Multi-objective Feature Selection Optimization Using NSGA-2

Hamza Rangwala  
 Department of Science  
 Wilfrid Laurier University  
 Waterloo, ON, Canada  
hamzarangwala51@gmail.com

Azam Bidgoli  
 Department of Science  
 Wilfrid Laurier University  
Waterloo, ON, Canada   
aasilianbidgoli@wlu.ca

ABSTRACT

Feature selection is critical in machine learning, particularly with high-dimensional datasets. This paper introduces an innovative approach using the NSGA-II (Non-dominated Sorting Genetic Algorithm II) for feature selection, aiming to minimize the number of features and classification error simultaneously. A custom function was implemented to facilitate this process. The K-Nearest Neighbors (KNN) classifier evaluated the selected features across multiple runs and folds to ensure robustness. The methodology involved optimizing feature selection on training sets and evaluating on validation sets, identifying Pareto front solutions, and testing on the test set for generalization capability. Results showed effective feature reduction while maintaining or improving classification accuracy. Key metrics, such as average training error, minimum validation error, and test error, were tracked to highlight consistency. Non-dominated binary vectors of optimal feature subsets were analyzed for further feature selection via voting. This study provides a robust framework for feature selection in high-dimensional data, leveraging evolutionary algorithms and cross-validation techniques to potentially enhance machine learning model performance.

CCS CONCEPTS

• Insert CCS text here • Insert CCS text here   • Insert CCS text here

KEYWORDS

Insert keyword text, Insert keyword text, Insert keyword text, Insert keyword text

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 INTRODUCTION

In the era of big data, the exponential growth of available information has led to increasingly high-dimensional datasets, posing significant challenges for machine learning applications. These datasets, characterized by many features, often include redundant, irrelevant, or noisy data, which can negatively impact the performance of machine learning models. Consequently, feature selection has emerged as a critical preprocessing step to enhance model accuracy, reduce computational complexity, and improve interpretability.

Feature selection aims to identify the most relevant subset of features that contribute significantly to the predictive accuracy of a model. By eliminating unnecessary features, the dimensionality of the data is reduced, leading to simpler models that are easier to interpret and faster to train. Moreover, effective feature selection can mitigate the risk of overfitting, thereby enhancing the generalization capability of the model on unseen data.

Traditional feature selection methods, such as filter, wrapper, and embedded techniques, have been widely used. However, these methods often struggle with the curse of dimensionality and may not efficiently explore the vast search space of potential feature subsets. To address these limitations, evolutionary algorithms have gained popularity due to their ability to perform global searches and handle complex optimization problems. Among these, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) has shown promise in multi-objective optimization tasks, making it a suitable candidate for feature selection.

This study introduces an innovative approach utilizing NSGA-II for feature selection, with a dual objective of minimizing the number of features and the classification error simultaneously. The NSGA-II algorithm is well-suited for this task as it can effectively balance the trade-off between these conflicting objectives by identifying a set of Pareto optimal solutions. Each solution on the Pareto front represents a potential feature subset that optimizes the trade-off between the two objectives.

To facilitate the feature selection process, a custom function was developed which is used to diversify the initial population based on the number of features and generating random values to set the cells of the individuals and integrated with the NSGA-II algorithm. The selected feature subsets were evaluated using the K-Nearest Neighbors (KNN) classifier, a widely used non-parametric method known for its simplicity and effectiveness. The performance of the KNN classifier, in conjunction with the selected features, was assessed across multiple runs and folds of the dataset, ensuring the robustness and generalizability of the results.

The methodology employed in this study involved dividing the dataset into training and validation sets. The feature selection process was optimized on the training set, and the selected features were then evaluated on the validation set to identify Pareto front solutions. These solutions were further analyzed for their performance on the validation set, and the subsets with the minimum classification error were tested on a separate test set to evaluate their generalization capability.

The results of this study demonstrated that the proposed approach effectively reduced the number of features while improving classification accuracy. Key performance metrics, such as average training error, minimum validation error, and test error which were tracked across multiple runs, highlighting the consistency and reliability of the feature selection process. Additionally, the non-dominated binary vectors representing the optimal feature subsets were saved and analyzed for further selection using a voting mechanism, providing valuable insights into the feature selection process.

In summary, this research offers a robust framework for feature selection in high-dimensional data, leveraging the strengths of evolutionary algorithms and rigorous cross-validation techniques. The findings contribute to the field by offering a scalable and effective solution for feature selection, potentially improving the performance of various machine learning models.

3 METHODOLOGY

The updated template, user manuals, samples, and required fonts, all are available at the URL <https://www.acm.org/publications/proceedings-template>. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

 (1)

**Continuation part of Paragraph Text** The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**



The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

1.1 Heading Level 2

In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] Patricia S. Abril and Robert Plant, 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan, 2007), 36-44. DOI: <https://doi.org/>10.1145/1188913.1188915.

[2] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI:https://doi.org/10.1145/567752.567774

[3] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. DOI:https://doi.org/10.1007/3-540-09237-4.

[4] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY..

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00