FAKE PAGES DETERMINATION IN THE SOCIAL NETWORK "VKONTAKTE" WITH THE ARTIFICIAL NEURAL NETWORKS

Tolmachev R.V., Voronova L.I

Moscow technical university of communications and informatics, Moscow, Russia (111024, Moscow, Avamotornaya st., 8a), e-mail: voronova.lilia@ya.ru

This article describes the program of binary classification of pages in the social network "VKONTAKTE" for real and fake pages with machine learning algorithms based on neural networks

Keywords - fake page, binary classification, artificial neural networks

«VKontakte» is the most popular social network in Russia and some CIS countries[1]. Recently, the problem of fake pages has become acute. A fake page is an account that contains false information about its owner. Fake pages are very important to define, but it is extremely difficult to calculate them [2].

The article describes the possibility of identifying fake pages by the method of machine learning based on neural networks.

Mathematical Statement of the Classification Problem

Let **X** be the set of descriptions of objects, and **Y** the set of numbers (or names) of classes. There is an *unknown target dependence* - the map $\mathbf{y}^*$: $\mathbf{X} \to \mathbf{Y}$, , whose values are known only on the objects $\mathbf{X}^m = \{ (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots , (\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$ of the final training sample. It is required to construct an algorithm $\boldsymbol{a}$: $\mathbf{X} \to \mathbf{Y}$ that can classify an arbitrary object $\mathbf{x} \in \mathbf{X}$.

We specify the notation and the characteristic space **x** with the help of numerical metrics as follows[5]:

$\mathbf{x}=\{x_j\}$, j=1,n; where **n** is the number of characters;

$\boldsymbol{y}= \{1, 0\}$ – vector of output values (real / fake page);

$\boldsymbol{x}^{(i)}, y^{(i)}$ – vector of characteristic values and the output value of the i-th training example, *i=1, m; where m* – is the number of training examples; α – learning rate.

Let compare each page with a vector **x,** containing 15 components. The component includes the name of the metric and its type (k-categorical, n-numeric) **x** = **{**(Sex, k), (Age, n), (Marital status, k), (City, k), (Knows languages, n), (Political preferences, k), (World view, k), in life, k), (The main thing in people, k), (Attitude to smoking, k), (Attitudes towards alcohol, k), (Posts on the wall, n), (Comments on the wall, n), (Laiki on the wall , n), (Picture on the wall, n).**}**

The mechanism for collecting metric values for real pages is described below, and characteristic values for a fake pages variety was randomly generated. To solve classification problems, a neural **MLP** architecture network is used[6]**.**

Multilayer perceptron (MLP) is a neural network without feedbacks. It assumes training by methods of direct propagation of the input signal and back propagation of the error. During the forward pass, all network weights are fixed. During the backward pass, all weights are adjusted according to the error correction rule, namely for each node the error value δ is calculated for the i-th node in the l-th layer, which is also the product of the values of the layer *j* of the transposed matrix of parameters with the errors of the next layer and the derivative of the probability function of the argument of the layer j [3].

The stages of solving are described below.

Data preparation

Collection of real data. To collect user data from the social network "Vkontakte" authors wrote a script in Python, which uses a special interface VK API, which allows us to receive information from the vk.com database by using http requests to a special server[7]. The structure of the response of each request is strictly specified, the results of the program are written to a CSV file.

Fake Data Generation. To generate fake page data, a script was written in Python is written, based on the given lists: Names, Surnames, Cities, Groups and random number generators. As a result, the program received a CSV file containing data on a thousand real users and several thousand generated. The first few columns of the training set are presented in Figure 1.

Row No.	id	label	Surname	City	Name	sex=M	sex=F	relationship=Single
1	1	0	Gordeev	Moscow	Roman	1	0	1
2	2	1	Smirnov	Rostov	Ivan	1	0	0
3	3	1	Perov	Perm	Nikolai	1	0	0
4	4	1	Ivanov	Ivanovo	Igor	0	1	0
5	5	0	Sidorov	Kazan	Nikita	0	1	1
6	6	1	Sobolev	Samara	Avram	1	0	0

Fig1. Training set

Classification solution process. RapidMiner(RM) is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, predictive analytics. It is used for research, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation.

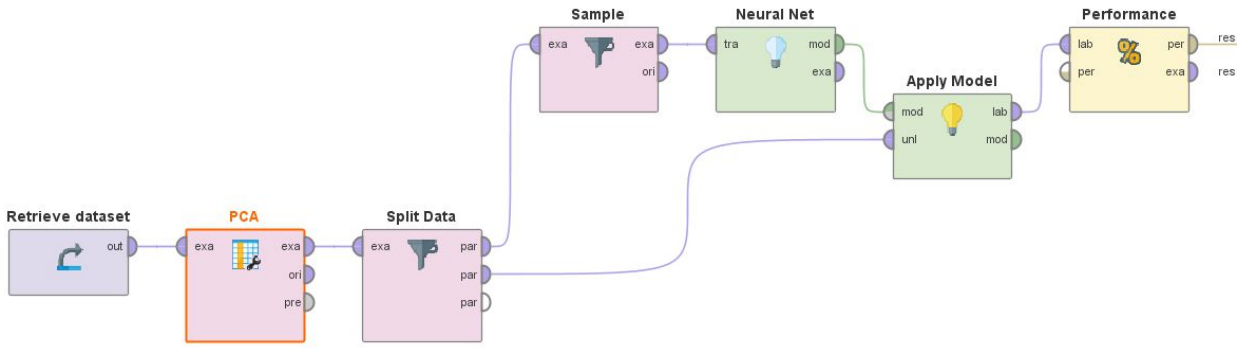

Fig. 2 Process of classification

RM helps to create solution process. A process is a collection of operators interconnected in a given order to perform the required analysis/data processing task. The operator performs actions on the input-output data ("ports"), data enters the input, data processed by the operator is output. The process for solving the page classification problem is shown in Figure 2.

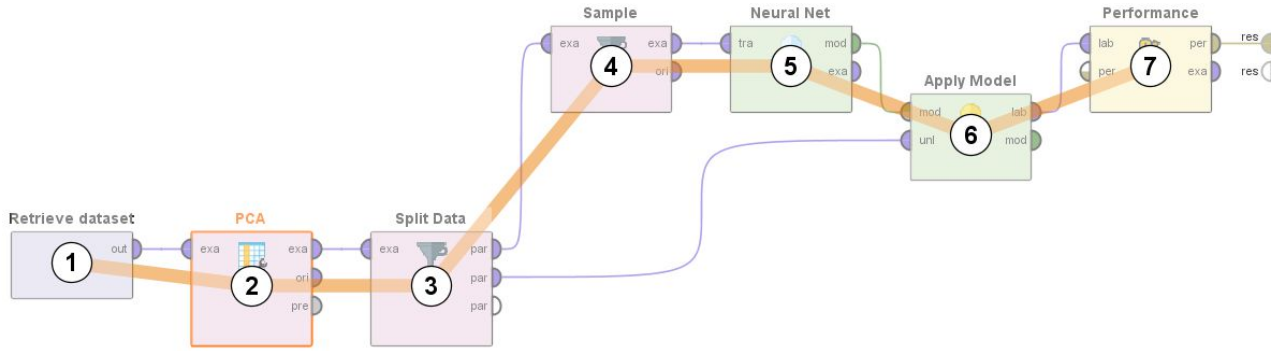

Fig.3 The sequence of actions

Figure 3 shows the sequence of actions in the RM process.

1) Loading data from the received dataset into the program memory

2) The Principal Component Analysis method is one of the most intuitively simple and often used methods for reducing the dimensionality of data and projecting them onto the orthogonal subspace of features.

3) Separation Train/Test. To train and test a neural network, the original data set is divided into a training and test sample, containing 0.7 and 0.3 data from the entire sample, respectively

4) Balancing Train. To train a neural network, approximately the same number of objects belonging to a particular class in the training sample is required.

5) Training of the neural networkю Parameters: n=15, activation function: sigmoid, $\alpha = 0.3$, Training cycles: 300

6, 7) The results. To assess the accuracy of a binary classification, it is common practice to create a confusion matrix. In a confusion matrix, classification results are compared to additional ground truth information. The strength of a confusion matrix is that it identifies the nature of the classification errors, as well as their quantities[4]. Figure 5 shows the numerical metrics that characterize the quality of the model-the error table.

	true 0	true 1	class precision
pred. 0	892	28	96.96%
pred. 1	0	880	100.00%
class recall	100.00%	96.92%	

Fig.5 Confusion matrix

References

1. http://buy-accs.ru/read/kak-sozdat-neubivaemyj-fejk-akkaunt-vkontakte

2.http://kak-delat-pravilno.ru/kak-pravilno-nazyvaetsja/kak-pravilno-nazyvaetsja-stranica-v-kontakte.html

3. http://robocraft.ru/blog/algorithm/560.html

4. http://blog.gramant.ru/2012/06/06/f1-measure/

5 Tolmachev R.V., Voronova L.I. Tematicheskaya klassifikaciya statej novostnogo resursa metodami latentno-semanticheskogo analiza// sovremennye naukoemkie texnologii. 2017. № 3. p. 55-60.

6. Voronova L.I., Tolmachev R.V. Razrabotka prilozheniya dlya kontent-analiza internet-publikacii// Stat'i dokladov 18 vserossijskaya studencheskaya nauchno-prakticheskaya konferenciya Nizhnevartovskogo gos.universiteta. 2016. p.1400-1404.

7. Tolmachev R.V., Voronova L.I.. Razrabotka prilozheniya dlya kontent-analiza internet-publikacij//Telekommunikacii i informacionnye texnologii. 2016. v. 3. № 1. p. 104-107.