13. Каршиев А. А., Маматкулова У. Е., Шобутаев К. С. РЕАЛИЗАЦИЯ КВАЛИМЕТРИЧЕСКОГО ПОДХОДА В УПРАВЛЕНИИ КАЧЕСТВОМ ОБРАЗОВАНИЯ СТУДЕНТОВ СОВРЕМЕННОГО УНИВЕРСИТЕТА //Европейский журнал исследований и рефлексии в области образовательных наук. – 2019. – Т. 2019.

14. Obid o’g, Assistent Salimov Jamshid, Assistent Abror Mamaraimov Kamalidin o'g, and Assistent Normatov Nizomiddin Kamoliddin o‘g. "Numpy Library Capabilities. Vectorized Calculation In Numpy Va Type Of Information." Eurasian Research Bulletin 15 (2022): 132-137.

15. Ziyoda, Maydonova, and Normatov Nizommiddin. "RAQAMLI IQTISODIYOTDA SUN'IY INTELLEKT TEXNOLOGIYALARINI TURLI SOHALARDA AVTOMATLASHTIRISH VOSITALARI." International Journal of Contemporary Scientific and Technical Research (2023): 246-250.

16. Nizomiddin, Normatov. "TA’LIMDA DASTURLASH JARAYONINI BAHOLASHGA ASOSLANGAN AVTOMATLASHTIRILGAN TIZIMNI TADBIQ ETISH." International Journal of Contemporary Scientific and Technical Research (2023): 24-28.

17. Kamoliddin o‘g’li, Normatov Nizomiddin, and Ergashev Sirojiddin Baxtiyor o‘g‘li. "ERWIN DASTURI YORDAMIDA IDEF0, IDEF3 VA DFD STANDAT DIAGARAMMALARIDAN FOYDALANIB TIZIM SIFATIDA YARATILGAN UNIVERSITETNING MONITORING BO ‘LIMI LOYIHASI." Новости образования: исследование в XXI веке 1.6 (2023): 378-386.

DATA PREPROCESSING TECHNIQUES IN MACHINE LEARNING

Raximov Nodir Odilovich,
Khasanov Dilmurod
Tashkent University of Information Technologies

Abstract. In this paper, importance of preprocessing and techniques in this field such as data cleaning, dimensionality reduction, smoothing, normalization are illustrated. During the research we mentioned some details of techniques above. However, our research includes only theoretical aspect of data preprocessing. The data preprocessing phase while arduous and time-intensive stands as the cornerstone of data science, possessing paramount significance. Neglecting the meticulous cleansing and structuring of data has the potential to undermine the integrity and efficacy of subsequent modeling endeavors.

Keywords: data preprocessing, data cleaning, normalization, exploratory data analysis, dimensionality reduction.

Introduction

When confronted with real-world data, Data Scientists invariably find it necessary to employ preprocessing techniques to enhance data usability. Such techniques serve the dual purpose of rendering the data more amenable for utilization within machine learning (ML) algorithms and mitigating complexity to forestall overfitting, ultimately yielding a superior model. Upon comprehending the nuances of your dataset and identifying primary data intricacies through Exploratory Data Analysis (EDA), the subsequent stage entails data preprocessing, which involves preparing the dataset for its application in a model. Ideally, one would hope for a dataset devoid of imperfections. Nevertheless, real-world data is inherently prone to various issues necessitating remediation. For instance, within an organizational context, inconsistencies such as typographical errors, missing data, disparate scales, and other anomalies frequently manifest. These real-world challenges must be rectified to enhance data utility and comprehensibility. This pivotal phase, where data is cleansed, and most issues are resolved, is aptly termed "data preprocessing"[2].

Preprocessing methods

Neglecting the data preprocessing step carries significant repercussions for subsequent machine learning model utilization. Many models are incapable of accommodating missing values, and various adverse data characteristics, including outliers, high dimensionality, and noise, can adversely impact model performance. Thus, by undertaking data preprocessing, the dataset is enhanced in terms of completeness and accuracy. This pivotal phase is indispensable for effecting essential data adjustments prior to supplying the dataset to the machine learning model, thereby ensuring the model's effectiveness and reliability[1,3].

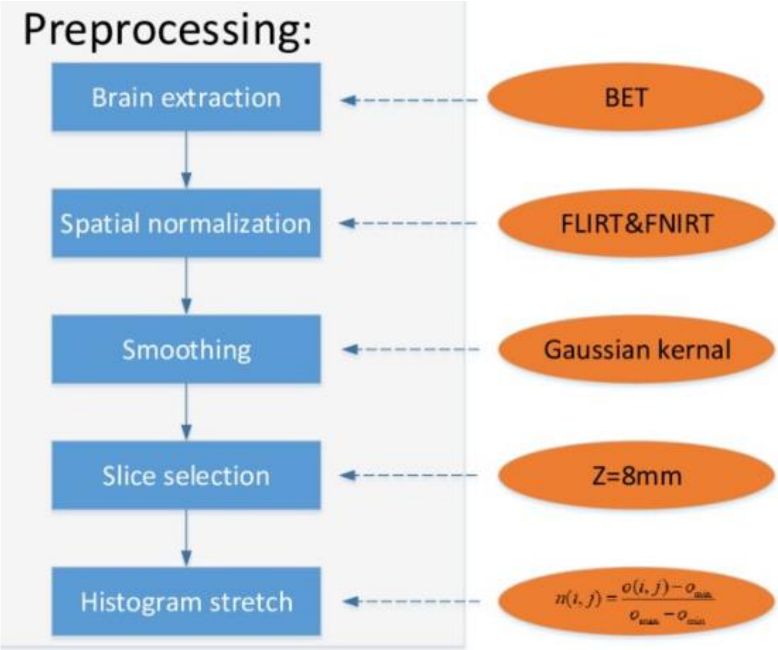

Fig 1. Typical architecture of preprocessing

Dimensionality reduction – is primarily concerned with the reduction of input features within training data. In the realm of real-world datasets, an abundance of attributes is typically encountered. Failing to curtail this multitude of features can potentially compromise the performance of a model when it is later applied to this dataset. Effectively reducing the number of features while preserving a significant portion of the dataset's variability yields several advantageous outcomes, including:

Conservation of computational resources: diminished feature space necessitates fewer computational resources during modeling.

Enhanced model performance: a streamlined feature set often contributes to improved model performance.

Mitigation of overfitting: Dimensionality reduction assists in preventing overfitting, wherein a model becomes excessively complex and memorizes training data, leading to a substantial drop in performance on test data.

Alleviation of multicollinearity: this technique helps alleviate multicollinearity, a condition characterized by high correlations among one or more independent variables. Moreover, the application of dimensionality reduction techniques contributes to the reduction of noise within the dataset. Now, let us delve into the primary approaches to dimensionality reduction that can be employed to enhance data suitability for subsequent analysis[4,6].

Feature selection pertains to the process of identifying and retaining the most salient variables (features) associated with the prediction variable. In essence, it involves selecting attributes that exert the greatest influence on the model.

Normalization is a data preparation method employed to standardize the values of attributes within a dataset, thereby enhancing the effectiveness and precision of machine learning models. The primary objective of normalization is to mitigate any potential biases and discrepancies stemming from variations in feature scales. Various forms of data normalization are available. Suppose you possess a dataset denoted as X, containing N rows (entries) and D columns (features). In this context, X[:,i] signifies feature i, while X[j,:] signifies entry j.

- Z normalization(Standardization);
- Min-Max normalization;
- Unit vector normalization.

For instance, Z normalization method Gaussian function is used to change the value of elements in dataset.

$$\hat{X}[:,i] = \frac{X[:,i] - \mu_i}{\sigma_i},\ \ \mu_i = \frac{1}{N}\sum_{k=1}^{N} X[k,i],\ \sigma_i = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(X[k,i]-\mu_i)^2} \quad (1)$$

In (1.1) standardization does not change the type of distribution:

$$\hat{X} = aX + b \rightarrow f_{\hat{X}}(x) = \frac{1}{|a|} f(\frac{x-b}{a}) \qquad (2)$$

This transformation sets the mean of data to 0 and the standard deviation to 1. In most cases, standardization is used feature-wise Min-Max normalization[9]:

$$\hat{X}[:,i] = \frac{X[:,i] - \min(X[:,i]}{\max(X[:,i]) - \min(X[:,i])} \qquad (3)$$

Z normalization result :

$$\begin{bmatrix} 13 & 16 & 19 \\ 22 & 23 & 38 \\ 47 & 56 & 70 \end{bmatrix} \Longrightarrow \begin{bmatrix} 0.013 & 0.051 & 0.1 \\ 0.155 & 0.172 & 0.43 \\ 0.586 & 0.741 & 0.92 \end{bmatrix}$$

Data cleaning - a paramount facet of the data preprocessing phase pertains to the identification and rectification of flawed and imprecise observations within the dataset, thereby augmenting its overall quality. This methodology entails the discernment of inadequacies, inaccuracies, redundancies, inconsequentialities, or void entries in the dataset. Following the identification of these anomalies, it becomes imperative to effectuate corrective measures, which may encompass data modification or exclusion. The strategic approach employed in this context is contingent upon the problem domain under consideration and the overarching objectives of the project. We will now delineate some of the prevalent data analysis challenges and expound upon the methods for their resolution. Noisy data typically encompasses inconsequential or erroneous entries within a dataset, which may manifest as meaningless data points, incorrect records, or duplicated observations. For instance, consider a scenario where a database column labeled 'age' contains negative values, rendering the observation nonsensical. Another circumstance pertains to the removal of extraneous or irrelevant data. Especially, in regression problems we have to work on real numbers, in this case normalization is one of the most important techniques[11].

Conclusion

Through the research above data preprocessing techniques help to improve accuracy of the prediction result. For example, Dimensionality reduction makes easier computational process and reduces execution time for prediction or analysis model. Moreover, other techniques such as normalization makes understandable dataset, so dataset values are transformed the mean of data to 0 and the standard deviation to 1. Some theoretical explanations are given according to preprocessing techniques' details in this article. The next researches, we will try to touch mathematical aspect of these processes.

References:

1. Jiang S., Li J., Zhang S., Gu Q., Lu C., Liu H. Landslide risk prediction by using GBRT algorithm: Application of artificial intelligence in disaster prevention of energy mining. Process. Saf. Environ. Prot. 2022, 166, 384–392.

2. N.Rahimov, D.Khasanov,“The application of multiple linear regression algorithm and python for crop yield prediction in agriculture”, Harvard educational and scientific review, Vol.2. Issue 1 Pg. 181-187.

3. N.Raximov, J.Kuvandikov, D.Khasanov, “The importance of loss function in artificial intelligence”, International Conference on Information Science and Communications Technologies (ICISCT 20222), DOI: 10.1109/ICISCT55600.2022.10146883

4. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, NY: Springer; 2009.

5. Oliver Theobald. Machine Learning for Absolute Beginners**.** – Scatterplot Press. 2017. pg.43-98

6. M.Tojiyev,O.Primqulov,D.Xasanov, “Image segmentation in OpenCV and Python, DOI:10.5958/2249-7137.2020.01735.8

7. N.Raximov, O.Primqulov, B.Daminova,“Basic concepts and stages of research development on artificial intelligence”, International Conference on Information Science and Communications Technologies (ICISCT), www.ieeexplore.ieee.org/document/9670085/metrics#metrics

8. N.Rahimov, D.Khasanov, “The mathematical essence of logistic regression for machine learning”, International Journal of Contemporary Scientific and Technical Research. Pg. 102-105.

9. Khasanov Dilmurod, Tojiyev Ma’ruf,Primqulov Oybek., “Gradient Descent In Machine”. International Conference on Information Science and Communications Technologies (ICISCT), https://ieeexplore.ieee.org/document/9670169

10. Babomurodov O. Zh., Rakhimov N. O. Stages of knowledge extraction from electronic information resources. Eurasian Union of Scientists. International Popular Science Bulletin. Issue. № 10(19)/2015. – pp. 130-133. ISSN: 2411-6467

11. N Raximov, M Doshchanova, O Primqulov, J Quvondikov. Development of architecture of intellectual information system supporting decision-making for health of sportsmen.// 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)Prasun Biswas, “Loss function in deep learning and python implementation”(web article), www.towardsdatascience.com, 2021.

12. Rahimov Nodir, Khasanov Dilmurod. (2022). The Mathematical Essence Of Logistic Regression For Machine Learning. https://doi.org/10.5281/zenodo.7239169

13. T. Maruf, "Hazard recognition system based on violation of the integrity of the field and changes in the intensity of illumination on the video image," 2022 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2022, pp. 1-3, doi: 10.1109/ICISCT55600.2022.10146933

14. Ma’ruf Tojiyev, Ravshan Shirinboyev, Jahongirjon Bobolov. Image Segmentation By Otsu Method. International Journal of Contemporary Scientific and Technical Research, (Special Issue), 2023. 64–72, https://zenodo.org/record/7630893

MULOHAZALAR VA MATRITSALARNING O‘ZORO BOG‘LANISHI

Maniyozov Oybek Azatboyevich
Toshkent axborot texnologiyalari univeristeti Farg‘ona filiali
maniyozovo@gmail.com

Annotatsiya: Ushbu maqolada mulohalar va matritsalar bog‘lanishi, mulohazalarni shifrlash, shirflangan ma’lumotlarni yechish, matritsalar ko’paytmasi hamda teskari matritsalarni Matlab programmasida ishlatish haqida ma’lumot berilgan.

Kalit so‘zlar: Mulohaza, sodda mulohaza, shifr matritsa, teskari matritsa, matritsalar ko‘paytmasi.