sciendo

RESEARCH PAPERS FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY
IN BRATISLAVA
2019, Volume 27, Number 45 DOI 10.2478/rput-2019-0029

IMPACT OF DATA NORMALIZATION ON CLASSIFICATION MODEL ACCURACY

Dmitrii BORKIN[1], Andrea NÉMETHOVÁ[1], German MICHAĽČONOK[1], Konstantin MAIOROV[2]

[1]SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS
ULICA JÁNA BOTTU Č. 2781/25, 917 24 TRNAVA, SLOVAK REPUBLIC
e-mail: andrea.peterkova@stuba.sk, dmitrii.borkin@stuba.sk, german.michalconok@stuba.sk

[2]KALASHNIKOV IZHEVSK STATE TECHNICAL UNIVERSITY,
DEPARTMENT OF COMPUTER SOFTWARE,
4260069 IZHEVSK, UL. STUDENČESKAJA 7, RUSSIAN FEDERATION
e-mail: po@istu.ru
Received 23 August 2019, accepted 8 October 2019, published 29 November 2019

Abstract

In this paper, we present the impact of the data normalization on the classification model performance. In first part of this paper, we present the structure of our dataset, where we discuss the features of the data set and basic statistical analysis of the data. In this research, we worked with the medical data about the patients with the Parkinson disease. In second part of this paper, we present the process of data normalization and the impact of scaling data on the classification model performance. In this research, we used the XGBoost model as our classification model. The main classification task was to classify whether the patient is ill with Parkinson disease or not. Since the data set contains more numerical parameters of different scaling, the main aim of this paper was to investigate the impact of the data normalization (scaling) on the performance of the classification model.

Key words

Data normalization, model accuracy, classification

INTRODUCTION

Data analysis is more and more frequently implemented into various areas, such as automation, finance or even healthcare. The data analysis can be performed in various methods and can have different objectives and goals. The main two objectives of data analysis using machine learning methods are classification and regression. In this paper, we are dealing with

the classification task of using an XGBoost classification model. The area of interest is healthcare and in particular, the data about patients with the Parkinson disease.

Machine learning methods perform classification tasks after learning how to classify new observings. The learning process is performed on the existing known data. However, the data parameters may differ in character, and, if they are numerical, they may also be in different units and scales. Some machine learning algorithms and methods may perform worse than others on raw data. One of the most important steps in the data mining process is the data pre-processing and especially data normalization (scaling). The aim of this paper is to present the impact of data normalization on the performance of the XGBoost classification model.

DATA ANALYSIS

In our paper, we are dealing with the medical data about the patients with the Parkinson disease. This data set consists of biomedical data and is divided into two main categories. Figure 1 shows the distribution of these two main categories across the whole data set. The categories are healthy people and patients with the Parkinson disease.

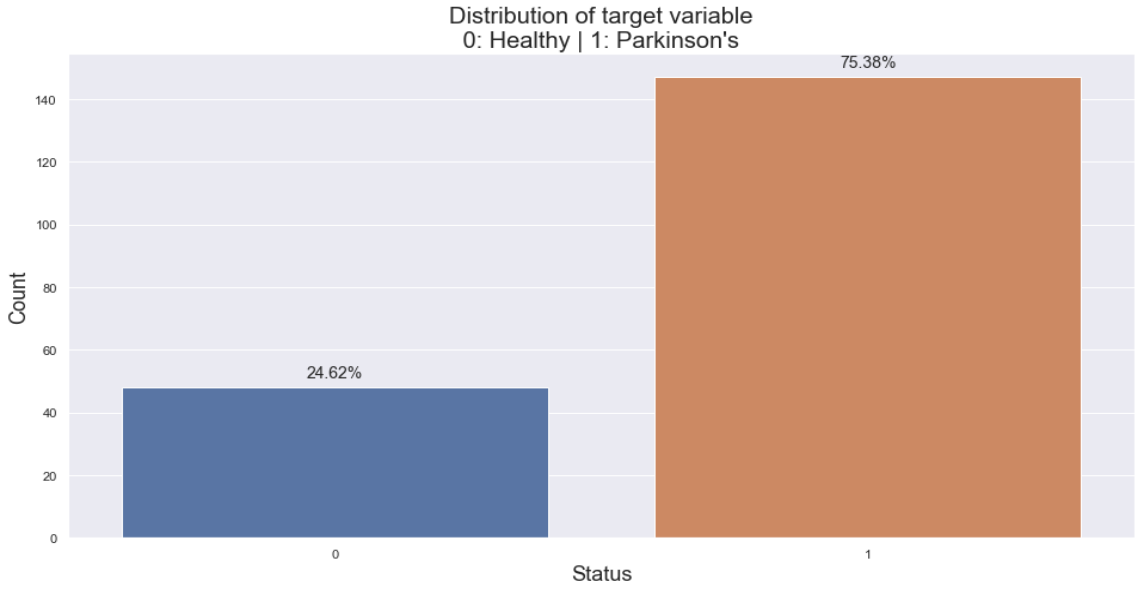

Figure 1 *The distribution of the target variable in the data set*

The whole dataset consists of 195 records and 23 biomedical parameters. The following Table shows the particular data parameters.

Table 1 Data set parameters	
PARAMETER	**DESCRIPTION**
Name	Subject name and recording number
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Several measures of variation in fundamental frequency
MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Several measures of variation in amplitude

NHR, HNR	Two measures of ratio of noise to tonal components in the voice
Status	Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE, D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1, spread2, PPE	Three nonlinear measures of fundamental frequency variation

Statistical indicators and character of the data set

Before we start applying the data normalization and classification methods, it is needed to perform classical statistical analysis of the data. For each parameter, we computed statistical indicators like mean, standard deviation, minimum value, maximum value and quantiles. Figure 2 shows the computed values for each indicator.

	mean	std	min	25%	50%	75%	max
MDVP:Fo(Hz)	154.228641	41.390065	88.333000	117.572000	148.790000	182.769000	260.105000
MDVP:Fhi(Hz)	197.104918	91.491548	102.145000	134.862500	175.829000	224.205500	592.030000
MDVP:Flo(Hz)	116.324631	43.521413	65.476000	84.291000	104.315000	140.018500	239.170000
MDVP:Jitter(%)	0.006220	0.004848	0.001680	0.003460	0.004940	0.007365	0.033160
MDVP:Jitter(Abs)	0.000044	0.000035	0.000007	0.000020	0.000030	0.000060	0.000260
MDVP:RAP	0.003306	0.002968	0.000680	0.001660	0.002500	0.003835	0.021440
MDVP:PPQ	0.003446	0.002759	0.000920	0.001860	0.002690	0.003955	0.019580
Jitter:DDP	0.009920	0.008903	0.002040	0.004985	0.007490	0.011505	0.064330
MDVP:Shimmer	0.029709	0.018857	0.009540	0.016505	0.022970	0.037885	0.119080
MDVP:Shimmer(dB)	0.282251	0.194877	0.085000	0.148500	0.221000	0.350000	1.302000
Shimmer:APQ3	0.015664	0.010153	0.004550	0.008245	0.012790	0.020265	0.056470
Shimmer:APQ5	0.017878	0.012024	0.005700	0.009580	0.013470	0.022380	0.079400
MDVP:APQ	0.024081	0.016947	0.007190	0.013080	0.018260	0.029400	0.137780
Shimmer:DDA	0.046993	0.030459	0.013640	0.024735	0.038360	0.060795	0.169420
NHR	0.024847	0.040418	0.000650	0.005925	0.011660	0.025640	0.314820
HNR	21.885974	4.425764	8.441000	19.198000	22.085000	25.075500	33.047000
status	0.753846	0.431878	0.000000	1.000000	1.000000	1.000000	1.000000
RPDE	0.498536	0.103942	0.256570	0.421306	0.495954	0.587562	0.685151
DFA	0.718099	0.055336	0.574282	0.674758	0.722254	0.761881	0.825288
spread1	-5.684397	1.090208	-7.964984	-6.450096	-5.720868	-5.046192	-2.434031
spread2	0.226510	0.083406	0.006274	0.174351	0.218885	0.279234	0.450493
D2	2.381826	0.382799	1.423287	2.099125	2.361532	2.636456	3.671155
PPE	0.206552	0.090119	0.044539	0.137451	0.194052	0.252980	0.527367

Figure 2 *Statistical indicators for each data set parameter*

Subsequently, we also computed the correlation between each parameter. The correlation matrix is shown in Figure 3.

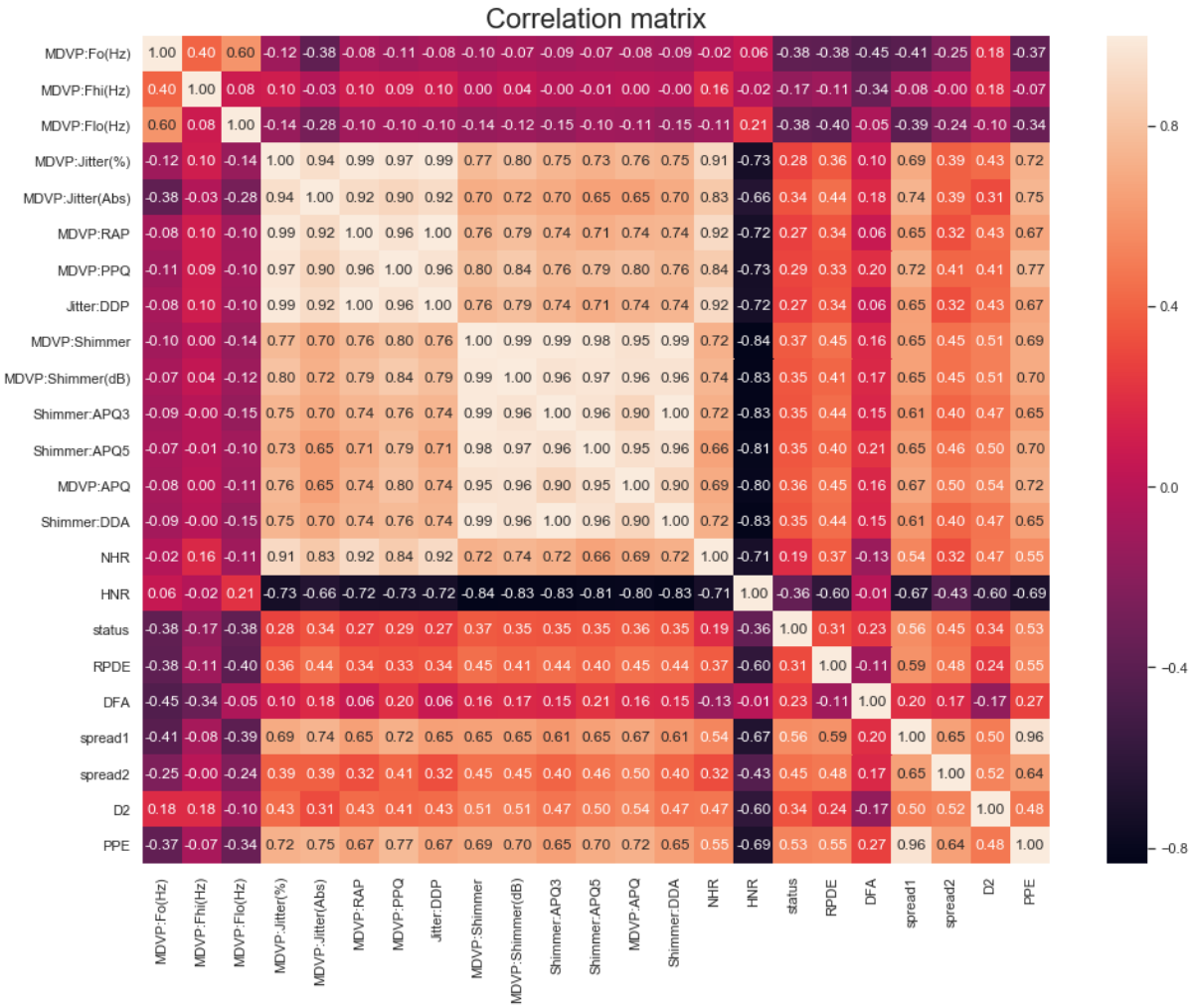

Figure 3 *Correlation matrix*

NORMALIZATION

In this part of our paper, we will present the method we used for data normalization. The range of values of raw data tends to have different scales. In such a case, in some machine learning algorithms, objective functions will not perform effectively without the data normalization. As an example, we can mention, that many classifiers and models calculate the distance between two points as the Euclidean distance. If one of the data parameters has a wide range of values, the computed distance will be governed by this particular feature. This is the reason, why the range of all features should be normalized (scaled) so that each feature will have values in same range.

Min-max normalization method

Min-max scaling, or also called min-max normalization, is known as the simplest method based on rescaling the range of values of the features to scale the range of [0, 1] or [−1, 1]. Selection of the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

where x is an original value and x‘ is the normalized value.

When we want to rescale a range between an arbitrary set of values [a, b], the formula can be described as follows:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}, \tag{2}$$

where a, b are the min and max values.

RESULTS

After performing the data normalization on our dataset, we wanted to compare the accuracy of the XGBoost classification model with normalized data, and also with the raw data set. XGBoost stands for “Extreme Gradient Boosting”. It is used for supervised learning problems, where we use the training data (with multiple features) x_i to predict a target variable y_i.

We computed the accuracy of the model with the following formula. The closer the accuracy value is to 1, the more accurate the model is.

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}. \tag{3}$$

The accuracy value for the case where we used the raw data set was equal to 0.976. On the other hand, the accuracy value for the case where we used the normalized (scaled) data was equal to 0.786.

CONCLUSION

In this paper, we compared the accuracy of an XGBoost classification model in two cases. In the first case, raw data set was used with original values, and, in the second case, normalized data was used. The data after normalization was in same range of values. The main objective of this paper was to investigate the impact of the data normalization on the classification model accuracy. As the results show, the XGBoost model performed better with the raw dataset, which confirms that the XGBoost method is not sensitive to linear transformation of the data.

However, this may be caused by a relatively small dataset or the character of the data. Since the results are not general, it can be useful to always investigate the accuracy parameter of the raw and normalized data.

Acknowledgement

This publication is the result of implementation of the Project: UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

This publication is the result of implementation of the VEGA Project 1/0673/15: Knowledge discovery for hierarchical control of technological and production pro- cesses supported by the VEGA.

This publication has been written thanks to support of the Operational Program Research and Innovation for the project: Research of advanced methods of intelligent information processing, ITMS code: NFP313010T570 co-financed by the European Regional Development Fund.

References

[1] TORLAY, L., et al. 2017. *Machine learning–XGBoost analysis of language networks to classify patients with epilepsy*. Brain informatics, 4.3: 159.
[2] ZHANG, Licheng; ZHAN, Cheng. 2017. Machine learning in rock facies classification: an application of XGBoost. In: *International Geophysical Conference, Qingdao, China, 17-20 April 2017*. Society of Exploration Geophysicists and Chinese Petroleum Society, p. 1371-1374.
[3] CHEN, Tianqi, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.
[4] WANG, Ling-Lie, YANG, Chen Ning. 1978. Classification of SU (2) gauge fields. *Physical Review D*, 17.10: 2687.
[5] FOODY, Giles, M. 2002. Status of land cover classification accuracy assessment. *Remote sensing of environment*, 80.1: 185-201.
[6] STEHMAN, Stephen, V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62.1: 77-89.
[7] LIU, Canran, FRAZIER, Paul, KUMAR, Lalit. 2007. Comparative assessment of the measures of thematic classification accuracy. *Remote sensing of environment*, 107.4: 606-616.