Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **1079:1084***

Data Pre-processing Issues in Medical Data Classification

Ashwini Tuppad[1], Shantala Devi Patil[2]

[1,2]School of Computer Science and Engineering, REVA University.
Email: tuppadashwini@gmail.com[1] , shantaladevipatil@reva.edu.in[2]

**Corresponding author's E-mail: tuppadashwini@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

With digitalization of data and the rise of World Wide Web, access to information has been very easy and affordable. Especially the Web and the Internet have boosted research activities by facilitating access to large, publicly available medical datasets under open access scheme. These developments have resulted in explosive amounts of data being generated varying in volume, variety and velocity thus referred to as big data. Availability of such medical big data has catalyzed the research in medical predictive analytics. However, the true value of such data can be derived only after subjecting it to careful processing and analysis before drawing inferences from it. Publicly available medical datasets have noise in the form of missing values, outliers and data inconsistencies, that may affect the results or outcomes negatively. Pre-processing of such data is essential to eliminate noisy elements and refine the data to be suitable for further analysis and processing. This paper signifies the need for data pre-processing and explains the data pre-processing pipeline with various underlying stages constituting it. It also presents a comparative analysis of various data pre-processing techniques for handling missing values and outliers in a dataset..

Keywords: *Data, Pre-processing, Missing data, Imputation, Outlier, Sampling*

1. Introduction

Recently, medical predictive analytics has gained a lot of momentum in research community. With technologies like machine learning, data mining, big data and cloud computing, analysis of medical or clinical data for decision making and improvement of healthcare services has been possible [1,2]. The scope of the field is very wide covering clinical decision support, hospital administration management, quality of care improvement, pharmacological innovations, public health monitoring and so on. In part, such a growth is attributable to easy availability of public medical datasets nowadays. However, owing to huge volume and variety of medical data, the task of processing and analysis of such data to capture useful insights and information itself is tedious and difficult [2,3].

Medical data commonly contain missing values and outliers [4-7]. Missing values arise when required medical data are either unavailable, unrecorded or are lost. Such values often affect the predictive algorithm accuracy negatively [8]. Handling missing values is thus an important pre-processing step. Outliers are data values that are abnormal or extreme, lying outside the defined range. They are considered as noisy data requiring elimination [7]. Outlier detection and elimination is another significant data pre-processing step. Apart from these, a notable and common issue in medical datasets is class imbalance problem where the majority of medical records in the dataset belong to one particular class leaving very few records in others [7]. In such cases, the accuracy of the predictive algorithm is compromised and the result is said to be pseudo-accurate. These issues can be handled by data pre-processing, wherein the data is subjected to noise elimination and refinement to obtain better prediction results [15,16]. Primarily data pre-processing comprises of four steps- data cleaning, data integration, data transformation and data reduction [4,5,9,11]. Fig. 1 shows the data pre-processing pipeline. It is not mandatory to have all stages of pre-processing pipeline as it depends on the dataset and purpose of research.

1. Data cleaning refers to elimination of noise from the data including missing values, outliers or inconsistent data values. Techniques for handling missing values include complete elimination,

mean substitution, probabilistic parameter estimation, kNN-imputation and so on [1,5]. Techniques for removal of outliers or inconsistent data include binning, regression, clustering etc. [4].

2. In data integration (or Data consolidation), relevant data from multiple sources are combined resulting in heterogeneity and redundant values. It aids in achieving better accuracy with more knowledge gathered from many sources. Two significant challenges resulting from data integration are processing of heterogeneous and redundant data. Correlation analysis can be used to discover redundancy and handle such values appropriately [4].

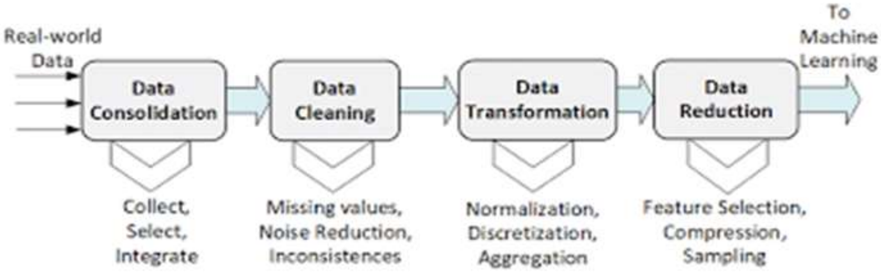

Real-world Data
Data Consolidation
Data Cleaning
Data Transformation
Data Reduction
To Machine Learning
Collect, Select, Integrate
Missing values, Noise Reduction, Inconsistences
Normalization, Discretization, Aggregation
Feature Selection, Compression, Sampling

Fig 1: Data Pre-processing Pipeline

3. Data reduction, the next step in pre-processing pipeline aims to reduce the size of the data without compromising on the quality. Reducing data to compact form helps in managing the space and time complexity in real time when datasets are of huge scale. Some notable data reduction techniques include dimensionality reduction and data compression [1,4,9,11].

4. The final step of data transformation involves converting the reduced data to a form that makes further analysis easy and straightforward. Common data transformation techniques include smoothing, normalization [10], aggregation etc. [4]. Usually, not all pre-processing steps are required for a particular dataset. The selection of pre-processing steps as well as techniques depend on the nature of dataset and its characteristics.

Handling Missing Values

Missing values occur in most of the datasets and if left untreated, have the ability to impact the predictive accuracy negatively [17]. Often their occurrence is attributed to either of the following cases- (a) data value unknown (b) data value known but unrecorded (c) data value lost [5]. If the dataset is small with lots of missing values, the results are adversely affected, biased and misinterpreted [2]. Popular techniques for handling missing values include-

a) Complete elimination of missing value records - Though it is not best strategy to completely eliminate records containing missing values from the dataset, it is still adopted when the dataset is very large with few missing values in some records. In such case, it provides easy and fastest solution for the problem [1].

b) Substitution of a constant- Missing values may also be replaced by a common constant value across all records of the dataset, irrespective of feature or the class [3,4].

c) Data Imputation- This technique refers to substitution of meaningful, approximate estimates for the missing values found in the dataset. Following is the explanation of the most common and preferred data imputation techniques [5]-

i. *Mean substitution-* In this technique, the missing values of a particular feature are substituted by the mean value(average) of the non-missing values of the same feature for a given class. The mathematical formula for finding the mean value of ‘n’ non-missing values of jth feature of class ‘k’ is-

$$\overline{x_{ij}} = \sum_{i:x_{ij}\in C_k} \frac{x_{ij}}{n_k}$$

(1)

ii. *Median substitution-* This technique involves imputing missing values with median i.e., the middle or central value of the dataset. It is more preferred over mean substitution when the dataset has skewness(distortion) and outliers are present, since the median provides a much better and correct representation of data than mean. The mathematical formula for median imputation value is-

$$\widehat{x_{ij}} = \boldsymbol{median}_{i:x_{ij}\in c_k} x_{ij} \qquad \textbf{(2)}$$

iii. *Mode imputation*- Here, the mode (most frequently occurring value) is used as a replacement for missing data. It is more suited when the data values to be imputed are categorical in nature.

Mode = Most frequently occurring value of the feature

d) kNN imputation- k-Nearest Neighbors (kNN) is instance-based machine learning algorithm, which has been popularly used for data imputation that preserves the prediction accuracy well. It finds the replacement for missing value in a record by choosing k nearest neighbors of that record using a distance metric. The nearest neighbors are chosen such that they have similar or proximal values for rest of the features in the missing value record. In case of numerical feature, the missing value is imputed with the average value of the same feature summed over nearest neighbor records. When it is a categorical feature, its missing value is replaced by the frequently occurring value among the k nearest neighbor records chosen. kNN imputation is highly successful but with larger datasets, the time complexity goes on increasing owing to the task of finding k nearest neighbors.

Other data imputation techniques reported in the literature are Multivariate Imputation by Chained Equations (MICE) based on logistic regression and predictive mean matching [10] for predicting replacement values for missing values. However, this technique works only when the missing values have the property of missing at random. Support Vector Regression Imputation [5], Bayesian Inference-based Imputation [1], Decision Tree Imputation [8], Miss Forest Imputation [12] method are also some of the imputation techniques adopted in previous research works.

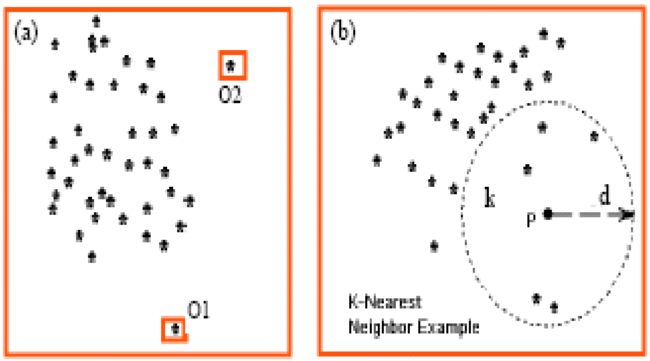
Outlier Elimination

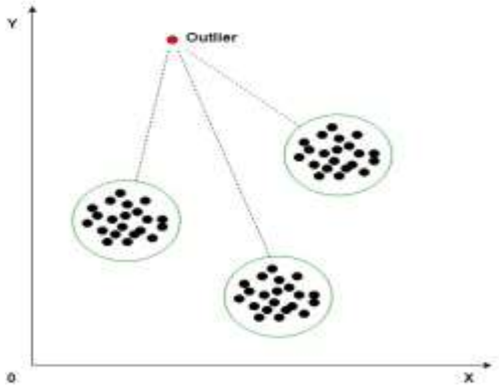
Outliers are considered to be noisy data which are basically extreme or abnormal data values located outside of the normal predefined range of features. They have the potential to hamper the prediction results if present in abundance and not managed properly [6,7]. A classification of outliers is provided in [13]-

a) Point outliers- Singular values of features pertaining to multidimensional data types that act as outliers.

b) Contextual outliers- These class of outliers depend on the context such time series, graphical or discrete data types They are not random abnormal or out-of-range values rather depend on contextual attribute values on which they are dependent.

c) Collective outliers- Here, instead of one instance acting as an outlier, a collection of data or instances together introduce abnormality.





(a) (b)

Fig. 2. Outlier detection – (a) Distance based outlier detection (b) Cluster based outlier detection

Popular approaches for outlier handling adopted by previous research works are explained below-

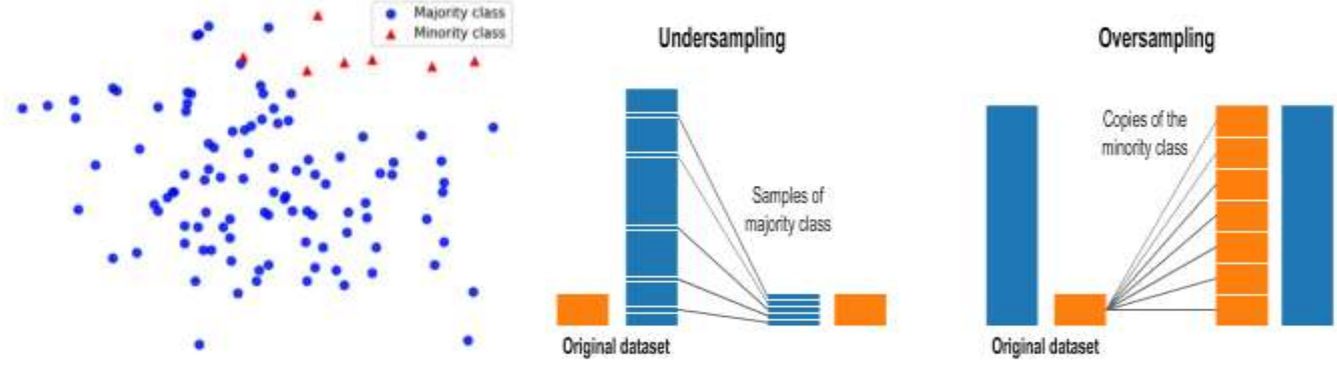
1) Feature Selection- This technique corresponds to selection of small set of highly representative features of a given dataset that eliminates less useful features. The selected subset of features are highly discriminative and are able to classify the records well. Dimensionality reduction is one means of feature selection. F score is also another strategy that computes the discriminative power of features [6]. They can be further categorized into filter or wrapper sub-techniques. Filter methods make use of the intrinsic feature properties and statistics to select relevant optimal feature subset, while wrapper methods involve use of Machine Learning classifier driven feature selection [1].

2) Instance Selection- This refers to selection of a subset of instances or samples from the dataset that better represent the different classes and maximize the chances of classification while minimizing the bias or abnormality. Also known as sampling, such techniques are based on statistics and select statistically good sample expected to likely increase the classifier performance [6].

3) Distance based outlier elimination- This technique uses distance metrics for computing similarity between features or attributes of the dataset. For any two valid features(non-outliers), a threshold value is predefined to denote its similarity range. The outlier detection thus involves scanning each of the features to check their similarity using distance metric like Euclidean distance. In case the Euclidean distance is greater than the threshold value, the feature is recognized as an outlier and is eliminated [Fig. 2(a)]. Example of distance-based outlier elimination technique is Tomek Links. They are basically a pair of points that represent neighbouring instances such that one belongs to majority class while another is minority class instance. Such two points a and b are said to form Tomek link if there does not exist any point c with d(a,c)<d(a,b) or d(b,c)<d(a,b). The instance of the majority class is removed as an outlier [10].

4) Cluster based outlier elimination- Another popular method of outlier handling is through clustering. Clustering involves grouping similar data in particular clusters which define the class to which data belong. In a classification problem, the records or instances in a dataset belong to one particular class and there may be two or more classes present. Based on the number of classes, clustering involves identifying similar data located around a central point or centroid. The same strategy is used for identifying outliers when outlier records does not fit into any of the clusters identified by their centroids [Fig. 2(b)]. The distance between outlier and centroid of any cluster considered is too large to define it as valid record, following which such outliers are eliminated. K-means clustering is most common cluster-based outlier handling technique [10].

Handling Class Imbalance Class imbalance is real problem apart from missing values and outliers. In huge datasets, often one of the class dominates in terms of number of examples it contains than the other one.



(a) (b)

Fig. 3. (a) Class imbalance problem showing majority and minority class instances (b) Under-sampling and Over-sampling techniques When there is high rate of class imbalance, classification results are not reliable. Statistical Sampling techniques are used to balance the dataset. Following are the most common sampling techniques used for data imbalance addressal [14] -

1) Random Under-Sampling- This technique involves resampling a training dataset by randomly deleting instances from majority class to balance the number of instances in minority class. Hence, the name Under-Sampling. However, it is vulnerable to loss of important information if applied many times.
2) Random Over-Sampling- This technique is also category of random sampling however, here randomly instances are selected from the minority class and duplicated to fit in majority class. The drawback is overfitting of data to the model under certain cases.
3) SMOTE (Synthetic Minority Oversampling Technique)- It is sampling technique which focuses on balancing minority class instances with those in majority class. As the name suggests, synthetic samples of the minority class are created such that the number of minority and majority classes are near equal. Hence, samples are new additions than redundant ones. the SMOTE approach is applied on over-sampled minority classes by creating synthetic examples instead of replicating or replacing already existing examples. The majority class samples remain unchanged, thus avoiding the overfitting problem.
4) Distribution-based Over-Sampling- This involves oversampling strategy for data balancing however, here synthetic examples are produced following the distribution of original dataset.

Hybrid Sampling techniques- These involve combination of under-sampling as well as over-sampling to balance the dataset while avoiding the problem of overfitting or data loss.

4. Conclusion

Data pre-processing is significant step in any classification problem. It is more required when the dataset has been collected from publicly available datasets curated from multiple data sources. This paper has presented overview of data pre-processing pipeline and its various stages along with their use. The most common and challenging problems of missing data, outliers and class imbalance found in public datasets has been discussed with various possible solutions to resolve them. The applicability of each of the pre-processing stages and techniques therein depend on the dataset used and the research objectives, though ultimate aim of remains noise elimination and data preparation for further analysis.

References:

Puneet Misra and Arun Singh Yadav, “Impact of Preprocessing Methods on Healthcare Predictions” 2nd INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND SOFTWARE ENGINEERING, 2019.

[2] E. T. Capariño, A. M. Sison and R. P. Medina, "A Modified Imputation Method to Missing Data as a Preprocessing Technique," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1-6, 2018.

[3] Arpitha, Manda & Ramalakshmi, K. & Ramachandran, Venkatesan, “Enhancement of accuracy on a medical dataset by the usage of different data preprocessing techniques. International Journal of Innovative Technology and Exploring Engineering”, 8, pp. 1-4, 2019.

[4] Bhaya, Wesam. (2017). Review of Data Preprocessing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. 10.3923/jeasci.2017.4102.4107.

[5] J. Sessa and D. Syed, "Techniques to deal with missing data", 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), pp. 1-4, 2016.

[6] Min-Wei Huang, Wei-Chao Lin, Chih-Wen Chen, Shih-Wen Ke, Chih-Fong Tsai and William Eberle, “Data Preprocessing Issues for Incomplete Medical Datasets”, Expert Sys: J. Knowl, Eng. 33, 5, pp. 432-438, 2016.

[7] Nonso Nnamoko, Ioannis Korkontzelos, “Efficient treatment of outliers and class imbalance for diabetes prediction”, Artificial Intelligence in Medicine, Volume 104, 2020.

[8] Y. Pristyanto and I. Pratama, "Missing Values Estimation on Multivariate Dataset : Comparison of Three Type Methods Approach," 2019 International Conference on Information and Communications Technology (ICOIACT), pp. 342-347, 2019.

[9] S. N. Singh and K. Kathuria, "Diabetes Diagnosis using different Data Pre-processing Techniques," 2018 4th International Conference on Computing Communication and Automation (ICCCA), pp. 1-4, 2018.

[10] E. Zeinulla, K. Bekbayeva and A. Yazici, "Effective diagnosis of heart disease imposed by incomplete data based on fuzzy random forest," 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-9, 2020.

[11] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, Guoyan Wang, “Data Processing and Text Mining Technologies on Electronic Medical Records: A Review”, Journal of Healthcare Engineering, 2018.

[12] Vinutha N, Pattar S, Sharma S, Shenoy PD, Venugopal KR, “A Machine Learning Framework for Assessment of Cognitive and Functional Impairments in Alzheimer's Disease: Data Preprocessing and Analysis”, J Prev Alzheimers Dis.7(2), pp. 87-94, 2020.

[13] A. Christy, G. Meera Gandhi, S. Vaithyasubramanian, “Cluster Based Outlier Detection Algorithm for Healthcare Data”, Procedia Computer Science, Volume 50, pp. 209-215, 2015.

[14] H. Benhar, A. Idri, J.L. Fernández-Alemán,, “Data preprocessing for heart disease classification: A systematic literature review”, Computer Methods and Programs in Biomedicine, Volume 195, 2020.

[15] Sarab AlMuhaideb, Mohamed El Bachir Menai, “An Individualized Preprocessing for Medical Data Classification”, Procedia Computer Science, Volume 82, pp. 35-42, 2016.

[16] Razavi A.R., Gill H., Åhlfeldt H., Shahsavar N. (2005) A Data Pre-processing Method to Increase Efficiency and Accuracy in Data Mining In: Miksch S., Hunter J., Keravnou E.T. (eds) Artificial Intelligence in Medicine. AIME 2005. Lecture Notes in Computer Science, vol 3581. Springer, Berlin, Heidelberg.

[17] T. Jayalskshmi and A. Santhakumaran, "Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks," 2010 Second International Conference on Machine Learning and Computing, pp. 109-112, 2010.