknown signals mixed in the unknown system [4,5,9]. There are many different methods and algorithms used in BSS task. They explore different properties of data like: independence [4,9], decorrelation [5,10], sparsity [12], smoothness [5,15], non-negativity [6,11] etc. In our case model results represent probability of system reaction. It means that our data are non-negative and therefore transformation associated with such properties was chosen. It is called non-negative matrix factorization. The non-negative matrix factorization aims at matrix decomposition into product of two matrices with non-negative elements. In our case we try to find such A and S where $a_{ij} \geq 0, x_{ik} \geq 0 \quad \forall i,j,k$ where $\mathbf{X} \approx \mathbf{AS}$. To obtain non-negative factors S and A we can apply the Image Space Reconstruction Algorithms (ISRA) which is derived form the squared Euclidean distance as the cost function [3,11]

$$D_F(\mathbf{X} \,\|\, \mathbf{AS}) = \frac{1}{2}\left\|\mathbf{X} - \mathbf{AS}\right\|^2 . \tag{5}$$

Minimizing the above cost function leads to following algorithms

$$\mathbf{A} \leftarrow \mathbf{A}.\times \mathbf{YX}^T./\,\mathbf{AXX}^T , \tag{6}$$

$$\mathbf{X} \leftarrow \mathbf{X}.\times \mathbf{A}^T\mathbf{Y}./\,\mathbf{A}^T\mathbf{AX} , \tag{7}$$

where .× and ./ denote component -wise multiplication and division respectively. As the starting point can be taken:

$$\mathbf{S} = \mathbf{X}, \tag{8}$$

$$\mathbf{A} = (\mathbf{I} + \mathbf{E}) + \ (\mathbf{I} + \mathbf{E})^{\mathrm{T}}, \tag{9}$$

where **I** is eye matrix and **E** is random form.

4 Component classification and generalized mixing

After basis component are estimated by e.g. NMF we need to classify them as destructive or constructive. The problem can be difficult task because obtained components might be not pure constructive or destructive due to many reasons like improper linear transformation assumption or other statistic characteristics than explored by chosen BSS method. Therefore particular component can have constructive impact on one model and destructive on the other or there may exist components destructive as a single but constructive in a group. To make the classification, for all the components' subset we check the impact after elimination of them on the final results. This procedure gives us information which component set elimination improves prediction mostly. This is equivalent with the finding of the mixing matrix $\hat{\mathbf{A}}$. It is the best matrix we can find by simple test with eliminating each combination of the components. However, the components can be not pure so their impact should have weight other than 0. It means that we can try to find the better mixing system than described by $\hat{\mathbf{A}}$. The new mixing system can be formulated more general than linear.

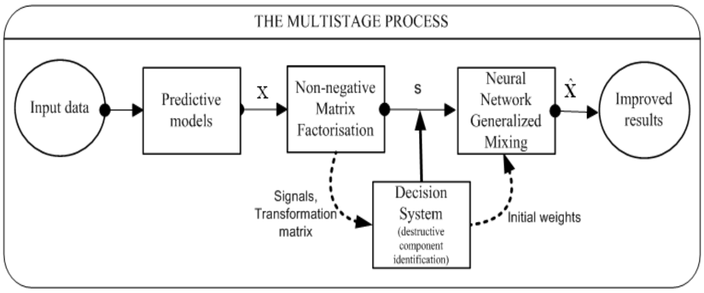
THE MULTISTAGE PROCESS
Input data
Predictive models
X
Non-negative Matrix Factorisation
s
Neural Network Generalized Mixing
X̂
Improved results
Signals, Transformation matrix
Decision System (destructive component identification)
Initial weights

Fig. 2. Scheme for model improvement.

For the general mixing system we can take MLP neural network [2,7]:

$$\hat{\mathbf{X}} = \mathbf{g}^{(2)}(\mathbf{B}^{(2)}[\mathbf{g}^{(1)}(\mathbf{B}^{(1)}\mathbf{S}+\mathbf{b}^{(1)})]+\mathbf{b}^{(2)}), \qquad (10)$$

where $\mathbf{g}^{(i)}(.)$ is a vector of nonlinearities, $\mathbf{B}^{(i)}$ is a weight matrix and $\mathbf{b}^{(i)}$ is a bias vector respectively for i-th layer, i=1,2. The first weight layer will produce results related to (4) if we take $\mathbf{B}^{(1)} = \hat{\mathbf{A}}$. But we employ also some nonlinearities and the second layer, so comparing to the linear form the mixing system gains some flexibility. If we learn the whole structure starting from system described by $\hat{\mathbf{A}}$ with initial weights of $\mathbf{B}^{(1)}(0) = \hat{\mathbf{A}}$, we can expect the results will be better, see Fig. 2.

5 Practical experiment

In practical experiment we have analysed the calls of clients with roaming services. If the unbilled amounts on their accounts grow quickly and there is a significant chance that the client will have the problems with an invoice settlement, he should be notified in advance. To estimate the non-payment-chance we need some tool anticipating, whether the customer will or will not pay. The data describe financial characteristic of the client and include 7607 observations for learning, 3802 for validating and 3802 for testing. About 22% of the customers have some payment problems.

The further models will be evaluated with the PCC measure (percentage of correctly classified). For the confusion matrix:

	Model classified as Positive	Model classified as Negative
Real True	TP	TN
Real False	FP	FN

the PCC measure will be calculated as:

$$\mathrm{PCC} = (\mathrm{TP} + \mathrm{FN}) / (\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}). \qquad (11)$$

Simple random marking 0 and 1 with probability 0.78 and 0.22 gives the PCC = 65.04%.

We have created three neural models of the MLP structure. The simplest model MLP2-7-1based on two variables: number of the outgoing calls blocks and time from last block. The hidden layer has hyperbolic activation function and the output layer – logistic. The aggregation functions are linear. The model achieved PCC = 77.96% on the testing set.

The second model is MLP7-7-1 with input like in MLP2-7-1 and also minimal invoice amount, period within the telecom, sum of all the invoices issued, sum of all the invoices paid, ratio of unbilled amount in the sum of all the clients invoices. The aggregation and activation functions are like in the model MLP2-7-1. The PCC is 78.22% on the testing set.

The third model is MLP9-10-1 with input like in the MLP7-71 model and moreover maximal invoice amount and average invoice amount. The aggregation and activation functions are like in the model MLP2-7-1. The PCC is 77.43% on the testing set.

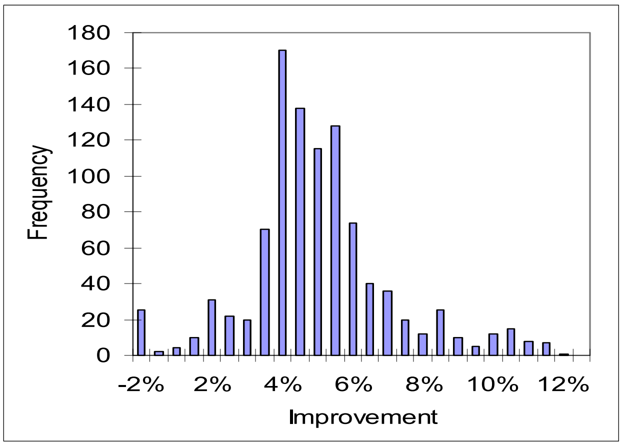

Fig. 3. Improvement rate (in %) of the PCC

Due to starting point randomization we have replicated the integration procedure 1000 times. In Fig. 3 you can observe the improvement results in the form of histogram. About 97% replications lead to the improvement of the PCC rate (improvement rate >0). The best result was about 12% and the average 5,3%. The experiment confirms that the procedure improves the model quality with statistical significance.

6 Conclusions

The adaptive dunning system for the telecom customers can includes the information integration obtained from many models of user behaviour. The ensemble method is based on decomposition procedure via non-negative matrix factorization what allows the identification of the fundamental components. The practical experiment confirms that the proposed procedure enhances the predictive power of the models and thus the validity of the approach, in general.

References:

[1] Breiman, L., Bagging predictors, *Machine Learning*, No. 24, 1996, pp. 123-140

[2] Bishop, C.M., *Neural networks for pattern recognition*, Oxford Univ. Press, Oxford UK, 1996

[3] Byrne, C., Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods. *IEEE Transactions on Image Processing*, IP-7, 1998, pp. 100-109

[4] Cardoso, J.F., High-order contrasts for independent component analysis. *Neural Computation,* No. 11, 1999, pp. 157-192

[5] Cichocki, A., Amari, S., *Adaptive Blind Signal and Image Processing*, John Wiley, Chichester, 2002

[6] Cichocki, A., Zdunek, R., Amari, S., New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* ICASSP2006. Toulouse, France, 2006

[7] Haykin, S., *Neural networks: a comprehensive foundation*, Macmillan, New York, 1994

[8] Hoeting, J., Mdigan, D., Raftery, A., Volinsky, C., *Bayesian model averaging: a tutorial,* Statistical Science 14, 1999, pp. 382-417

[9] Hyvärinen, A., Karhunen, J., Oja, E., *Independent Component Analysis*, John Wiley, 2001

[10] Jolliffe, T., *Principal Component Analysis*, Springer-Verlag, 1986

[11] Lee, D.D., Seung, H.S., Learning of the parts of objects by non-negative matrix factorization. *Nature*, No. 401, 1999, pp. 788-791

[12] Li, Y., Cichocki, A., Amari S., Sparse component analysis for blind source separation with less sensors than sources, *Fourth Int. Symp. on ICA and Blind Signal Separation*, Nara, Japan, 2003, pp. 89-94

[13] Mitchell, T., *Machine Learning*, McGraw-Hill, Boston, 1997

[14] Szupiluk, R., Wojewnik, P., Zabkowski, T., Model Improvement by the Statistical Decomposition. Artificial Intelligence and Soft Computing Proceedings, *LNCS*, Springer-Verlag, Heidelberg, 2004, pp. 1199-1204

[15] Szupiluk, R., Wojewnik, P., Zabkowski, T., Prediction Improvement via Smooth Component Analysis and Neural Network Mixing, *ICANN 2006. Lecture Notes in Computer Science*, Springer Berlin-Heidelberg, 2006, pp. 133-140

[16] Yang, Y., Adaptive regression by mixing, *Journal of American Statistical Association 96*, 2001.