Feature Selection Using Law of Total Variance with Fast Correlation-Based Filter

Nur Atiqah Mustapa
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah,
Gambang, Pahang, Malaysia.
atiqahnurmustapa@gmail.com

Azlyna Senawi
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah,
Gambang, Pahang, Malaysia.
azlyna@ump.edu.my

Chuan Zun Liang
Centre for Mathematical Sciences, Universiti Malaysia Pahang Al-Sultan Abdullah,
Gambang, Pahang, Malaysia.
chuanzunliang@ump.edu.my

2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS) | 979-8-3503-1093-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSECS58457.2023.10256367

***Abstract*—The increased dimensionality of data poses a formidable obstacle to completing data mining tasks. Due to the extraneous features associated with high-dimensional data, processing and analysis took longer and were less precise. As a pre-processing phase in the analysis of data mining tasks, feature selection is effective at reducing dimensionality, removing irrelevant characteristics, increasing accuracy, and enhancing the readability of the results. This research proposes the law of total variance with fast correlation-based filter (LTVFCBF) as a new feature selection method. LTVFCBF chose the significant features by identifying relevant features and remove redundant features among the relevant ones. The analysis was conducted with ten datasets of varied dimensionality to evaluate the performance of the proposed LTVFCBF and validated using four classifiers: *K*-nearest neighbours, Naïve Bayes, support vector machine, and bagging. The LTVFCBF and LTV methods have been compared in terms of the number of selected features, classification accuracy, and execution time. In overall, the suggested LTVFCBF has the potential to minimize the dimensionality of data by selecting a lower number of significant features with better accuracy. However, it requires a slightly higher execution time compared to LTV. Aside from that, LTVFCBF can achieve comparable accuracy with faster execution time when less than half of the original features are maintained. The proposed method can produce a promising outcome and may be regarded as an effective filter approach for feature selection.**

Keywords—feature selection, dimensionality reduction, law of total variance, Pearson correlation coefficient, filter feature selection

I. INTRODUCTION

In recent years, the number of instances and attributes of data has increased in numerous applications. This increment is because the capacity to produce and gather data has developed rapidly, and the quantity of data has grown by over 50% every month compared to the previous years [1]. Its magnitude may pose a significant difficulty for many machine learning algorithms regarding scalability and learning efficiency. For example, high-dimensional data can contain a significant amount of unnecessary information, significantly hindering the performance of learning algorithms. Considering this, feature selection is now crucial for machine learning applications involving high-dimensional data.

Feature selection is a common technique for dimensionality reduction, which seeks to reduce the number of features in a dataset [2]. Feature selection focuses on optimizing the original features by preserving the primary data in the dataset to facilitate subsequent analysis. This method eliminates unneeded or superfluous features without altering their original characteristics. In various scenarios, it is necessary to minimize the dimensional space and evaluate significant elements without altering the original nature of the datasets [3]. Feature selection has been a fruitful area of study and helps reduce extraneous features, enhance the efficiency of learning tasks, and enhance learning performance [4][5][6].

There are three basic categories of feature selection models: filter, wrapper and hybrid model [7][8]. In removing unnecessary and redundant features, the filter model is not dependent on the machine learning algorithm. Instead, this model evaluated the significance of a subset of features based on the inherent properties of the dataset. This model is more computationally efficient and does not inherit any bias from the classifier, as it was created without actively optimizing the performance of any particular classifier. In contrast to the filter model, the wrapper model evaluates the relevance of feature subsets using a machine-learning algorithm. Wrapper model typically discovers characteristics that are better suit to the predefined learning algorithm, resulting in greater learning performance, but it is typically more computationally expensive than the filter model. When a high number of features are present, the filter model is typically selected due to its computational efficiency. The hybrid model focuses primarily on merging filter and wrapper models in order to get the greatest possible performance with a specific learning algorithm as that can be achieved by the wrapper while having the same time complexity as by the filter model.

This work focuses on the filter model and seeks to build a new feature selection method that can efficiently remove irrelevant and redundant feature with low computational complexitiy while suffers minimal loss in classification accuracy. The *K*-nearest Neighbours (KNN), Naïve Bayes (NB), support vector machine (SVM), and bagging (BG) will be utilized to validate the performance of the proposed strategy. These classifiers were selected because of their efficacy and applicability in evaluating feature selection methods [9].

II. RELATED WORK

All feature selection methods pass through subset generation, evaluation, stopping criterion, and result validation during the selection procedure. Subset generation is a technique that uses a search strategy to generate selectable feature subset candidates. The search strategy is one of the essential aspects of subset feature generation while looking for the ideal subset [8]. There are three main strategies: exhaustive search, sequential search, and random search. The significance of each proposed feature subset is then assessed using either independent or dependent criteria.

The filter technique selects features by independently analyzing potential feature subsets, such as correlation and information gain [10][8]. The filter method is more