A MACHINE LEARNING ALGORITHM FOR MULTI-SOURCE HETEROGENEOUS DATA WITH BLOCK-WISE MISSING INFORMATION

Jayshree Ghorpade-Aher*

Research Scholar, Department of Computer Engineering, Pune Institute of Computer Technology, SPPU,
Assistant Professor, SCET, Dr. Vishwanath Karad MIT WPU, Pune, Maharashtra, India
jayshree.aherghorpade19@gmail.com

Dr.Balwant Sonkamble

Professor, Department of Computer Engineering, Pune Institute of Computer Technology, SPPU,
Pune, Maharashtra, India
basonkamble@pict.edu

Abstract

The heterogeneous data coming from varied sources is integrated to induce learning as it may have complementary information. The growing research for multi-source heterogeneous data analysis has extended the traditional methods of single source homogeneous data. One of the major challenges is processing the heterogeneous data while integrating the information from different sources with various missing data patterns. The potential features act as the useful predictors, but rejecting the observations with missing data values will lose the information which may create biased conclusions affecting the model strength for uniform decisions. A general imputation technique should be preferred for its computational simplicity and capability to induce a diminutive bias in the dataset. The proposed 'Heterogeneous Data and Weight Algorithm' constructs the robust model averaging techniques and feasible estimations with advanced Machine Learning techniques that implements the weighted probability approach to enhance the performance of the model with better predictive power.

***Keywords*: Machine Learning; heterogeneous data; block-wise missing; weights; probability.**

1. Introduction

Analyzing heterogeneous data with multiple sources is one of the imperative facets of computing, that is driven by data with the diverse features and dimensions. The data-driven computing techniques assist the data with effective feature engineering algorithms to explore and expedite the variation in the input data. A steady and understandable Machine Learning model can be constructed enabling proper generalization with reduced feature-set those contribute more towards the outcome. One of the simple ways is to merge the data coming from various sources and model them together to improve the performance. The meaningful information can be extracted from the variety data by analyzing the useful attributes to interpret appropriate features. While dealing with the multi-source heterogeneous data, identifying the missing values is one of the snags with an unexpected hidden challenge that affects the datastores. The real-time information like Electronic Health Records (EHRs) mostly contains such values. The analysis of integrated healthcare data that comes from multiple sources has difficulties to handle the missing data. The advanced Machine Learning algorithms along with the statistical models, discriminate analysis, etc. function better on complete observations for which the acquaintance of the outcome variable is essential to steer with this missing data. The incomplete data needs to be identified for the multi-source heterogeneous analysis. This missing data may have various possibilities such as the data value was erased accidently, the data entry of the respective value was overlooked or skipped, it may not be applicable to the instance, the data value was of no interest for the instance, networking failures while storing the data value, inaccurate data sources of few observations are not considered, few data collection procedures may be expensive to be applied to each and every instance, and others. Thus, the missing data mostly with high dimensions pose a threat and potential challenge to the model due to redundant information or corrupted entries. To construct a stable and coherent learning model with upright understanding, it is censorious to perform suitable data processing for handling the missing data values along with the feature engineering techniques.

1.1 Multi-source Heterogeneous Data with Block-wise Missing Information

In the real world, one of the simplest approaches is integrating the multi-source heterogeneous data, where all the sources are treated equally significant by overlooking the relationship within the source and between the sources. Also, the imputation techniques must be tailored considering the proportion of missing data with the domain of interest for the dataset. These missing data patterns needs to be handled correctly with various imputation techniques to improve the predictive power of the model. The block-wise missing information is processed with various statistical techniques by imputing the appropriate values. Most of the advanced Machine Learning techniques rely on the effective feature selection subset with different learning strategies. The removing or ignoring of missing values as shown in Fig. 1, will strongly lead to huge loss of data with biased outcomes.

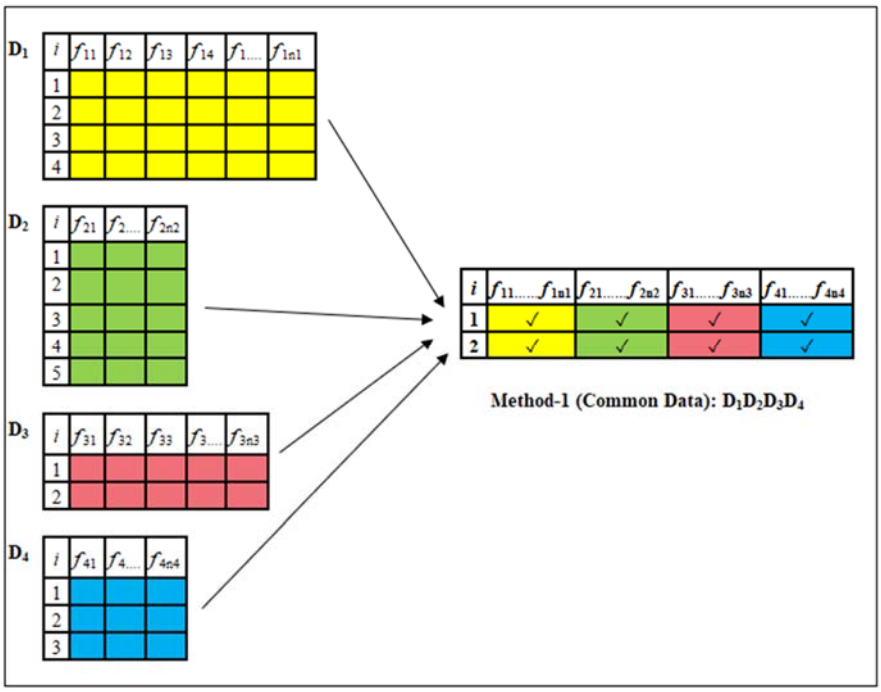

Fig. 1. Method-1: Multi-source Heterogeneous Data with Common Data

Method-1generates the common data observations with all the necessary features through the inner join of the four data sources. The healthcare applications demand for the integration of multi-source heterogeneous data. D_1, D_2, D_3 and D_4 be the four varied heterogeneous data sources with maximum features f_1n_1, f_2n_2, f_3n_3 and f_4n_4 respectively. The corresponding dimensional representation of D_1 data source is shown in Eq. (1), D_2 data source is shown by Eq. (2), Eq. (3) represents data source D_3, and Eq. (4) characterizes data source D_4. The integrated dimensional representation for the model with all the features for method-1 is depicted in Eq. (5).

$$D_1 = \{f_{1i} \mid 1 \le i \le n_1\} \ \forall f_{1i} \in \mathbb{R} \quad (1)$$

$$D_2 = \{f_{2i} \mid 1 \le i \le n_2\} \ \forall f_{2i} \in \mathbb{R} \quad (2)$$

$$D_3 = \{f_{3i} \mid 1 \le i \le n_3\} \ \forall f_{3i} \in \mathbb{R} \quad (3)$$

$$D_4 = \{f_{4i} \mid 1 \le i \le n_4\} \ \forall f_{4i} \in \mathbb{R} \quad (4)$$

$$D_1D_2D_3D_4 = \{f_{ij} \mid 1 \le i \le 4 \ \& \ [1 \le j \le fn_1 \vee 1 \le j \le fn_2 \vee 1 \le j \le fn_3 \vee \mid 1 \le j \le fn_4]\} \ \forall f_{ij} \in \mathbb{R} \quad (5)$$

Data processing is a time-consuming task that directly impacts the model building with effective performance. The technique of replacing or inserting these missing values is known as imputation. There are different imputation methods to overcome such challenges as the missing data will deceive to erroneous inferences and significantly affect the decisions. The heterogeneous multi-source data considered together appears in a block as either available (✓) or missing(x) as depicted in Fig. 2 and produced by Method-2 through the outer join of the four data sources. Different imputation techniques such as pair-wise deletion, complete observation deletion, zero imputation, central tendencies like imputations with arithmetic mean or geometric mean, mode substitutions,

model-based imputations like k-nearest neighbours (kNN), etc. are the statistical methods used to replace the missing data. The data intensive computing will explore the advanced processing techniques and solutions, which will help to establish the relation amongst various types of data to leverage the potential power of this data.

Although, various approaches are designed to handle the missing data, still the imputation methods remain unsuccessful in achieving the behaviour of the missing data, with arbitrarily dispersed values across the integrated data and thus appear block-wise. The significant loss of data would outcome instability in the results. While analyzing the multi-source heterogeneous data, it is difficult to estimate and approximate the data values with their importance as it becomes hard to know which data has significant contribution for the outcome or which data should be neglected for a specific application. These problems pose challenges and limit the accomplishment of the model towards the resultant.

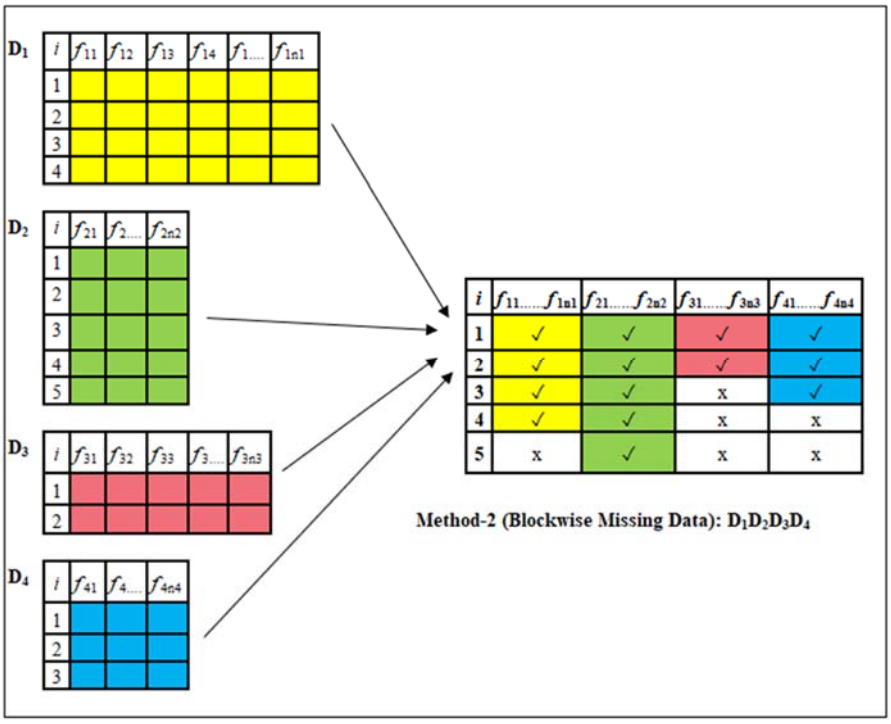

Fig. 2. Method-2: Multi-source Heterogeneous Data with Block-wise Missing data

The observations for the four varied heterogeneous data sources, D_1, D_2, D_3 and D_4 with maximum features f_1n_1, f_2n_2, f_3n_3 and f_4n_4 are ob_{1r1}, ob_{2r2}, ob_{3r3}, ob_{4r4} respectively. The missing element data-value dv_k for k^{th} observation with feature f_{ij} is represented in Eq.6.

$$dv_k = \begin{cases} data\ value\ \exists\ f_{ij} : \checkmark \\ data\ value\ \neg\exists\ f_{ij} : \text{x} \end{cases} \tag{6}$$

where, $k \leftarrow \{maximum\ (ob_{1r1} \vee ob_{2r2} \vee ob_{3r3} \vee ob_{4r4})\}$

The lack of useful attribute data values will affect the predictive power of the model. The proposed technique is consistent to smoothly handle the missing informative data since the 'Heterogeneous Data and Weight Algorithm' reduces the selection bias by aggregating additional samples across various missing patterns as compared to single imputation methods. The proposed algorithm works on the principle of sub-model generations from the multi-source heterogeneous data with model averaging based on effective probability weight estimations. The multiple sub-models based on the weight algorithmic techniques fasten the execution process by considering the available data sources and reducing the computational complexity. The experimentation is performed with the healthcare data that comes from multiple sources for analyzing the Type-2 Diabetes Milletus (T2DM) disease. The recent study by the American Diabetes Association (ADA) has discussed that the mortality rate has grown up for the patients with Diabetic Mellitus (DM) [Gerui et. al., (2021)] and thus needs intensive medical care [American Diabetes Association, (2021)] with multiple features. The previous research work is discussed in section 2. The next section 3, represents the research study of the proposed algorithm. Discussion in section 4 gives the experimental analysis. The conclusion section lastly ends with the concluding remarks.

2. Previous Research Study

The data must be explored at minute level to extract most of the information for appropriate predictions. The heterogeneous real-world multi-source data may have partial or incomplete data for few of the attributes or observations [Fei Xue and Annie Qu, (2021)]. Thus, it's crucial to build effective techniques to impute the missing data [Ghorpade and Sonkamble, (2020)]. Various methods and techniques are used to overcome the missing value data problems [Qiujun Lan and Shan Jiang, (2021)]. Consider the Electronic Health Record (EHR) data of the patients, where the digital values are preferred for the medical diagnosis. But many a times there are lacunae for these electronically recorded data due to networking problems while storing the data, data collection from sensors, failure of power supply, unable to record few medical parameters because of some unidentified reason or irrelevance of that parameter for the patient condition, etc. The state of art study applies that data is either missing due to unidentified reasons or may be few identifiable explanations with arbitrary and accidental causes. Imputing techniques if not properly applied to the missing data [Panda and Adhikari, (2020)] with its distribution may lead to unwanted bias in the dataset with misguiding results for further analysis. Many of the techniques do not consider the relationship between the attributes, which can be one of the better ways to generate the value of the missing data. One of the main concern and problems while processing or integrating the multi-source heterogeneous data is combining the data from different sources that generates various missing data patterns.

Corresponding to the influence of the extraneous or intervening attributes on the response or target variable, missing data can be alienated into structures. The three categories of missing data [X. Yu and Q. Wu, (2021)] are:

- Missing completely at random (MCAR: if the probability of missing value in an attribute is independent of the attribute itself and on any other attribute on the set);
- Missing at random (MAR: when the probability of a missing value on a sample 'X' is not dependent on 'X' but it obeys a pattern in the data-set and hence can be generated using other attributes);
- Not missing at random (NMAR: when the probability of missing values on 'X' completely depends on 'X' itself). The NMAR can be imputed using other samples or attributes in the data-set, whereas MCAR and MAR are recoverable. Various imputation processes such as statistical methods with mean and median techniques are used for computation purpose to replace the missing values.

The traditional methods alone do not perform well to handle the missing data. Dealing with incomplete or missing values for a dataset while maintaining the data reliability [Jan et. al., (2020)] may be a challenge. The real-time data mostly contains the missing values and proper substitutions to this data with suitable values is an important task. The simplest way will be replacing the missing values with some guess that best suits the data variable or completely remove the particular observation if most of the attribute values are absent. But these solutions will not work and unfortunately lead to biased outcomes which in turn will affect the decision making for the model with invalid evaluations. One of the imputation techniques is k-nearest neighbour (kNN) algorithm [Zhang S., (2018)] that deals with missing values and handles the data in an effective manner. The Average Imputation (AI) technique [Sharma et. al., (2021)] replaces the missing data value with the average of the conforming features over the complete dataset but it leads to variation in the standard deviation measures. The Model based Missing value Imputation using Correlation (MMIC) [Zahin et. al., (2018)] is an effective imputing technique for the numerical and categorical attribute values. Other methods such as Expectation Maximization Imputation (EMI) [Zhao and Duangsoithong, (2020)] calculates the imputation values by exploring the average mean and covariance matrix of the dataset. But the problem with this method is that it uses the highly correlated data amongst the attributes. Predictions of the final outcome can be improved with penalized regression techniques, where multiple sub-models will act as the candidate models and perform the weight calculations [Ghosh and Yuan, (2009)]. Particularly the model's output must be estimated correctly with objective function, residuals, standard errors, parameter investigations, etc. to analyze its sensitivity to minor agitations in the input due to missing data values or unfair data distributions to avoid the ambiguities in the results. Thus, the research gap identified states that there is a necessity to study and explore the concerns for model uncertainty. Models should be facilitated with optimal strategies by combining various base learners and sub-models to average approximations and adapt to the data patterns.

3. The Proposed Technique-Algorithm

Data either generated or collected, produces information which is analyzed to produce suitable decisions. It acts as a source of knowledge that discovers new patterns to assist the decision makers and predict optimal results. The advances in technology have facilitated the collection of multiple source heterogeneous data in different applications. The single sourced homogeneous data has categorical and numerical variables or features. As discussed in the literature study, the missing data [Salgado, (2016)] must be handled with suitable techniques to overcome the aforementioned challenges that may lead to bias analysis & wrong estimations. The multi-source

heterogenous data when integrated together for healthcare applications pose serious issues and complications with missing data patterns. The experimentation is performed with the healthcare multi-source heterogeneous data. The first phase of the research study for feature selection techniques with the proposed 'Ensemble Bootstrap Genetic Algorithm (EnBGA)' [Ghorpade-Aher and Sonkamble, (2022)] produced the effective features with the data driven approach. Each of the single source data (Diagnosis, Medication, Transcript & Lab data) generated their effective features with the proposed feature selection algorithm. The multi-source heterogeneous data has different types of features. Integrating such heterogeneous data with the effective features is a challenge as block-wise missing data is generated since few of the attribute's values are missing. Handling this missing data patterns due to failure of record data or unavailability of the data is an important task for building a robust Machine Learning model. These missing values can significantly affect the efficacy and performance of the classification model. Thus, an effective technique should be developed to impute these missing values.

3.1 Imputation Techniques

One of the most commonly used method is Zero Imputation (ZI) technique, where all the missing values are replaced with `zeros' and then the data is processed effectively. The Mean Imputation (MI) technique fills the missing value with the arithmetic average of the data values of that feature itself. The mean imputation is applicable for numerical variables and the mode imputation works for nominal variables. But the problem with these two techniques is that it approximates the distribution function with distortion which reduces the quality of the analyzed data and its performance [Garciarena and Santana, (2018)]. The kNN imputation as proposed by (Zhang, (2018)], is effective for handling the multi-source heterogeneous data. The missing data is replaced with the mean of the k-values coming from the 'k' most similar complete observations or samples. One of the best uses of the kNN algorithm is that given enough data it considers the correlation structure of the data and thus make better predictions [Xiang, (2015)]. The correlation among data attributes needs to be explored and understood. The traditional imputation techniques can severely hamper the classification model and distribution of the attributes which will have computational complexities.

Fig. 3 depicts the data distribution density plots of multi-source heterogeneous data for Feature#1 of Transcript data source with various missing data imputations applied to the actual data as shown in Fig. 3(a) with the kernel density estimate plot, which visualizes the probability estimations of the data variables. Fig. 3(b) represents the data distribution plot with Zero imputation technique. Fig. 3(c) represents the data distribution plot with Mean imputation technique. Fig. 3(d) represents the data distribution plot with Median imputation technique. Fig. 3(e) represents the spread of data by applying the kNN imputation technique.

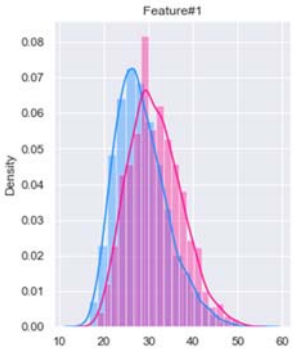

Fig. 3. (a) Actual Data

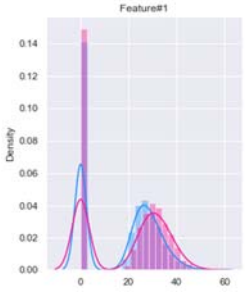

Fig. 3. (b) Zero Imputation

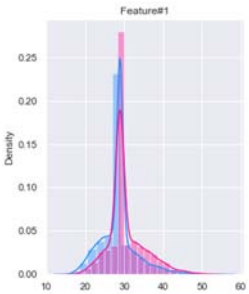

Fig. 3. (c) Mean Imputation

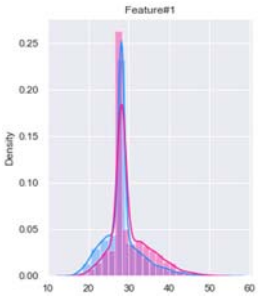

Fig.3. (d) Median Imputation

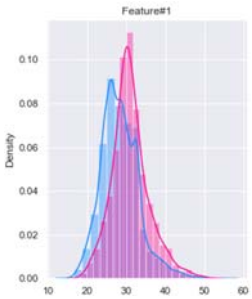

Fig. 3. (e) kNN Imputation

Fig. 3. Data distribution plots for Feature#1 with Missing Imputations

Fig. 4. displays the data distribution density plots of multi-source heterogeneous data for Feature#2 with various missing data imputations applied to the actual data that is shown in Fig. 4(a). As shown in Fig. 4(b) the data distribution plot with Zero imputation technique is executed. Fig. 4(c) represents the data distribution plot with Mean imputation technique. Fig. 4(d) represents the data distribution plot with Median imputation technique. Fig. 4(e) represents the spread of data by applying the kNN imputation technique.

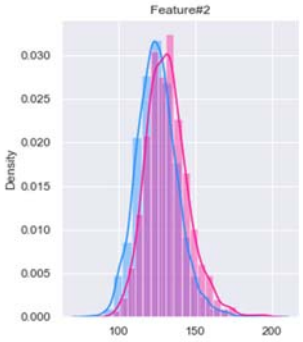

Fig. 4. (a) Actual Data

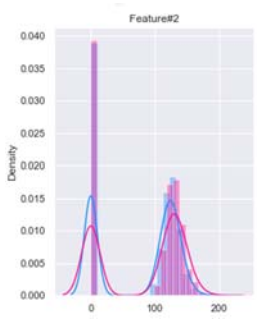

Fig. 4. (b) Zero Imputation

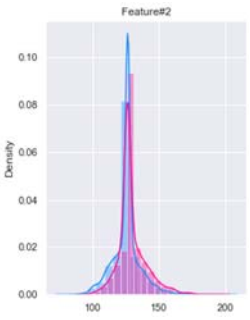

Fig. 4. (c) Mean Imputation

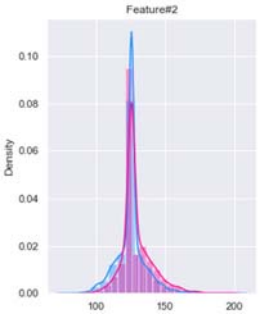

Fig. 4. (d) Median Imputation

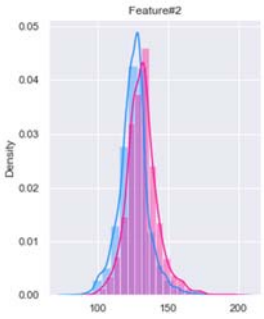

Fig. 4. (e) kNN Imputation

Fig. 4. Data distribution plots for Feature#2 with Missing Imputationss

The varied data distribution spread displayed in Fig. 3 and Fig. 4 exhibits the bias which is induced due to missing data patterns in heterogeneous data. The state-of-art study shows the research gap where increase in missing percentage of data leads to bias and variance that needs to be minimized to ensure revamped decision making. The mean and median imputations introduce the variables between '0' & '1', leading to the shift of binary data to continuous data which results in unnecessary distinctions that misleads the decision making with wrong conclusion. These unfair data distributions must be addressed to avoid the ambiguities in the final results.

3.2 Model Estimations with Weighted Probability

Multiple models of different Machine Learning classifiers are used for analysing the performance of the proposed 'Heterogeneous data & Weight Algorithm' to reduce generalization error. The block-wise missing data can be appropriately handled by exploring various patterns of the missing data and building respective sub-models. Various proposed weight calculation techniques such as *Equal* weight algorithm, *Rank* weight algorithm, *Absolute* weight algorithm, *Inverse* weight algorithm and *Composite* weight algorithm are designed for producing appropriate weights along with the probabilities for the sub-models. The *ROC_score* gives better performance than accuracy measure for most of the uneven data. It is based on the optimal threshold with predicted scores.

The performance of the proposed Machine Learning Model for 'Heterogeneous data & Weight Algorithm' is analysed with various classifiers. The statistical methods have revealed that penalized regression studies are being used to improve the predictions on outcomes by implementing model averaging [Alhorn et. al., (2021)] ideas. The sub-models have the ability to lessen the model uncertainty. The Ridge regression method imposes a penalty on the coefficients and diminish the ordinary least squares estimates of the regression. The purpose of model selection is to determine a model with f_k that best fits the data with 'n' sub-models as shown in Eq. (7) for the k^{th} model, where, $1 \leq i \leq n$.

$$logit\ P(Y_i) = f_k(X_{ki}, \beta_k) \tag{7}$$

The model averaging technique combines more than one feasible model that are built with weight updation. Let w_1, w_2, ..., w_n denote the weights assigned to the multiple sub-models of the model averaging estimator [Sun et. al., (2022)]. It can be observed that the prediction risk (which comes from the estimation error) of the model averaging technique is broken down into the between-model risk based on the calculated weights for the multiple candidate models.

The proposed research work considers the multi-source heterogeneous data uncertainty and suggests the aggregation of possible set of sub-models that assures the precise scientific principles. The proposed model averaging algorithm normalizes the coefficients over the identified sub-models depending on the different patterns of missing data and the determined weighted probability. The algorithm designs the advanced weight calculation Machine Learning techniques that emanates optimal predictive results adapting to the biases and variations in the input data with diverse research approaches. The experimental analysis shows that the estimated model weights as presented in Fig. 5, outperforms the traditional imputing techniques.

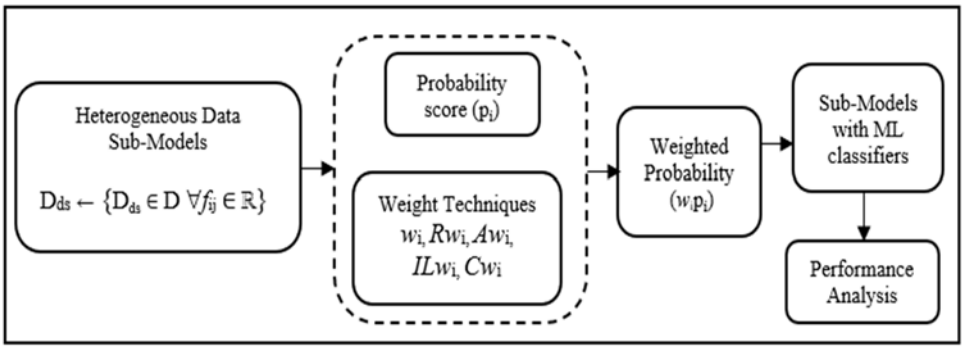

Fig. 5. Proposed Weighted Probability Techniques

'D' be the input heterogeneous data with multiple data sources $D_{ds} \in D$ as shown in Eq. (8). The effective feature sub-set for each of the source data is generated using the algorithm 'Ensemble Bootstrap Genetic Algorithm (EnBGA)'. The optimal cutoff or threshold for the *ROC_score* is determined and the probabilities for each of the sub-models are calculated. The *ROC_score* of the '*n*' sub-models is determined with A as shown in Eq. (9).

$$D = \int_1^4 D_{ds} \tag{8}$$

$$A = \{A_i \mid 1 \leq i \leq n\} \tag{9}$$

The proposed weight calculation techniques such as *Equal* weight Eq. (10), *Rank* weight Eq. (11), *Absolute* weight Eq. (12), *Inverse Loss* weight Eq. (13) and *Composite* weight Eq. (14) are applied with model averaging method. Further, the proposed algorithm 'Heterogeneous Data and Weight algorithm' aims at determining the sub-models to process the multi-source data with weighted probability approach, that makes the overall model reliable and robust.

3.3 Proposed 'Heterogeneous Data and Weight Algorithm'

Heterogeneous Data and Weight Algorithm

Step-1: Effective Features ← $\{f_{1n1} \in D_1, f_{2n2} \in D_2, f_{3n3} \in D_3, f_{4n4} \in D_4\}$

Step-2: D ← *Data sub-models* $\{D_1\ D_2,\ D_1D3,\ .....,\ D_1D_2D_3D_4\}$

Step-3: For each candidate sub-models $D_i \in D$

(3.1): $D_i \leftarrow$ *[input* (X_i)*, target* (y_i)*]*
(3.2): $A_i \leftarrow$ *{ROC_score for* X_i*}*
(3.3): $p_i \leftarrow$ *{Predict probabilities for* D_i*}*

Step-4: $A \leftarrow \{A_1, A_2, ..., A_n\}$

Step-5: Estimate Weight (w_i)

(5.1): *Equal* weight:

- For 'k' ∈ A:

$$w_k = \frac{\sum_{i=1}^{n} A_i}{n} \tag{10}$$

(5.2): *Rank* weight:

- $R_i \leftarrow \{\text{Rank } (A_i)\ \forall\ 1 \le i \le n\}$
- $RR_i \leftarrow \{\text{Reciprocal of } R_i\ \forall\ 1 \le i \le n\}$
- For 'k' ∈ RR_i:

$$Rw_k = \frac{k}{\sum_{i=1}^{n} RR_i} \tag{11}$$

(5.3): *Absolute* weight:

- For 'k' ∈ A:

$$Aw_k = \frac{k}{\sum_{i=1}^{n} A_i} \tag{12}$$

(5.4): *Inverse loss* weight:

- $L_i \leftarrow \{\text{Loss of } M_i\ \forall\ 1 \le i \le n\}$
- $IL_i \leftarrow \{\text{Inverse of } L_i\ \forall\ 1 \le i \le n\}$
- For 'k' ∈ IL_i:

$$ILw_k = \frac{k}{\sum_{i=1}^{n} IL_i} \tag{13}$$

(5.5): *Combined* weights:

$$Cw_k = \frac{(Aw_k + ILw_k)}{2} \tag{14}$$

Step-6: Calculate (w_i * p_i) for each of the sub-models

Step-7: Determine the performance metrics & perform the comparative analysis

4. Experimental Analysis

The meaningful information can be extracted from the variety data by analysing the useful attributes to interpret appropriate features. Various methods are used to build a model, but to estimate the best parameters that will optimize the predictive performance is one of the needs with current advanced techniques. The data processing for heterogeneous data necessitates the awareness to know the effect of biases along with variations to produce appropriate better insights. The proposed technique for Heterogeneous Data and Weight algorithms works on sub-model generations from the multi-source heterogeneous data. There are various Machine Learning techniques those are used to analyze the complex relationships of the multi-source data. Python programming is used for experimentation on spyder tool. The selected effective features for the online medical dataset in the form of Electronic Health Records [Alkundi and Momoh, (2020)] are used to generate the sub-models. The dataset consists of multiple sources such as Medication data, Diagnosis data, Transcripts data and Lab data. The experimentation result helps to analyse the data sub-models with weighted probabilities to explore the performance of the model. *ROC_score* is a useful metrics as it estimates the optimal threshold and is scale invariant. The comparative performance analysis of the proposed `Heterogeneous Data and Weight Algorithm' with *ROC_score* for the multi-source data with weighted probability approach on Actual data and varied Missing Imputes is shown in Table 1

using the classifiers Logistic Regression and Random Forest. Also, analysis is done using Gradient Boost and Decision Tree as shown in Table 2.

Technique	Logistic Regression (LR)			Random Forest (RF)		
	40% Missing	20% Missing	Actual Data	40% Missing	20% Missing	Actual Data
Zero Impute	0.71021	0.75583	0.80194	0.71614	0.74194	0.78803
Mean Impute	0.74173	0.77883	0.80238	0.73689	0.76237	0.80910
Median Impute	0.73332	0.77424	0.80143	0.72185	0.75888	0.79037
kNN Impute	0.74011	0.77907	0.80294	0.73991	0.76704	0.80656
Proposed Algorithm	0.76197	0.79059	0.81921	0.75527	0.78508	0.80970

Table. 1 Comparative performance analysis of Heterogeneous Data with LR and RF

Technique	Gradient Boost (GB)			Decision Tree (DT)		
	40% Missing	20% Missing	Actual Data	40% Missing	20% Missing	Actual Data
Zero Impute	0.72062	0.76387	0.80925	0.59969	0.59488	0.59264
Mean Impute	0.74862	0.78214	0.82540	0.60030	0.59203	0.61044
Median Impute	0.74094	0.78435	0.80818	0.58774	0.58995	0.59245
kNN Impute	0.74684	0.78492	0.82600	0.58935	0.58777	0.60447
Proposed Algorithm	0.76196	0.79846	0.81375	0.66976	0.72490	0.74772

Table. 2 Comparative performance analysis of Heterogeneous Data with GB and DT

Various proposed weight calculation techniques as discussed in the algorithm are designed for producing appropriate weights along with the probabilities for the sub-models.

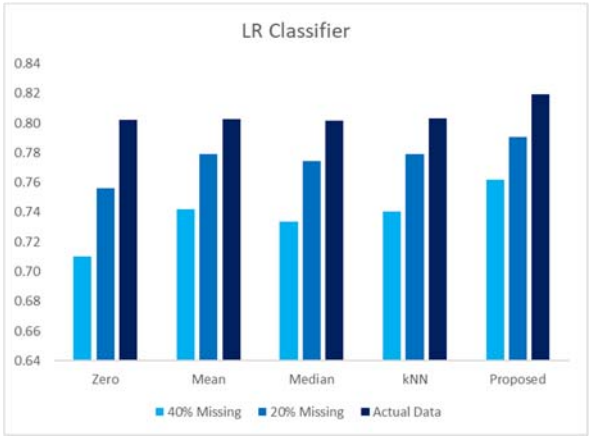

Fig. 6. Performance analysis of missing data patterns for Heterogeneous Data with Logistic Regression

It is observed that the proposed algorithm performs better for different block-wise missing data patterns and improves the predictive power of the model as shown in Fig. 6 using the Logistic Regression algorithm. Fig. 7 depicts the performance using Random Forest algorithm.

Fig. 7. Performance analysis of missing data patterns for Heterogeneous Data with Random Forest

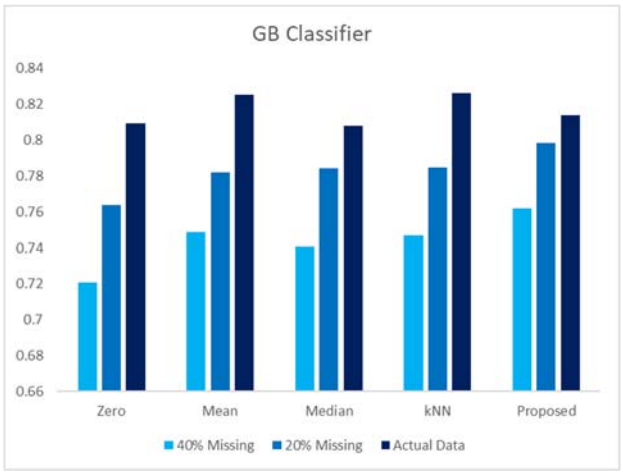

Fig. 8. Performance analysis of missing data patterns for Heterogeneous Data with Gradient Boost

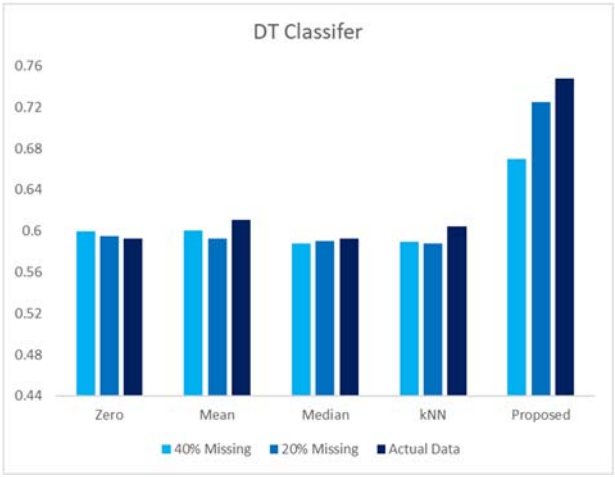

Fig. 9. Performance analysis of missing data patterns for Heterogeneous Data with Decision Tree

Fig. 8 depicts the performance analysis using the Gradient Boost algorithm, whereas Fig. 9 visualizes the performance applying the Decision Tree algorithm.

5. Conclusions

The proposed technique, 'Heterogeneous Data and Weights Algorithm' is a robust technique for handling missing data patterns for multi-source data. This method performs better when compared with the different missing treatments on block-wise data with sub-models. The multiple sub-models based on the proposed weight algorithms fasten the execution process by considering the available data sources. The experimental analysis

depicts that the sub-models implement the weighted probability approach to improve the overall performance of the model and increases the predictive power. The proposed method enforces additional weight on estimating functions from groups with either rarer missing values or more accurate imputation. The analysis of heterogeneous data appeals that these plausible sub-models along with their determined weighted probabilities, reduces the missing-data error and improves predictions with appropriate model selection uncertainty. Thus, the proposed weighted probability and model-averaging techniques explores the predictions based on the diverse patterns of sub-model and their predictive bias along with variance. The designed algorithm for heterogeneous data is highly effective with its impact towards more missing data patterns. Other Machine Learning classifiers along with Deep Learning techniques can be used further to analyze the missing data patterns for the multi-source heterogeneous data.

References

[1] Ghorpade-Aher J., Sonkamble B., (2022), "Effective Feature Selection Using Ensemble Techniques and Genetic Algorithm," in: Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, Vol 236, Springer, Singapore, Web of Science. https://doi.org/10.1007/978-981-16-2380-6 32.
[2] X. Yu and Q. Wu, "Multi-source Heterogeneous Data Association Technology to Build Public Safety Big Data Integration Research," 2020 International Conference on Big Data Economy and Information Management (BDEIM), 2020, pp. 17-20, doi: 10.1109/BDEIM52318.2020.00012.
[3] Fei Xue and Annie Qu, "Integrating multi-source block-wise missing data in model selection," May'2021, arXiv.org, arXiv:1901.03797
[4] Gerui Li, Ze Chen, Zhan Lv, Hang Li, Danqi Chang, Jinping Lu, "Diabetes Mellitus and COVID-19: Associations and Possible Mechanisms," International Journal of Endocrinology, Volume 2021, Article ID 7394378, April 2021.
[5] J. Ghorpade and B. Sonkamble, "Predictive Analysis of Heterogeneous Data – Techniques & Tools,' 2020 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, pp. 40-44, May 2020.
[6] American Diabetes Association, "Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021," Diabetes Care 2021, Volume-44, Issue-Suppl. 1, pp.S15 - S33, Jan 2021.
[7] Qiujun Lan, Shan Jiang, "A method of credit evaluation modeling based on blockwise missing data Springer,"Applied Intelligence, Feb'2021, doi.org/10.1007/s10489021-02225-5
[8] M. Z. Jan, J. C. Munoz and M. A. Ali, "A novel method for creating an optimized ensemble classifier by introducing cluster size reduction and diversity," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2020.3025173, September 2020.
[9] B. S. Panda and R. Kumar Adhikari, "A Method for Classification of Missing Values using Data Mining Techniques," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132935.
[10] Zahin, S. A., Ahmed, C. F., & Alam, T. "An effective method for classification with missing values," Applied Intelligence, Springer, Science+Business Media, 2018, doi 10.1007/s10489-018-1139-9
[11] Y. Zhao and R. Duangsoithong, "Empirical Analysis using Feature Selection and Bootstrap Data for Small Sample Size Problems,' IEEE 16th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Pattaya, Chonburi, Thailand, pp. 814-817, January 2020.
[12] Sharma, N., Dev, J., Mangla, M. et al., "A Heterogeneous Ensemble Forecasting Model for Disease Prediction," Springer, New Generation Computing, pp.1-15, January2021.
[13] D. Ghosh, Z. Yuan., "An improved model averaging scheme for logistic regression," J Multivar Anal., NIH Public Access, 100(8), 2009, pp.1670-1681, doi:10.1016/j.jmva.2009.01.006
[14] Zhang S., "Nearest Neighbor Selection for Iteratively kNN Imputation," Journal of Systems and Software, Vol. 85, No. 11, pp. 2541–2552, 2018 doi:10.1016/j.jss.2012.05.073
[15] Garciarena, U., & Santana, R. "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers, " Expert Systems with Applications, Vol. 89, pp. 52–65, 2018, doi:10.1016/j.eswa.2017.07.026
[16] Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., & Ye, J., "Multi-source learning with block-wise missing data for Alzheimer's disease prediction," Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, doi:10.1145/2487575.2487594
[17] Alamin Alkundi, Rabiu Momoh, "COVID-19 infection and diabetes mellitus," Commentary, Journal of Diabetes, Metabolic Disorders & Control, vol. 7, issue 4, pp.119-120, November'2020.
[18] Zhang, Zili & Yang, Pengyi, "An Ensemble of Classifiers with Genetic Algorithm Based Feature Selection," IEEE Intelligent Informatics Bulletin, Vol-9,2008, pp. 18-24
[19] Cátia M. Salgado, Carlos Azevedo, Hugo Proença, Susana M. Vieira, 'Missing Data,' Secondary Analysis of Electronic Health Records, 2016, Springer, Cham., ISBN: 978-3-319-43740-8, pp.143-161
[20] Alhorn, K., Dette, H. & Schorning, K., "Optimal Designs for Model Averaging in non-nested Models," Sankhya A 83, 745–778 (2021). https://doi.org/10.1007/s13171-020-00238-9
[21] Yuying Sun, Xinyu Zhang, Alan T.K. Wan, Shouyang Wang, "Model averaging for interval-valued data," European Journal of Operational Research,Volume 301, Issue 2, 2022, Pages 772-784, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2021.11.015.

Authors Profile

Jayshree Ghorpade-Aher, has completed her full time M.E. (Computer Engineering) from Pune Institute of Computer Technology (PICT), SPPU, Pune and also pursuing her Ph.D. in Computer Engineering from PICT. She is working as an Assistant Professor in School of CET, Dr. Vishwanath Karad MIT World Peace University, Pune. Her areas of interests are Machine Learning & Soft Computing.

Dr. Balwant Sonkamble, has completed his Ph.D. in Computer Science & Engineering from SGGSIE&T, Swami Ramanand Tirth Marathwada University, Nanded. He is working as a Professor in the Department of Computer Engineering, PICT, SPPU, Pune. His areas of interests are Artificial Intelligence, Machine Learning & Speech Processing.