Journal of Data Science and Modeling, Vol. 2, No. 1, 209-218, December 2023

Research Manuscript

Feature Selection in High Dimensional Datasets based on Adjacency Matrix

Behrang Ebrahimi [1], Negin Bagherpour[1]*

[1]Department of Engineering Sciences, University of Tehran, Tehran, Iran.

Recieved: 30/07/2024 Accepted: 09/11/2024

Abstract: Feature selection enhances classification and clustering by improving machine learning performance and reducing computational costs through the removal of irrelevant and redundant features. However, many existing methods neglect complex feature relationships and fail to capture high-order dependencies, often due to traditional algorithms' limitations in handling nonlinear relationships. This paper introduces a novel feature selection algorithm based on an adjacency matrix designed for supervised data. The algorithm works in three steps: first, it eliminates irrelevant features by evaluating each feature's correlation with its class. Next, it assesses pairwise feature relationships, constructing an adjacency matrix of selected features. Finally, clustering techniques are applied to group the adjacency matrix into k clusters, where k represents the number of desired features; the most representative feature from each cluster is chosen for further analysis. This approach systematically addresses both linear and nonlinear dependencies, enabling more efficient and accurate feature selection and overcoming limitations in existing methods.

Keywords: Adjacency matrix, Feature selection, Mutual information, High dimension

Mathematics Subject Classification (2010): 68P10, 68T09.

*Corresponding Author: negin.bagherpour@ut.ac.ir

1. Introduction

Feature selection is a critical preprocessing step in high-dimensional data analysis, which aims to identify relevant features and remove irrelevant, redundant, and noisy ones. In the current era of information explosion, diverse datasets such as images, texts, and medical microarrays are generated at an unprecedented speed, leading to challenges such as the curse of dimensionality, overfitting and reduced model interpretability. Feature selection acts as a powerful tool to address these challenges by reducing the dimensionality of the data, preserving the inherent data structure and increasing the performance of the learning algorithm. By carefully selecting the most informative features, this step can significantly improve the accuracy, efficiency and interpretability of machine learning models. Such a comprehensive feature selection algorithm is suitable for a wide range of applications, from image recognition to natural language processing Guyon and Elisseeff (2003), Liu and Yu (2005).

There exist feature selection methods for both supervised and unsupervised data. Supervised methods assess feature association by analyzing feature class correlations, using labeled data to identify features that are most strongly associated with a target variable or class. In contrast, unsupervised techniques evaluate feature importance based on data variance and separability without relying on class labels Dy and Broadley (2004), He *et al.* (2005). Although each of the existing algorithms often focuses on only one of these approaches, the potential benefits of integrating both perspectives to create more robust and effective feature selection strategies are also notable.

Direct processing of high-dimensional data not only increases the computational complexity and memory requirements, but also leads to suboptimal performance due to the presence of irrelevant and noisy features. Studies show that the number of intrinsic features of high-dimensional data are usually much lower than the whole dimension and only a subset of features are truly informative for tasks such as clustering and classification Ding *et al.* (2020). Including the irrelevant features in the dataset might significantly impair the performance of the learning algorithm, leading to overfitting, poor generalization and increased computational costs. Feature selection appears to be critical preprocessing step for high-dimensional data because it facilitates dimension reduction by removing irrelevant and redundant features while preserving essential data features Ding *et al.* (2020).

In this paper, we present a new supervised feature selection method that includes three key steps. The first step involves measuring feature class correlations using the nonparametric mutual information-based (MINE) dependency measurement technique Reshef *et al.* (2014). MINE is a powerful tool for detecting

non-linear dependencies between variables and is suitable for evaluating the association of features in complex datasets. Then the measured correlations are normalized and the features covering a proper portion of the cumulative distribution function are selected. This portion serves as a hyperparameter and is adjustable by the user. In the second step, the pairwise relationships between the features selected from the first step are evaluated to construct an adjacency matrix, which is then merged with the diagonal matrix obtained in the previous step. In this step, almost all interactions and higher-order dependencies between features are captured, and a more comprehensive representation of feature connectivity is provided. Finally, in the third step, a clustering technique is applied to the adjacency matrix, where the number of clusters is equal to the number of selected features. The most representative feature from each cluster is selected as a part of the final feature set, which ensures the proper diversity of selected features.

The remainder of this paper is organized as follows: Section 2 reviews related work on feature selection algorithms. Section 3 presents the proposed feature selection method in detail. Section 4 evaluates the performance of the proposed method on some benchmark datasets and compares it with state-of-the-art algorithms. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Works

Over the last few decades, extensive research has been conducted on feature selection as one of the most important steps in machine learning and data mining. Recently numerous algorithms and techniqueshave been proposed to tackle the challenges of high-dimensional data analysis most of which focus on simple filtering methods that rank features based on their individual relevance to the target variable or class Guyon and Elisseeff (2003), Liu and Yu (2005). These methods are computationally efficient and can handle large-scale datasets, but often fail to capture complex interactions and dependencies between features.

To overcome the limitations of filter methods, wrapper methods were introduced. They use a specific learning algorithm as a black box to evaluate feature subsets and select the most informative ones Kohavi and John (1997). Wrapper methods can capture interactions and dependencies, but are computationally expensive and may be overfitted to the specific learning algorithm used. On the other hand, embedded methods combine the advantages of both filter and wrapper methods by incorporating feature selection as part of the model training process Guyon and Elisseeff (2003), Chandrashekar and Sahin (2014). These methods are more efficient than wrapper methods and can capture feature interactions, but

are often specific to a particular learning algorithm and may not generalize well to other tasks or datasets.

Recently, deep learning-based feature selection methods have attracted attention due to their ability to automatically learn feature representations from raw data Wu (2021). These methods use the powerful feature extraction capabilities of deep neural networks to identify relevant features in an end-to-end manner, without relying on manual feature engineering. However, deep learning-based methods require large amounts of labeled data and computational resources and may not perform well in scenarios with limited data or complex feature interactions. Despite the significant progress made in feature selection research, there are still several challenges and limitations that need to be addressed. One of the most important challenges is the lack of consideration for higher-order feature interactions, as most existing methods only investigate pairwise interactions Zhao *et al.* (2021). Additionally, many feature selection algorithms are sensitive to noise and outliers in the data, which can lead to suboptimal feature subsets Gu *et al.* (2012).

In this paper, our goal is to present a new method for supervised feature selection that combines the strengths of the filter approach and overcomes its weakness of considering only pairwise feature interactions. Our method uses a nonparametric mutual information MINE-based dependency measurement technique to capture nonlinear feature class correlations, and it combines a clustering-based approach to select a diverse set of informative features. By carefully designing the feature selection process and incorporating domain-specific knowledge, we believe that our method can achieve advanced performance on a wide range of high-dimensional data analysis tasks while providing insight into the structure and properties of the underlying data.

3. Proposed Method

In this section, we present our novel supervised feature selection method, which consists of three key steps: feature-class correlation measurement, pairwise feature relationship evaluation, and feature clustering. The method is designed to identify the most relevant features while eliminating irrelevant, redundant, and noisy ones, thereby improving the performance and interpretability of machine learning models.

Step 1: Feature-Class Correlation Measurement
The initial step of our method involves assessing the correlation between each feature and the target class. Partial correlation is a suitable method for revealing the correlation between two variables while holding other variables constant. Con-

sidering each feature's role separately makes partial correlation a reliable benchmark; however, it does not consider nonlinear relations between features. To overcome this issue, mutual information is introduced Reshef *et al.* (2014), using the nonparametric mutual information-based (MINE) dependency measurement technique. MINE is a robust tool for detecting nonlinear dependencies among variables, making it well-suited for evaluating feature associations in complex datasets.

After computing the MINE coefficient for each feature-class pair, we standardize the values using min-max normalization. This normalization process ensures that feature-class correlations are standardized and easily comparable. Subsequently, we select features that encompass a specified percentage, denoted as "P," of the cumulative distribution function of the normalized MINE scores. This parameter, akin to a hyperparameter, allows us to control the balance between the accuracy of feature selection and the number of features chosen. Adjusting the value of "P" enables us to fine-tune the trade-off, facilitating the exclusion of irrelevant features and the identification of the most pertinent features related to the class.

Step 2: Pairwise Feature Relationship Evaluation
In the second step of our method, we evaluate the pairwise relationships between the features selected in the previous step. This approach is motivated by the recognition that feature interactions and dependencies can provide valuable information for feature selection beyond what can be obtained from feature-class correlations alone. To quantify these pairwise relationships, we construct an adjacency matrix denoted as A, where each element $A[i, j]$ represents the MINE statistic between features i and j. Notably, the diagonal elements of the matrix A are augmented by adding the normalized MINE scores calculated in the initial step, which reflect individual feature-class correlations and feature interpretability. By incorporating pairwise feature relationships into the adjacency matrix, we aim to capture complex interactions and higher-order dependencies between features, providing a more comprehensive representation of feature connectivity. This information can be particularly useful in scenarios where features are highly correlated, such as in cases of multicollinearity or feature redundancy.

Step 3: Clustering Features
In the last phase of the method, we use a clustering technique on the adjacency matrix A to find a diverse set of informative features. The number of clusters matches the chosen number of features, represented by k. We employ a common clustering method like k-means or hierarchical clustering to group features into k clusters based on their relationships with each other.

After clustering, we pick the feature closest to the center of each cluster as the cluster's representative feature. This selection ensures that each chosen feature

is closely related to its cluster's features and less related to features outside the cluster. By selecting one feature from each cluster, we avoid redundancy and focus on informative and complementary features.

The result of our method is a set of k features ready for use in various machine learning tasks like classification or regression. The value of k can be adjusted to meet specific needs, such as reducing dimensionality or considering computational constraints. We refer to this algorithm as FSAM standing for Feature Selection by Adjacency Matrix. Its flowchart is presented in Figure 1. We note that one of the most important characteristics of FSAM is the possibility of parallel implementation. More precisely, computing correlations can be done in parallel while the comparison of the weights and selection of the most important features are independent and can be done in parallel as well. This leads to the successful prediction of features in high dimensional datasets. This can be seen in Section 4.

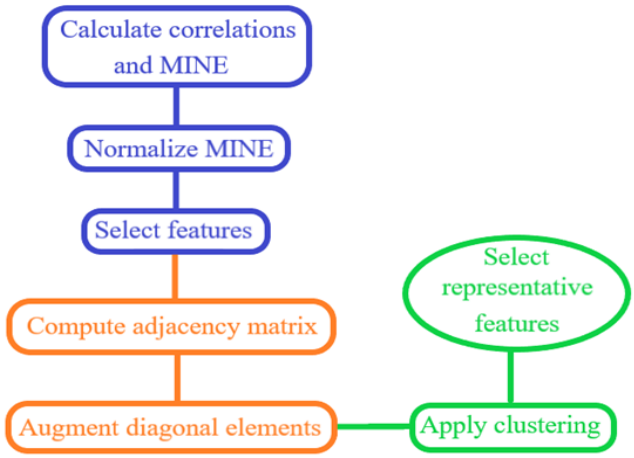

Figure 1: FSAM flowchart

4. Experiments and Results

In this section, we present the results of our experiments on benchmark datasets to evaluate the performance of our proposed feature selection method. We compare our method (FSAM) with state-of-the-art feature selection algorithms and demonstrate its effectiveness in obtaining higher accuracy and efficiency. In fact, FSAM,

Table 1: Number of samples and important features in benchmark datasets

	Number of samples	Number of important features
Optical	5620	64
Cylinder	541	22
Musk	476	166
Mice	180	77

when combined with all learning algorithms**,** including logistic regression, neural networks, XGBoost, support vector machines, and random forests, outperforms other feature selection methods. Figure 2 presents the two best learning algorithms for each of the three well-known datasets:

- Optical Recognition of Handwritten Digits
- Cylinder Bands
- Musk (version 1)
- Mice Protein Expression

The number of features and samples for each dataset is presented in Table 1.

The presented results confirm the efficiency of FSAM in choosing the most important features more accurately. We note that the selected datasets represent different types based on the number of features and samples.

Furthermore, the most important characteristics of these three datasets are the high number of features in Musk (version 1), the low number of features in Cylinder Bands, and the high number of samples in Optical Recognition of Handwritten Digits. This means that our tests cover a variety of dataset types and confirm the efficiency of our algorithm on datasets with both low and high numbers of features as well as a high number of samples. Moreover, as shown in Figures 2(a)–2(d), FSAM outperforms other methods with a considerable difference in accuracy. However, in Figures 2(e) and 2(f), while all methods achieve very high accuracy, FSAM still outperforms others.

A well-known benchmark for investigating the efficiency of feature selection algorithms is F_{score} which is defined as Ting (2011):

$$\text{recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}},$$

$$\text{precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}},$$

$$F_{\text{score}} = 2 \ \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

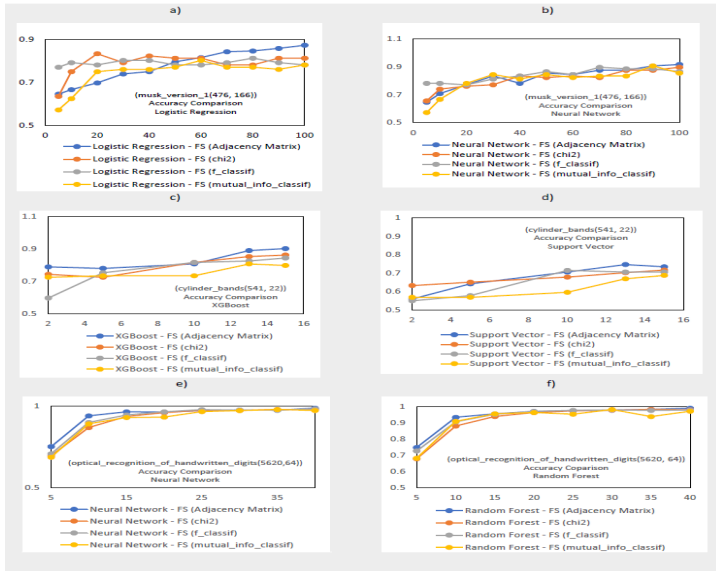

Figure 2: Comparison of FSAM to other FS methods

In Table 2 the F_scores for benchmark datasets are reported. In each dataset, the average results for different classification strategies, including logistic regression, neural networks, support vector machines, and random forest are reported.

Table 2: Comparison on F_scores across benchmark datasets

Dataset	Method	F_{score}	Dataset	Method	F_{score}
Optical	Adjacency	**0.9823**	Cylinder	Adjacency	**0.8858**
	Chi2	0.9496		Chi2	0.8764
	F_classif	0.9647		F_classif	0.8532
	Mutual	0.9388		Mutual	0.8661
Dataset	Method	F_{score}	Dataset	Method	F_{score}
Musk	Adjacency	**0.9272**	Mice	Adjacency	0.8531
	Chi2	0.8965		Chi2	0.8346
	F_classif	0.8749		F_classif	**0.8671**
	Mutual	0.8834		Mutual	0.8478

This evaluation confirms the efficiency of our proposed algorithm in selecting the best features with higher precision and recall on test datasets.

5. Concluding Remarks

In this study, we introduced a new supervised feature selection approach that utilizes the strengths of the filter method while addressing its limitations, particularly its focus on pairwise feature interactions. Our method uses a nonparametric technique (MINE) to capture nonlinear correlations between features and incorporates a clustering-based strategy to select a diverse set of informative features. Our experiments on standard datasets demonstrate the effectiveness of our approach in increasing model accuracy and efficiency compared to existing feature selection techniques.

There are several potential research avenues arising from our work. First, we plan to extend our method to accommodate unsupervised, semi-supervised, and weakly supervised scenarios by incorporating domain-specific knowledge to guide feature selection. Second, we intend to investigate the interpretability of our method, especially in the context of complex models such as deep neural networks. By elucidating the data structure and feature importance, we believe our approach can contribute to the development of transparent and reliable machine learning systems.

Finally, we are eager to explore the practical application of our feature selection method in real-world domains such as healthcare, finance, and environmental science. By collaborating with domain experts and incorporating their expertise into the feature selection process, we anticipate tailoring our method to specific applications and providing valuable insights for decision -making and problem -solving.

Acknowledgment

We acknowledge the referees for their time and efforts to improve our research.

References

Chandrashekar, G., Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, **40(1)**, 16-28.

Ding, D., Yang, X., Xia, F., Ma, T., Liu, H., and Tang, C. (2020). Unsupervised feature selection via adaptive hypergraph regularized latent representation learning. *Neurocomputing*, **378**, 79-97.

Dy, J. G., Broadley, S. A. (2004). Feature selection for unsupervised learning. *Journal of machine learning research*, **5**, 845-889.

Gu, Q., Li, Z., Han, J. (2012). Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725.*

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, **3**, 1157-1182.

He, H., Bai, Y., Garcia, E. A., Li, S. (2005). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322-1328, IEEE.

Kohavi, R., John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, **97(1-2)**, 273-324.

Liu, H., Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, **17(4)**, 491-502.

Reshef, D. N., Reshef, Y. A., Mitzenmacher, M., Sabeti, P. C. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, **111(9)**, 3354-3359.

Ting, K.M. (2011). Precision and Recall. *In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning*. Springer, Boston, MA, DOI: 10.1007/978-0-387-30164-8_652.

Wu, H.(2021) A Deep Learning-Based Hybrid Feature Selection Approach for Cancer Diagnosis. *Journal of Physics*, Conference Series, 1848.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H. (2021). Advancing feature selection research. *ASU Feature Selection Repository*, **4**.