Genetic algorithm feature selection resilient to increasing amounts of data imputation

Maryam Kebari, Annie S. Wu
University of Central Florida
Orlando, Florida, USA

Abstract

This paper investigates the robustness of a genetic algorithm (GA) in feature selection across a dataset with increasing imputed missing values. Feature selection can be beneficial in predictive modeling to reduce computational costs and potentially improve performance. Beyond these benefits, it also enables a clearer understanding of the algorithm's decision-making processes. In the context of real-world datasets that can contain missing values, feature selection becomes more challenging. A robust feature selection algorithm should be able to identify the key features despite missing data values. We investigate the effectiveness of this approach against two other feature selection algorithms on a dataset with increasingly imputed values to determine whether it can sustain good performance with only the selected features. Our results reveal that compared to the other two methods, the features selected by GA resulted in better classification performance across different imputation rates and methods.

We explore a genetic algorithm (GA) for feature selection in predictive modeling, demonstrating its effectiveness and robustness against increasing amounts of missing data. This approach enhances the speed and accuracy of classification algorithms by effectively selecting relevant features despite incomplete datasets, addressing the challenge of working with real, often incomplete data. Previous research showed this GA method also maintains robust feature selection with smaller training sets (Norat, Wu, and Liu 2023). In this work, we examine the impact of data imputation method and the amount of data imputed on the ability of this GA approach to perform feature selection. Questions that we ask include: How does increasing amounts of imputed data affect the set of features selected? How does data imputation method affect the set of features selected? We examine the GA's performance relative to these two questions and compare our observations with the feature selection performance of two methods (selected from two other categories of feature selection algorithms). We will examine what are the features selected and how data imputation affects consistency. In addition, we will compare the performance of a neural network using the selected features to build a predictive model.

Methodology and Experimental setup

We examine six different data imputation methods. Mean/Mode/Median, KNN(Pedregosa et al. 2011; Troyanskaya et al. 2001), Linear Regression (Pedregosa et al. 2011), Bayesian Ridge(Pedregosa et al. 2011), and MICE(Van Buuren and Groothuis-Oudshoorn 2011).

The GA we use belongs to the wrapper feature selection category. To ensure a thorough study, we compare it to two additional methods the Chi-squared method from the filter category and the LASSO method from the embedded category. Our dataset consists of 40,662 data points. Each data point has 25 features, and a binary label indicating whether it represents a PT office receiving more than the median standardized payment amount (target class).

To compare the performance of different feature selection methods, we perform classification using data containing only the *selected features* and compare the results with those obtained using *all features* in the dataset.

To test the effectiveness of the selected features, we use a feedforward neural network model with three layers: two hidden layers with ReLU activation and one output layer with a softmax activation function. The model is trained using categorical cross-entropy as the loss function and the Adam optimizer.

For every data imputation method and each feature selection technique, we conduct experiments with varying percentages of missing data, ranging from 0.20% to 0.80%. The missing data is generated by randomly removing data point values.

Results

We assess classification performance across various data removal levels (0, 40%, 80%) and imputation methods, examining the impact of feature selection on Accuracy, Precision, Recall, F1-score, and ROC-AUC score. Performance ratios, depicted in heatmaps from Figure 1 to Figure 5, compare outcomes using selected features against using the entire dataset. These heatmaps illustrate ratios with color intensity, where numbers above one, highlighted in yellow, signify improved performance with feature selection.

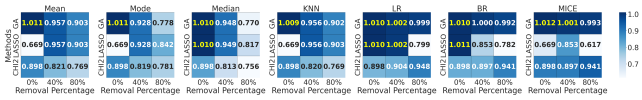


Figure 1: Heatmaps of Various Imputation Methods (Accuracy)

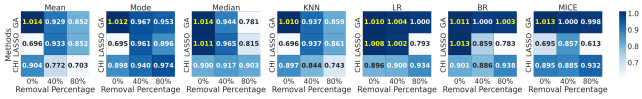


Figure 2: Heatmaps of Various Imputation Methods (Precision)

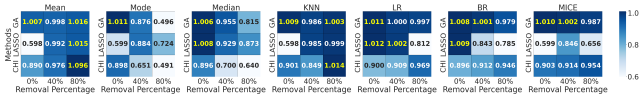


Figure 3: Heatmaps of Various Imputation Methods (Recall)

Accuracy, precision, recall, F1-score, and ROC-AUC metrics all show similar trends as detailed in Figures 1, 2, 3, 4, and 5, respectively. Performance typically decreases with more data removal, especially using simple imputation methods. However, GA maintains strong results across metrics, particularly with complex imputations like LR and MICE, demonstrating high robustness. LASSO's performance varies, closely matching GA initially then dropping, while chi-squared's performance is inconsistent, improving with complex imputations. Overall, GA exhibits exceptional stability and high scores, even with significant data removal.

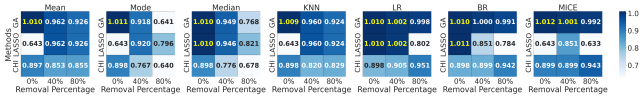


Figure 4: Heatmaps of Various Imputation Methods (F1-Score)

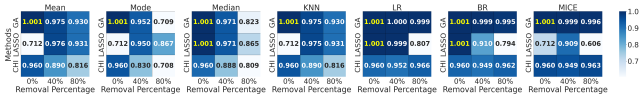


Figure 5: Heatmaps of Various Imputation Methods (ROC-AUC)

GA's performance declines more with simpler univariate imputation methods as data removal increases, but remains stable with complex multivariate strategies, closely mirroring performance with all features intact. Conversely, LASSO occasionally matches GA but is notably inconsistent, with effectiveness varying significantly with dataset completeness. We look at examples from GA to understand the results better. In Figure6 to Figure 8, the $x-axis$ represents the different features. We have ten lines in each feature group. Each line indicates a single run. The $y-axis$ represents the corresponding feature weights. In Figure 6, we see the GA feature

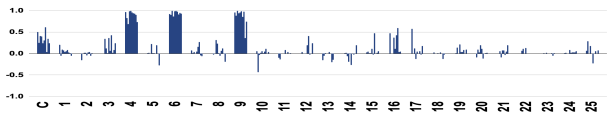
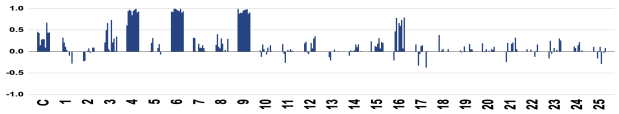
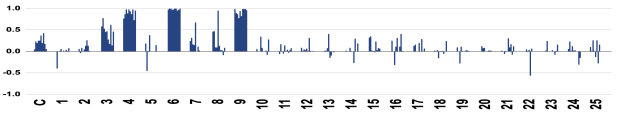


Figure 6: GA feature weights with No imputation

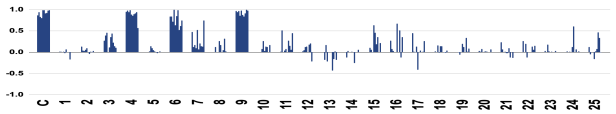


(a) GA feature weights 40% missing, imputed via MICE

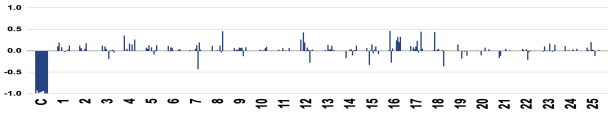


(b) GA feature weights 80% missing, imputed via MICE

Figure 7: GA feature weights imputed via MICE



(a) GA feature weights 40% missing, imputed via Median



(b) GA feature weights 80% missing, imputed via Median

Figure 8: GA feature weights imputed via Median

weights of the ten runs with 0% missing values. It selects features 4, 5, and 9 as the best features. In Figure 7a and 8a, GA can identify the same key features in both, although in the Median method, the weight of the constant value increases. When the data removal percentage goes up to 80%, GA can still identify the key features with the MICE method; however, when using the Median method, GA's ability to find the features diminishes.

Conclusion

This paper examines the robustness of GA based feature selection with increasing amounts of imputed data in predictive models. Feature selection, crucial for reducing computational costs and improving model performance, becomes challenging with incomplete data, often addressed through imputation. The study tests GA feature selection on datasets with various imputation methods, comparing its performance against methods like Chi-squared and LASSO. Results show GA maintains strong performance across various imputation scenarios and metrics (accuracy, precision, recall, F1, and ROC-AUC), even with significant data removal, particularly with complex imputation techniques like MICE, underscoring its robustness with heavily imputed data.

References

Norat, R.; Wu, A. S.; and Liu, X. 2023. Genetic algorithms with self-adaptation for predictive classification of medicare standardized payments for physical therapists. *Expert Systems with Applications* 218:119529.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; and Altman, R. B. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics* 17(6):520–525.

Van Buuren, S., and Groothuis-Oudshoorn, K. 2011. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software* 45:1–67.