Generalized Denoising Coder with Multi-modal Approach for Handling Incomplete Information in Public Health Big Data Analytics

Kamlesh Kumar Yadav[1], Dhablia Dharmesh Kirit[2]

[1]*Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India.*

[2]*Research Scholar, Department of CS & IT, Kalinga University, Raipur, India*

KEYWORDS

Generalized Denoising Coder, Multi-modal, Public Health Big Data Records, Deep Learning, Incomplete Information

ABSTRACT

A data-driven health model incorporates many missing and incomplete values, and the effectiveness of a health model that relies on characteristics not captured by the consumer is low. A Deep Learning (DL) health paradigm aims to improve the precision of learning weights. A DL-based health model is trained by optimizing the weights to maximize accuracy. If the learning process is excluded, the accuracy of applying a DL-based health model to the user's circumstance may be reduced, and it becomes the responsibility of people to manage Public Health Big Data Records (PHBDR). To handle incomplete information, this study suggested a generalized denoising coder with a multi-modal approach in public health big data records (GDC-MM-PHBDR). The suggested technique uses a Generalized Denoising Coder (GDC), which estimates the output values following the relevant input values using Neural Networks (NN). The National Health and Nutrition Examination Study (NHNES) provided data for this investigation. This approach allows for the estimation of incomplete information in the PHBDR. Multimodality, which the PHBDRs provide, enables compiling data for a single item from many sources. The layout of the GDC is multi-modal. The GDC-MM-PHBDR approach continuously demonstrates excellent performance, attaining maximum accuracy across all noise levels. It begins at 98.31% accuracy with a noise factor of 0.05 and significantly decreases to 94.5% with a noise factor of 0.25.

1. Introduction

Healthcare data has been generated in several forms throughout the age of big data. Personalized wellness tracking devices, such as those, allow for the gathering of data that is customized explicitly to a person's everyday activities. Similarly, the fast advancement of laboratory procedures results in the production of large quantities of sequencing data. Nevertheless, these emerging forms of information are more prone to incomplete values than conventional tabulated medical information obtained from an upcoming observational or randomized study [1].

Bad things can happen when values are missing from a big data study. Predictions can be less accurate and cause bias to creep into the decision-making process [2]. This is especially annoying when important decisions need to be made in the healthcare field. Several useful ways to deal with missing values in tabular static data include estimation using mean, median, or mode. These methods might not work well with certain data types, like time-based, tabular, image, or genetic data. To ensure the model's quality and longevity, it is important to use modern methods. Nielson [3] says that incomplete data can be put into three groups: incomplete at random (IAR), missing at random (MAR), and incomplete not at random (INAR).

DL-based methods are being used increasingly to solve lost value issues these days. These methods have been shown to improve estimation accuracy [10]. DL-based systems can also deal with complex missing trends and information structures, like time data that needs to be arranged in a certain way and picture data that needs to be organized based on geography [5]. Because they are so effective and easy to change, DL-based interpolation models are becoming increasingly popular in many situations, from predicting in-patient mortality to finding Alzheimer's disease (AD) early on. Many reviews only look at methods that aren't deep learning or only look at neural networks as a single type of method. The details in these studies aren't clear enough to help researchers figure out if DL-based models will work with their data.

When data is combined in public health systems and utilized for training or assessment, issues such as extracting relevant characteristics and lacking certain variables become apparent [4]. Public health data representing various characteristics is organized as individual modalities to address this issue [8]. Including different variables in each model leads to variances in learning effectiveness or modeling outcomes [7]. The single-modal model may be transformed into a multi-modal model to address each modality's limitations based on the data features. Various single-modal models are necessary depending on the user's circumstances. In contrast, a comprehensive multi-modal model that encompasses them faces the challenge of adequately catering to the diverse needs of users.

As far as we know, there has not been a comprehensive evaluation of DL-based techniques for filling in missing values in various forms of healthcare big data [6]. To fill this void, an evidence mapping assessment [12] has been proposed that investigates the use of models depending on the kind of data [11]. The objective is to assist researchers in various clinical fields in handling missing values using DL-based approaches. Furthermore, using an NN for estimate enables the model to acquire the characteristics from the data autonomously, hence reducing the need for human involvement. Therefore, this work proposes a method for predicting incomplete information, particularly in PHBDR, using a multi-modal GDC in healthcare big data.

Generalized Denoising Coder with Multi-modal Approach for Handling Incomplete Information in Public Health Big Data Records (GDC-MM-PHBDR)

The health statistical information includes an individual's behaviors, genetic predisposition, illnesses, and medical records. NHNES refers to collecting information obtained through several means of data collection. Data is gathered through public health surveys, health examinations, and nutritional assessments following each approach's specific data collection process. Incomplete information is encountered in several circumstances throughout the implementation process. The outcomes of big data analysis in public health vary depending on the pre-processing methods used when there are incomplete details in several transactions. This necessitates a suitable pre-processing method that mitigates the impact of insufficient data on the outcomes. This work proposes a system for approximating incomplete data in large-scale healthcare using a GDC. The NHNES dataset comprises several variables, and the classification approach allows for various combinations of these variables in a multi-modal format. Enhanced data analysis and DL results may be anticipated by effectively handling incomplete information. The parameters used to calculate such incomplete information for PHBDR are shown in Fig. 1.

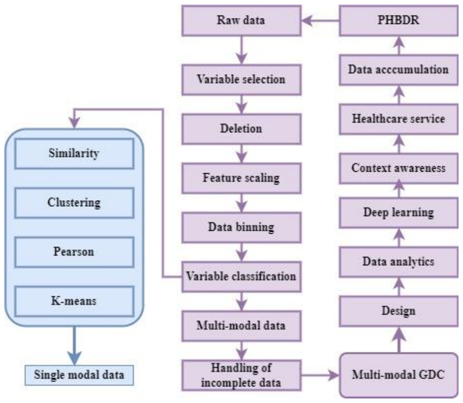

Figure. 1 GDC-MM-PHBDR framework

Pre-processing

This research utilizes the initial data released by NHANES from 2016 to 2022, which includes information from wellness, medical testing, and nutrition studies. The NHNES dataset has no information for the class value, necessitating accurate processing. If there are many scales for continuous parameters, it might lead to excessively high values or the integration of weights into 0. Inconsistent weight learning may occur due to differences in size, necessitating extra characteristic scaling [9]. Feature scaling is a method used to standardize the assortments of constraints to be consistently scaled. This methodology uses regression or clustering analysis to assess the effect of a specific attribute in both statistical analysis and NN modeling. Thus, data binning converts parameters into a certain kind using the feature scale. Binning is a method that classifies continuous features into breaks, sometimes called bins. NHNES uses information binning and characteristic classification to categorize variables for different units, such as age, height, and blood pressure. The parameters are categorized into eight intervals, ranging from 0%-100%

Finding the incomplete information using GDC-MM-PHBDR

The GDC used in traditional self-encoders represents a modification of the learning methods. To recover the original data without any noise, this self-encoder introduces random noise to the input data free from noise. This procedure involves iteratively modifying the input value by introducing random variations and restoring it to its original form. Before data is entered, an autoencoder selects a random number from the input and converts it to 0. In NN learning, incomplete information is often treated as 0.

Similarly, noise is estimated to be near zero and reintegrated into the original data during denoising autoencoder training when incomplete information is used. Thus, in the case of missing data, an NN trained with a non-zero predicted value will output a value of 0. An NN requires several hidden layers to convey the modality precisely based on the data attributes. In addition, various shapes may be generated by using a GDC and stacking hidden layers in several layers.

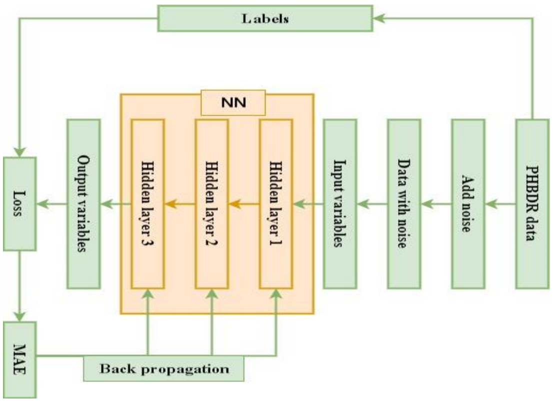

Figure. 2 Assessment of incomplete information using the GDC for PHBDR

The approach used in this context divides the learning process into unsupervised and supervised learning. Early GDCs were developed using a Restricted Boltzmann Machine (RBM) as a deep belief network. This technique addresses the challenges of extensive computational requirements, limited minimum, and disappearing gradient difficulties. Currently, using diverse propagation functions and optimizing schemes enhances the accessibility of learning when using backpropagation. This work included doing experiments using a supervised GDC. The input x' consists of data that contains randomly produced noise, with 25% of the values being absent and represented as 0. The label data comprises the original data x, which does not include any incomplete information. The GDC is taught via backpropagation, where the Mean Absolute Error (MAE) between the output x^ and the generated x is used as the loss function. This training is performed in a one-time learning process.

Fig. 2 depicts using the GDC to estimate the missing information for PHBDR. As seen in Fig. 2, the initial data undergoes adding noise and is then transformed into training data. Moreover, the weights in the NN are trained by using the mistake in the output values, with the real data serving as a label. Fig. 2 is a graphical representation of a GDC with a single input layer, three hidden levels, and a solitary output layer. The source data transactions have an incomplete information rate of roughly 25%. During the process of GDC learning, the error or noise factor gradually grows by 0.05, beginning with an initial value of 0.05 and continuing until it reaches a final value of 0.25. The activation function and optimizer used in this case are a Rectifier Linear Unit (RLU).

2. Results and discussion

The evaluation information in PHBDR is analyzed using the suggested GDC-MM-PHBDR, K-NN, Singular Value Decomposition (SVD), and column means (c-mean) methods and then ordered into the learning data. Within the c-mean scheme, any missing value is substituted with the average value of each column. Subsequently, the supervised learning technique is used to construct a model using the same DL approach to assess the resulting model. The NHNES data is pre-processed to provide 16,427 instances of public health big data evidence.

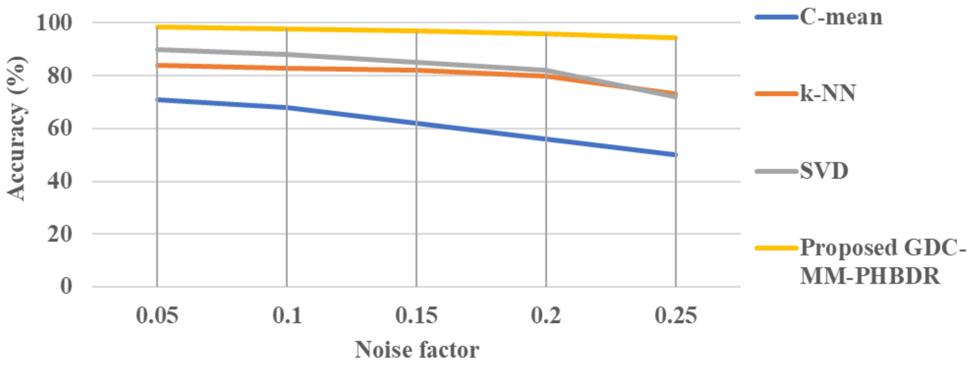

Figure. 3 Accuracy (%) of various DL methods for handling incomplete information in PHBDR

Fig. 3 depicts the accuracy of various DL methods for handling incomplete information in PHBDR. As the noise level grows from 0.05 to 0.25, the accuracy of all approaches typically drops. The GDC-MM-PHBDR approach continuously demonstrates excellent performance, attaining maximum accuracy across all noise levels. It begins at 98.31% accuracy with a noise factor of 0.05 and significantly decreases to 94.5% with a noise factor of 0.25. On the other hand, the other approaches exhibit more pronounced accuracy declines as the noise level increases: C-mean decreases from 71% to 50%, k-NN from 84% to 73%, and SVD from 90% to 72%. The results demonstrate that the suggested GDC-MM-PHBDR technique is superior in handling missing information in PHBDR compared to standard methods, making it more resilient and effective.

3. Conclusion and future scope

A Deep Learning (DL) health paradigm aims to enhance the accuracy of learning weights. A health model based on deep learning is trained by optimizing the weights to maximize accuracy. Excluding the learning process may diminish the accuracy of applying a DL-based health model to the user's circumstance. In such cases, the responsibility of managing Public Health Big Data Records (PHBDR) falls on individuals. This study proposed a generalized denoising coder with a multi-modal approach designed to handle incomplete information in public health big data records (GDC-MM-PHBDR).

Furthermore, the PHBDRs provide multimodality, which makes it possible to gather data for a single object from many sources. The GDC has a multi-modal organizational structure. Traditional self-encoders use a GDC that modifies the learning techniques. This self-encoder adds random noise to the noise-free input data to retrieve the original data without any noise. In this process, the input value is modified repeatedly by adding random changes and then returning it to its initial state. The GDC-MM-PHBDR method consistently achieves maximum accuracy at all noise levels, exhibiting exceptional performance. With a noise value of 0.05, it starts at 98.31% accuracy and drops dramatically to 94.5% with a noise level of 0.25. According to the results, the suggested GDC-MM-PHBDR approach works better than existing methods in resolving missing data in PHBDR, leading to higher efficacy and resilience.

Reference

[1] Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *Jama*, *322*(24), 2377-2378.

[2] Van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. Journal of personality assessment, 102(3), 297-308.

[3] Nielson, J. L., Cooper, S. R., Seabury, S. A., Luciani, D., Fabio, A., Temkin, N. R., ... & Track-TBI Investigators. (2021). Statistical guidelines for handling missing data in traumatic brain injury clinical research. Journal of neurotrauma, 38(18), 2530-2537.

[4] S. Neelima, Manoj Govindaraj, Dr.K. Subramani, Ahmed ALkhayyat, & Dr. Chippy Mohan. (2024). Factors Influencing Data Utilization and Performance of Health Management Information Systems: A Case Study. Indian Journal of Information Sources and Services, 14(2), 146–152. https://doi.org/10.51983/ijiss-2024.14.2.21

[5] Xu, D., Sheng, J. Q., Hu, P. J. H., Huang, T. S., & Hsu, C. C. (2020). A deep learning–based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients. IEEE Journal of Biomedical and Health Informatics, 25(6), 2260-2272.

[6] Mehak, S., Himanshi., & Sanju, S. (2024). Privacy-enhancing Blockchain Solutions for the Healthcare Sector: Efficient Message Sharing and Robust Big Data Protection. Journal of Internet Services and Information Security, 14(2), 85-97.

[7] Jäger, S., Allhorn, A., & Bießmann, F. (2021). A benchmark for data imputation methods. Frontiers in big Data, 4, 693674.

[8] Mohamed, K.N.R., Nijaguna, G.S., Pushpa, Dayanand, L.N., Naga, R.M., & Zameer, AA. (2024). A Comprehensive Approach to a Hybrid Blockchain Framework for Multimedia Data Processing and Analysis in IoT-Healthcare. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 94-108. https://doi.org/10.58346/JOWUA.2024.I2.007

[9] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining New York, NY, USA: Springer, 2015, pp. 59–139.

[10] KAVITHA, M. "A ku Band Circular Polarized Compact Antenna For Satellite Communications." National Journal of Antennas and Propagation 2.2 (2020): 15-20.

[11] Peralta, M., Jannin, P., Haegelen, C., & Baxter, J. S. (2021). Data imputation and compression for Parkinson's disease clinical questionnaires. Artificial Intelligence in Medicine, 114, 102051.

[12] Kearney, A., Rosala-Hallas, A., Rainford, N., Blazeby, J. M., Clarke, M., Lane, A. J., & Gamble, C. (2022). Increased transparency was required when reporting imputation of primary outcome data in clinical trials. Journal of Clinical Epidemiology, 146, 60-67.

[13] Dong, J., Wu, H., Zhou, D., Li, K., Zhang, Y., Ji, H., ... & Liu, Z. (2021). Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China. Journal of Medical Systems, 45(9), 84.