Comparative Analysis of Feature Selection Methods with a Focus on Genetic Algorithm for Improved Predictive Modeling

By: Guillermo Munoz-Perez

UNIVERSIDAD DE CHILE

Introduction

Within the realm of computational science, the handling of high-dimensional data has become a crucial process in the ability to comprehend and garner meaningful insights. Feature selection methods are pivotal in machine learning for enhancing model performance by reducing dimensionality and eliminating irrelevant features. This report introduces several feature selection methods categorized as filter, embedded, and hybrid approaches: Chi-Square Test (CST), Ridge Regression (RR), and Genetic Algorithm (GA). Our research begins with an in-depth analysis of the Genetic Algorithm (GA), a technique inspired by the process of natural selection. GA employs a population of candidate solutions that evolve over generations to optimize a given objective function. We explore how these methods can be used to select optimal features in a classification dataset. Additionally, we compare the performance of GA with traditional feature selection methods such as Chi-Square Test and Ridge Regression. The report concludes with a comprehensive overview of our research findings and discusses future directions in the field of feature selection.

Methodology

In the first few weeks, our research focused on understanding and implementing various feature selection methods. We began with the Chi-Square Test (CST), then moved on to Ridge Regression (RR), and finally explored the Binary Genetic Algorithm (GA) as a feature selection method. For each method, we integrated the selected features into a classifier and compared the results.

Each feature selection method was implemented using the following structured approach:

1. Data Preprocessing:
- Normalize the data to ensure better results.

2. Feature Selection:
- Apply the feature selection method to the normalized data.
- Create a new dataset comprising only the selected features.

3. Model Evaluation:
- Input the new dataset into a classifier such as Neural Network or K-NN.
- Measure performance metrics like accuracy, precision, F1 score, and recall.

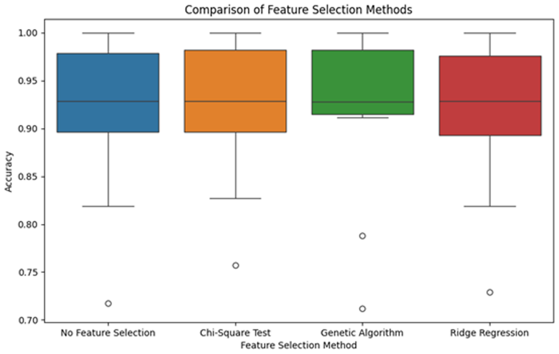
This process was repeated for each feature selection method to understand its strengths and weaknesses when applied to high-dimensional datasets.

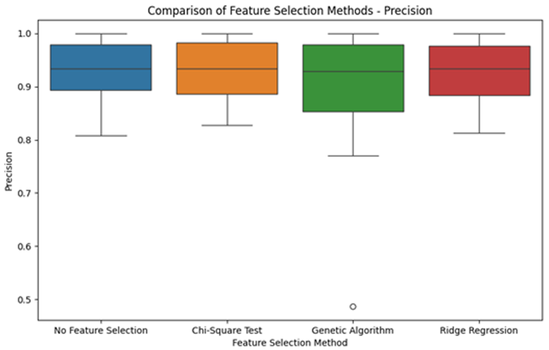
Results

In this section, we compare the performance of four feature selection methods: No Feature Selection, Chi-Square Test, Genetic Algorithm, and Ridge Regression, using four metrics: Accuracy, Precision, Recall, and F1 Score. The comparison results are illustrated in Figures 1 to 4.

Accuracy (Figure 1)

The box plot in Figure 1 shows that the median accuracy for all four methods is similar, with some variability across datasets. The Friedman test (p = 0.409) indicates no statistically significant difference in accuracy among the methods. The Nemenyi post-hoc test confirms this, showing no significant pairwise differences.





Conclusion

In this study, we explored and compared various feature selection methods, focusing on the Genetic Algorithm (GA) and its integration with traditional techniques like Chi-Square Test (CST) and Ridge Regression (RR). In conclusion, our research demonstrates that the Genetic Algorithm, particularly when integrated with traditional methods, offers a powerful tool for feature selection in high-dimensional datasets. This hybrid approach not only improves model accuracy but also provides a versatile solution applicable to various types of data and classification tasks.

References

[1] W. J. Yu, X. Y. Liu, and K. W. Wong, "An improved binary particle swarm optimization for feature selection," Applied Soft Computing, vol. 9, no. 1, pp. 552-558, Jan. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494609002270. [Accessed: Jul. 25, 2024].

[2] D. Wallis, "Comparing classifiers (Friedman and Nemenyi tests)," Medium, Oct. 3, 2021. [Online]. Available: https://medium.com/@diogeneswallis/comparing-classifiers-friedman-and-nemenyi-tests-32294103ee12. [Accessed: Jul. 25, 2024].