Feature Selection Methods for Data Dimensionality Reduction

Sara Dehghani [1], Razieh Malekhosseini [1*], Karamollah Bagherifard[1], Seyed Hadi Yaghoubyan[1]

[1] Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran

*** Corresponding author email address**: malekhoseini.r@iau.ac.ir

Received: 2025-09-02 **Revised:** 2026-01-01 **Accepted:** 2026-01-05 **Initial Publish:** 2026-02-15 **Final Publish:** 2026-07-01

Abstract

Despite creating opportunities, platforms with large-scale data also pose significant computational challenges. An issue with high-dimensional data is that in many cases, not all of the data's features are important or vital for uncovering the knowledge hidden within it. Due to this, reducing the dimensionality of data remains a significant topic in many areas of data mining. Using feature selection techniques is one effective method for reducing the dimensionality of data. During the process of feature selection, a subset of the original features is selected by eliminating irrelevant and redundant features. This article analyzes and categorizes different feature selection techniques from different perspectives. After that, it provides an overview of data clustering concepts and categorizes different clustering algorithms. This article also investigates the use of optimization algorithms in feature selection methods and presents methods based on this approach. Next, this article compares and analyzes feature selection methods, emphasizing their strengths and weaknesses.

Keywords: *Big data, Feature clustering, Feature selection, Optimization algorithms, classification.*

How to cite this article:

Dehghani, S., Malekhosseini, R., Bagherifard, K., & Yaghoubyan, S. H. (2026). Feature Selection Methods for Data Dimensionality Reduction. Management Strategies and Engineering Sciences, 8(4), 1-11.

1. Introduction

Recent decades have seen significant growth in large-scale datasets due to the rapid development of computers and information technologies. Simultaneously, the demand for high-speed, accurate applications that rely on large-scale datasets has surged. Data mining links artificial intelligence, machine learning, statistics, and databases to analyze and process vast amounts of data [1, 2]. Data mining aims to extract knowledge from datasets and present it in a structured format that can be readily comprehended and utilized for future applications. A significant challenge for data mining tasks like identifying patterns is when datasets have high dimensionality. This occurs when the number of features or variables in the dataset is much larger than the number of patterns or observations available. High-dimensional datasets can impair classifier performance in two ways. Increasing data dimensions lead to greater computational demands.

Additionally, models constructed with high-dimensional data have inferior generalization capabilities and are more prone to overfitting [3-5]. Reducing problem dimensions can enhance classification algorithm performance while reducing computational complexity. Numerous feature selection methods rely on heuristic and evolutionary approaches to avoid an upsurge in computational complexity. Feature clustering is an effective method for reducing dataset dimensionality, where the initial features are divided into clusters, and a selection of features is chosen from each cluster. Performing clustering ensures that the characteristics present within a cluster are alike, but they differ from the traits found in different clusters.

Subsequent sections will present techniques for dimensionality reduction and classify various feature selection methods based on multiple criteria. The article will introduce dimension reduction methods, categorize and define various feature selection methods, explain the methods based on each solution, and compare them, highlighting their respective pros and cons.

2. Definitions

Numerous features commonly characterize data. Some data features may be unimportant or noisy for the intended

data mining application. Irrelevant and redundant features in a dataset can reduce the performance of machine learning algorithms and increase computational complexity [3, 5, 6]. Reducing dimensionality is a crucial activity in machine learning and data mining tasks.

3. Data Dimensionality Reduction Techniques

Data dimension reduction methods can be categorized into two groups:

- Feature extraction methods: They extract features from a multi-dimensional space and map them to a lower-dimensional space. Two categories of feature extraction methods exist: linear and nonlinear.
- Feature selection methods: Aim to decrease data dimensionality by selecting a subset of the original features.

The process of feature selection aims to choose a set of features from the initial ones and discard those that are irrelevant or duplicative.

Figure 1 depicts the contrast between feature selection and feature extraction techniques. Figure 1(a) demonstrates how feature selection involves choosing a subset of initial features. In contrast, Figure 1(b) demonstrates the creation of a new set of features through feature extraction. Here, n represents the initial features and m represents the reduced features $m < n$.

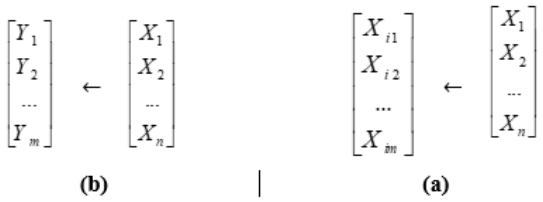

Figure 1. Shows the contrast between feature selection and feature extraction methods [7]

4. Feature Selection Methods using Search Strategies

Every feature selection method comprises two primary stages: creating candidate subsets and assessing those subsets. Various subsets are created by applying different search strategies, and their usefulness is assessed based on a specific criterion. The two stages are iterated until the stopping condition is reached. There are five categories of feature selection methods that can be classified according to their search strategy [8, 9].

1. Forward Selection: A feature is added greedily to an initially empty feature set in each iteration
2. Backward Elimination: It begins with all features included and removes one feature at a time in each iteration.
3. Stepwise Forward Selection: Each iteration adds or removes a feature greedily, starting from an initial set of features.
4. Stepwise Backward Elimination: It begins with all the features and, in each step, one feature is added or removed greedily.
5. Random Mutation: It begins with a random feature set, and a feature is randomly included or excluded in each step.

1. Figure 2 illustrates that the feature selection process comprises of four key stages: generating subsets, evaluating subsets, determining when to stop, and validating the final results.

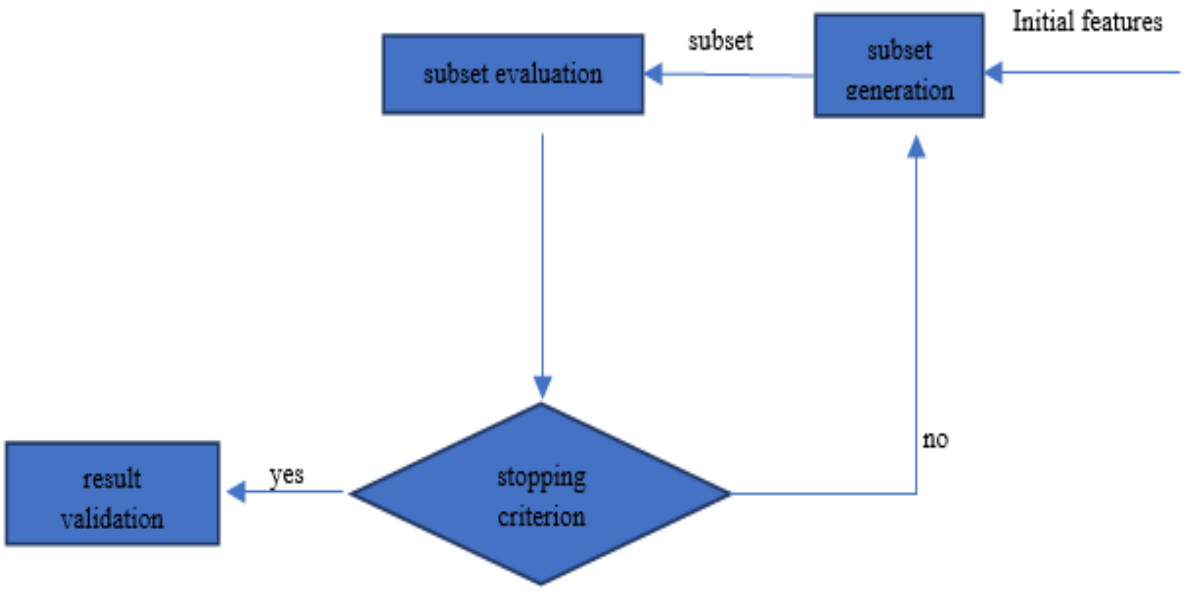

Figure 2. Stages of Feature Selection Process [10]

In Figure 2, it is shown that in every search process iteration, a new subset of the primary features is formed. This subset's fitness is evaluated using a particular criterion. The creation and assessment of subsets are continued repeatedly until a pre-defined endpoint is reached. Upon process completion, the top feature subset is picked and verified on the test dataset.

As shown in Figure 3, in this article we try to examine the feature selection methods from three different perspectives.

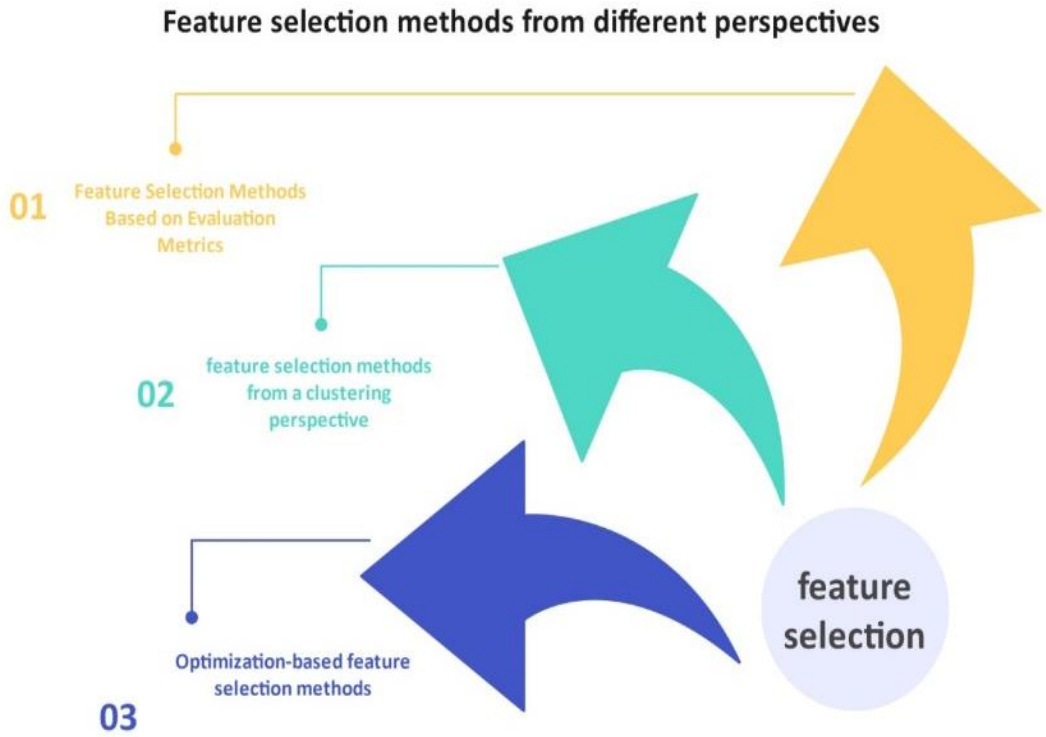

Figure 3. Feature selection methods from different perspectives

5. Feature Selection Methods Classification Based on Evaluation Metrics

There are two categories of feature selection methods: feature ranking and subset selection, which are classified based on how features are evaluated [11]. Feature ranking assigns a score to each feature based on a certain criterion, removing insufficiently scored ones. Subset selection methods search through the set of potential subsets to find the best one. The optimal subset is found by searching through all possible feature subsets, with a size of 2^n where n is the number of initial features. Feature selection methods shown in Figure 4 are typically classified into four main types based on the evaluation criterion: Filter, wrapper, embedded, and hybrid [4, 12].

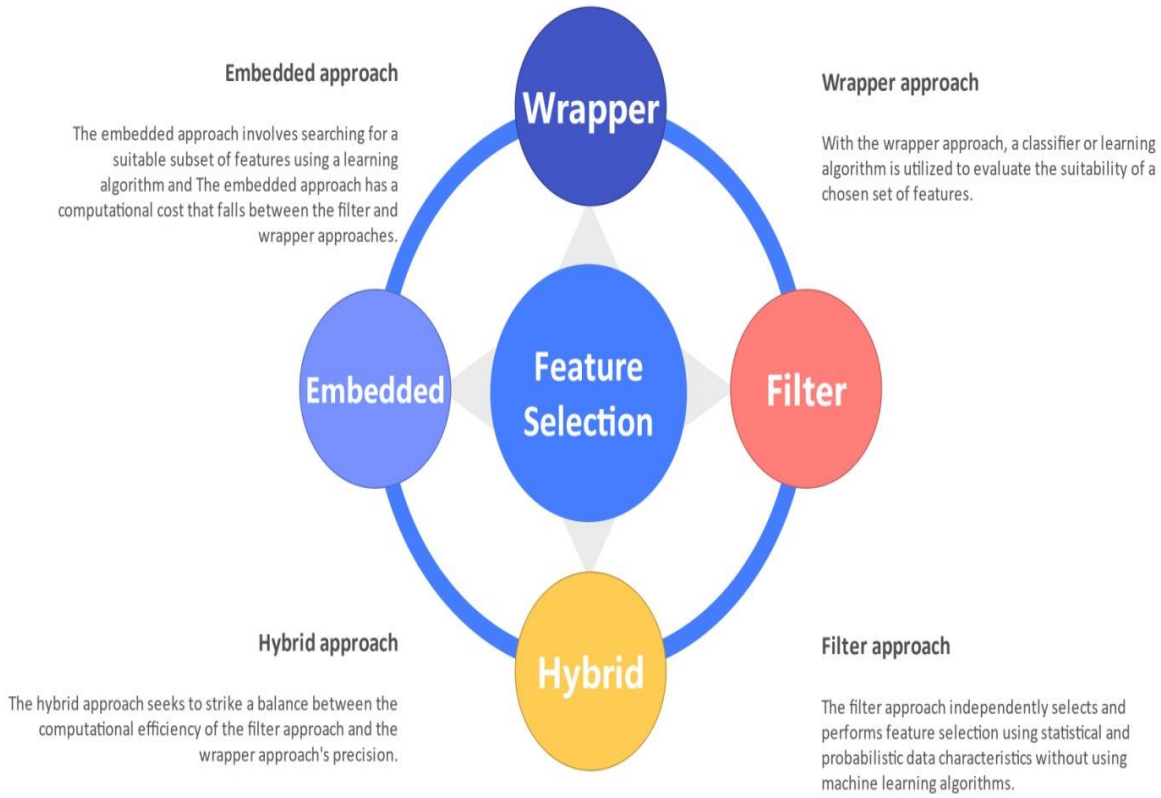

Figure 4. Feature Selection Methods Based on Evaluation Metrics

6. Wrapper approach

With the wrapper approach, a classifier or learning algorithm is utilized to evaluate the suitability of a chosen set of features. In this approach, a search technique is used to discover the most favorable set of features. A classifier is trained and tested to evaluate the quality of a generated feature subset at each step of the search process. The best feature subset is ultimately chosen as the final subset.

7. Filter approach

The filter approach independently selects and performs feature selection using statistical and probabilistic data characteristics without using machine learning algorithms. In other words, this approach employs the intrinsic data properties to evaluate features. The filter approach, which does not employ machine learning algorithms, is computationally faster than wrapper-based approaches and is ideal for high-dimensional datasets. Figure 5 displays the general scheme of this approach, which resembles wrapper-based methods but evaluates different subsets based on an independent criterion instead of assessing the generated feature subsets from the learning algorithm.

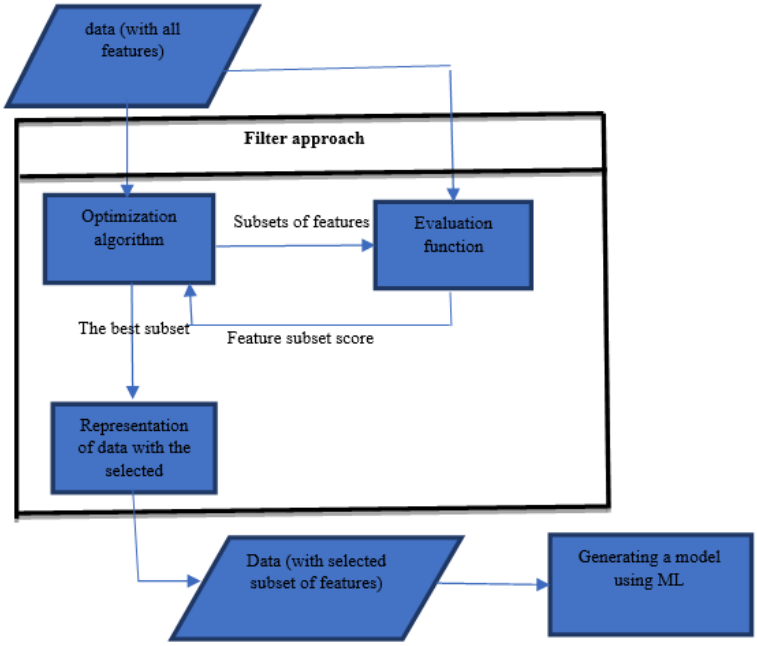

Figure 5. Overview of the filter approach

8. Hybrid approach

The hybrid approach seeks to strike a balance between the computational efficiency of the filter approach and the wrapper approach's precision. This is achieved by employing a proposed algorithm. The aim is to develop a method that is both efficient and effective. Many hybrid feature selection methods perform the feature selection process in two stages. Many hybrid feature selection methods perform the feature selection process in two stages. The hybrid approach often involves a two-stage feature selection process. Firstly, the filter approach is used to decrease the initial feature set, and then the wrapper approach is applied to select the final feature set from the reduced feature set.

9. Embedded approach

The embedded approach incorporates feature selection as an integral component of the learning algorithm. The embedded approach involves searching for a suitable subset of features using a learning algorithm. The embedded approach has a computational cost that falls between the filter and wrapper approaches. As previously stated, the wrapper approach evaluates each candidate subset using the classification accuracy of a preselected learning algorithm. Wrapper-based methods are computationally complex, which is a significant issue. The embedded approach aims to save computation time by integrating feature selection into the training process. The embedded approach, similar to the wrapper approach, relies on the learning algorithm used for feature selection.

10. *Categorizing* feature selection methods from a clustering perspective

There are two main types of feature selection methods: supervised and unsupervised [10, 12, 13]. Supervised feature selection methods work with labeled training data, where each example has a feature vector and a corresponding class label. In contrast, unsupervised feature selection methods operate on unlabeled data. Supervised feature selection methods are considered more reliable and perform better than unsupervised methods due to the use of class labels [12, 14]. Unsupervised feature selection is a challenging area that has received considerable attention in many studies.

11. Feature selection methods based on the clustering of features

There are four categories of feature selection methods: wrapper, filter, objective function optimization, and feature clustering [15-17]. To apply these solutions in feature selection, two important factors need to be considered. A similarity metric needs to be introduced to measure feature similarity. Secondly, it is necessary to specify a clustering algorithm to use these solutions in feature selection. One effective solution for reducing dataset dimensions is feature

clustering, where initial features are clustered, and a number of features are then selected from each cluster. Features are clustered based on their similarity, and dissimilarity between features in different clusters. Datasets with very high dimensions are often encountered in applications like text classification and bioinformatics. Two popular datasets used for text classification, 20 Newsgroups and Reuters21578, comprise more than 15,000 features each. High-dimensional datasets can pose a significant challenge to classification. Clustering the features can be useful in eliminating redundancies in original features and providing preliminary analysis, particularly in datasets with high dimensions [12].

12. Optimization-based feature selection methods

Studies indicate that identifying the perfect subset is a computationally complex problem [10, 12, 15]. The straightforward method to find the best subset involves examining all feasible subsets through an exhaustive search. As evaluating all possible subsets is inefficient, we require a computationally feasible solution with adequate usefulness.

Various search algorithms have been proposed for the feature selection problem to find a globally optimal solution within a feasible time frame. Researchers have shifted their focus towards heuristic and metaheuristic algorithms. Heuristic search methods provide faster algorithms that balance computational complexity with solution quality. These methods can find the solution in an acceptable time, but they can't ensure the discovery of the global optimal solution. Various algorithms with different approaches have aimed to address this problem of finding the optimal global solution, which is the best subset of main features. These algorithms explore the problem space and prioritize good solutions in order to find the optimal solution. Metaheuristic algorithms have effectively decreased the chances of being trapped in a local optimum by adopting this approach. Population-based optimization algorithms, including Ant Colony Optimization (ACO) [18], Genetic Algorithm (GA) [19], and Particle Swarm Optimization (PSO) [20], have been extensively studied in the context of feature selection using metaheuristic approaches.

13. Feature selection methods based on evaluation metrics

IG is a machine learning method that is widely used and based on information theory [21]. IG refers to the quantity of information that a feature contributes to a classification system. Equation (1) is used to determine the IG of feature A in relation to a pattern set S.

$$\mathrm{IG(S|A)} = \mathrm{E(S)} - \sum_{\mathrm{v}\in\mathrm{Values(A)}} \mathrm{P_v} \,.\, \mathrm{E(S_v)} \qquad (1)$$

The set of all feasible values for feature A is represented by Values (A). P_v denotes the possibility of patterns in S that have value v for feature S, and S_v is the subset of patterns in S that have the value v for feature A. $E(S)$ measures the disorder of the pattern set S. Equation (2) defines the entropy of variable X.

$$\mathrm{E(X)} = \sum_{\mathrm{v}\in\mathrm{Values(X)}} - \mathrm{P_v} \log_2(\mathrm{P_v}) \qquad (2)$$

P_v denotes the likelihood of patterns in S having a value of v for variable X. Features with high values are often selected by the information gain (IG) method. Features selected based on high information gained on the training data may not have strong predictive power on the test data. The gain ratio (GR) [22] and symmetric uncertainty (SU) [23] were proposed to resolve this problem.

GR is an effective measure in feature selection. The gain ratio measure represents the degree and uniformity with which a feature splits the data patterns. Equation (3) defines the gain ratio measure.

$$\mathrm{GainRatio(S,\ A)} \equiv \frac{\mathrm{IG(S|A)}}{\mathrm{E(A)}} \qquad (3)$$

$IG(S|A)$ denotes the information gain of feature A, and $E(A)$ denotes the entropy of feature A. The highest gain ratio value corresponds to the best rank in this method.

Symmetrical Uncertainty (SU) is an evaluation metric that addresses the bias toward selecting features with high information gain and scales the value between zero and one. Equation (4) is used to calculate the SU of feature A.

$$\mathrm{SU(S,\ A)} \equiv 2\left[\frac{\mathrm{IG(S|A)}}{\mathrm{E(S)+E(A)}}\right] \qquad (4)$$

The entropy of set S is denoted by $E(S)$ and the entropy of feature A is denoted by $E(A)$. When $SU = 0$, set S and feature A are completely independent, while $SU = 1$ indicates complete dependence between them. Features are chosen for this technique based on their strongest association with set S.

The Gini index (GI) [21] is a method for splitting based on impurity that works better with continuous values. Equation (5) can be used to compute the Gini index for a set of patterns S.

$$\text{GiniIndex(S, A)} \equiv \text{Gini(S)} - \sum_{v \in \text{Values(A)}} \text{P}_v.\ \text{Gini(S}_v\text{)} \quad (5)$$

Equation (6) can be used to compute $Gini(S)$.

$$\text{Gini(S)} = 1 - \sum_{v \in \text{Values(S)}} (\text{P}_v)^2 \quad (6)$$

The Gini index will have its maximum value if the subsets generated by dividing patterns in S based on feature A belong to only one class. Any feature with the highest Gini index value is regarded as appropriate.

The Fisher score (FS) [24] is a feature selection method that is supervised and aims to minimize the distance between patterns in the same category and maximize the difference between patterns in different classes. This metric calculates the ratio of the dispersion of patterns among various categories to the dispersion of patterns within each category. Features showing such distinguishing characteristics are given a higher score by this measure. Equation (7) is utilized to compute the Fisher score of feature A.

$$FS(S,A) = \frac{\sum_{v \in V\text{alues }(S)} n_v(\bar{A}_v - \bar{A})^2}{\sum_{v \in V\text{alues }(S)} n_v(\sigma_v(A))^2} \quad (7)$$

The mean and standard deviation of patterns within class V concerning feature A are represented by $\sigma_v(A)$ and $\overline{A}_v$, respectively. n_v is the number of patterns with class label V while $\overline{A}$ refers to the overall average of the pattern set linked to feature A. The final subset of features consists of those with the highest Fisher Score after computing this measure for all the features.

The term variance (TV) [25] is the simplest unsupervised measure for evaluating features. The representation of a feature is considered strong if it has a high variance. Features having high Term variance are regarded as significant sources of information. Equation (8) defines the measure of term variance.

$$\text{TV(S, A)} = \frac{1}{|S|}\sum_{j=1}^{|S|} (\text{A(j)} - \bar{\text{A}})^2 \quad (8)$$

$|S|$ shows the total number of patterns, while $A(j)$ represents the value of feature A in the j-th pattern.

Laplacian Score (LS) [26] is a graph-based technique that can be utilized for feature selection in both supervised and unsupervised scenarios. The LS score views the data space as a graph and assumes that if two data points are in proximity to one another, they likely belong to the same category. Feature selection is conducted using the local structure of the data space in this technique. Equation (9) is utilized to compute the LS score of feature A for a given set of patterns S.

$$\text{LS(S, A)} = \frac{\sum_{i,j}(\text{A(i)} - \text{A(j)})\ \text{S}_{ij}}{\sum_i (\text{A(i)} - \bar{\text{A}})\text{D}_{ii}} \quad (9)$$

$A(i)$ represents the value of feature A in the i-th pattern, while $\bar{A}$ is the mean of feature A. The proximity relationship between patterns, which is calculated using equation (10), is represented by S_{ij}, and D is a diagonal matrix where $\sum_j S_{ij}$.

$$\text{S}_{ij} = \begin{cases} e^{\frac{x_i - x_j}{t}}, & \text{if } x_i \text{ and } x_j \text{ are neighbors} \\ 0, & \text{Otherwise} \end{cases} \quad (10)$$

Two adjacent patterns, x_i and x_j, are deemed neighbors if one of them is among the K-nearest neighbors of the other. In addition, t is a constant coefficient in this case.

The mRMR method [27] is a popular filter-based multivariate feature selection technique. In the selection process of features, this technique considers both the relevance and redundancy of each feature. The aim of the mRMR criterion is to eliminate duplicate information among features by using their mutual information. The maximum relevance criterion assesses the appropriateness of features by computing the mutual information between features and the target class. By utilizing incremental search methods, a subset of features that is nearly optimal can be found using the combination of these two criteria, known as mRMR. The objective of the optimization algorithm is to find a subset of features that is nearly optimal by using the formula (11) and performing an incremental search when $\tilde{A}_m$ represents a set of m selected features using the mRMR method.

$$\max_{A_j \in \tilde{A}_n - \tilde{A}_{m-1}} \left[MI(A_j; C) - \frac{1}{m-1} \sum_{A_i \in A_{m-1}} MI(A_j; A_i) \right] \quad (11)$$

$\tilde{A}_n$ denotes the complete set of features containing n features in this equation. C denotes the target class, while A_j represents the j -th feature in the feature set. $MI(A_j; C)$ and $MI(A_j; A_i)$ represent the mutual information between the j-th feature and the i-th feature, respectively.

Equation (12) is used to calculate the mutual information between two random variables X and Y.

$$MI(X; Y) = \sum_{x \in X,} \sum_{y \in Y} P_{x,\,y} \log \frac{P_{x,y}}{P_x . P_y} \quad (12)$$

P_x represents the probability of variable X taking on the value X, while $P_{x,y}$ represents the joint probability of variable X taking on the value x and variable Y taking on the value y.

The RRFS method [28] is a new filter-based feature selection technique that evaluates the relevance and suitability of features based on their importance and redundancy. It can perform feature selection in both supervised and unsupervised modes. Features are initially sorted by a specific evaluation criterion to determine their appropriateness, and the top-ranking one is picked as the initial feature in this approach. In every iteration of the algorithm, one feature is chosen. If its similarity to the recently chosen feature is lower than a particular threshold, these steps are repeated until the required number of features is reached. Equation (13) is employed to calculate the mean absolute error, which is the evaluation criterion for suitability in the unsupervised mode.

$$MAD(S, A) = \sum_{j=1}^{|S|} |A(j) - \bar{A}| \quad (13)$$

The Fisher score obtained from equation (7) is the suitable criterion for the supervised case.

MC [29] is a commonly used method for measuring similarity between two features. Equation (14) yields the mutual correlation between two features, A_i and A_j.

$$MC(A_i, A_j) = \frac{\sum_{k-1}^{|S|} A_i(k) A_j(k) - |S| \bar{A}_i \bar{A}_j}{\sqrt{\left(\sum_{k-1}^{|S|} A_i(k)^2 - |S| \bar{A}_i^2\right)\left(\sum_{k-1}^{|S|} A_j(k)^2 - |S| \bar{A}_j^2\right)}} \quad (14)$$

$A_i(k)$ and $A_j(k)$ represent the values of i-th and j-th features in k-th pattern, while $\overline{A_i}$ and $\overline{A_j}$ are the mean values for the set of patterns corresponding to A_i and A_j respectively. A total dependence between the two features is indicated by $MC(A_i, A_j) = \pm 1$. $MC(A_i, A_j) = 0$ implies that A_i and A_j are independent of one another. Haindl et al. introduced a mutual-dependency-based technique for selecting features. Their approach involves calculating the mean mutual dependency for each feature, and subsequently removing the one with the highest value in each iteration. The process is repeated until a stopping condition is reached, and the remaining features are regarded as the final subset.

Table 1 presents a comparison of previous methods, along with their respective benefits and drawbacks.

Table 1. Comparison of Feature Selection Methods: Evaluating Advantages and Disadvantages in Classification

Feature Selection methods		Method	Advantages	Disadvantages
Filter	single-variable	Fisher Score (FS) (Gu et al., 2012)	Fast, classifier-independent.	Disregarding feature relationships, low profitability.
		Laplacian score (LS) (He et al., 2005)	Quick, classifier-agnostic, and able to select features without supervision.	Can pick alike and duplicate features, but performance is somewhat low.
		Information Gain (IG) (Raileanu & Stoffel, 2004)	Classifier-agnostic with effective removal of irrelevant features.	Limited to supervised learning feature selection and has high computational complexity.
		Term Variance (TV) (Theodoridis et al., 2010)	Quick with unsupervised feature selection ability.	Potential to choose similar and duplicate features, but performance is somewhat weak.
		Gain Ratio (GR) (Mitchell, 1997)	Quick, classifier-agnostic, and efficient with statistical and mathematical techniques.	Not considering feature relationships leads to greater computational complexity as compared to other single-variable approaches.

	multiple-variable	Minimum Redundancy Maximum Relevance (mRMR) method (Peng et al., 2005)	Considering feature relationships and not dependent on any specific classifier.	Greater computational complexity than single-variable approaches.
		Random Redundancy Feature Selection (RRFS) (Ferreira & Figueiredo, 2012)	Able to eliminate irrelevant and redundant features at the same time by considering feature relationships.	The two-stage process allows for the removal of some features in the first stage, which may result in reduced accuracy.
		Random Subspace Method (RSM) (Lai et al., 2006)	Achieves high performance by considering feature similarity.	This approach focuses on reducing feature similarity, which may lead to the selection of less related features and a subsequent reduction in accuracy.
		feature selection based on genetic algorithm method (GAFS) (Wang et al., 2018)	Introducing a novel approach of multiple populations in the genetic algorithm has enhanced the effectiveness of this technique relative to prior methods.	The algorithm's convergence rate will be notably slower in datasets with high dimensions.
Wrapper	Greedy	The merging of spiral-shaped mechanism with particle swarm optimization method for selecting features is delineated in the journal article "Expert Systems with Applications" (Chen et al., 2013)	Choosing the most informative features for classification problem-solving.	A potential for becoming trapped in local optima.
Hybrid		LMFS, A feature selection method that integrates filter and wrapper approaches (Zhang et al., 2014)	This method proposes a new evaluation function that uses statistical methods to calculate feature similarity and classification accuracy to measure the relationship between the selected subset and the target class. The resulting feature subset will have minimal similarity and maximum association with the target class.	This method has a higher computational complexity than filter and wrapper methods and relies on the classifier.
Embedded		Embedded SVM-based method (Guyon et al., 2002)	It has lower computational complexity than wrapper methods.	It has higher computational complexity than filter methods and relies on the SVM classifier.
		Using an embedded method based on Laplacian criteration (Y. Zhang et al., 2014)	This method can perform feature selection in unsupervised scenarios and has higher accuracy than conventional filter methods.	The computational complexity of this method is high due to the use of hierarchical clustering in similarity calculation and it is somewhat dependent on the classifier.

Table 2 presents the classification methods from the perspective of clustering.

Table 2. Classification of feature selection methods based on clustering

Method	Authors	Feature selection methods
Embedded SVM-based method	Guyon et al, 2002	Supervised
the decision tree-based solution for feature selection in detecting machine rotation errors	Sugumaran et al, 2007	
Fisher Score	Gu et al, 2012	
an incremental feature selection algorithm together with a Naive Bayes classifier	Bermejo et al, 2014	
a decision tree and the particle swarm optimization algorithm to detect spam as an illustration	Zhang et al, 2014	
a straightforward technique, which optimizes both feature selection and SVM parameter optimization at the same time	Faris et al, 2018	
a two-layer feature selection technique that utilizes both wrapper and embedded methods to generate a suitable subset of predictors	Amini & Hu, 2021	
Distribution-based clustering	Pereira et al, 1994	Unsupervised
Using distribution-based clustering for document classification	Baker et al, 1998	
A two-stage method based on the KNN principle	Mitra et al, 2002	

Incremental clustering method for feature clustering	Jiang et al, 2010
New clustering method with automatic cluster number determination	Cheung and jia, 2012
Hypergraph clustering-based method	Zhang et al, 2012
Unsupervised feature selection based on graph theory	Mandal et al, 2013
Combining feature clustering and rough set theory	Pacheco et al. 2017

Table 3 shows different feature selection methods from the perspective of the optimization algorithm used.

Table 3. Classification of feature selection methods based on optimization algorithm

Authors	ABC	PSO	GA	GWO	WOA
Sharawi et al, 2017					
Wang et al, 2018					
Xie et al, 2019					
Vijayanand & Devaraj, 2020					
Duarte & de Carvalho, 2020					
Abdel-Basset et al, 2020					
Al-Tashi et al, 2020					
Mafarja et al, 2020					
Rostami et al, 2021					
Rashno et al, 2022					
Thaher et al, 2022					
Song et al, 2022					
Zhou & Hua, 2022					
Shreem et al, 2022					
Riyahi et al, 2022					
Zhong et al, 2023					

14. Conclusion

In data mining, numerous datasets have a small number of patterns compared to a large number of features. In numerous cases, the performance of a classification system can be negatively impacted by irrelevant and redundant features. Feature selection is a crucial technique to address this issue. The article begins by discussing the classification and examination of feature selection methods. Four categories of feature selection methods were identified based on their evaluation criteria: embedded, wrapper, filter, and hybrid methods. In the wrapper approach, a learning algorithm is used to assess created subsets, whereas filter-based methods use general characteristics of features and statistical analysis to evaluate potential subsets. The feature selection process in this approach does not involve the use of a learning algorithm. Datasets often have too many irrelevant and duplicate features, which can damage classification performance. Feature selection is a crucial method for addressing this issue. Filter-based methods are simple and computationally efficient, while wrapper-based methods have higher accuracy because of their use of learning algorithms. Hybrid methods aim to blend the benefits of wrapper approaches and filter to create a feature selection technique that is both efficient and yields high-quality feature subsets. Embedded methods, a novel approach for feature selection, have been developed in recent times. Embedded methods appoint a learning algorithm to perform an optimization search for the best feature subset. To avoid increasing computational complexity, many feature selection techniques leverage evolutionary and metaheuristic algorithms.

Continuing the discussion on data clustering, we have covered the classification of clustering algorithms. As previously stated, clustering involves organizing data according to their similarities. Feature clustering reduces dataset dimensions by grouping important features into clusters and selecting a few from each cluster. Features within a cluster are similar to one another, while features in different clusters are dissimilar.

Authors' Contributions

Authors equally contributed to this article.

Acknowledgments

Authors thank all participants who participate in this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

References

[1] M. Alirezaei, S. T. A. Niaki, and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines," *Expert Systems with Applications,* vol. 127, pp. 47-57, 2019.

[2] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Systems,* vol. 166, pp. 198-206, 2019.

[3] Y. Liu and Y. F. Zheng, "FS_SFS: A novel feature selection method for support vector machines," *Pattern recognition,* vol. 39, no. 7, pp. 1333-1345, 2006.

[4] J. M. Cadenas, M. C. Garrido, and R. MartíNez, "Feature subset selection filter–wrapper based on low quality data," *Expert systems with applications,* vol. 40, no. 16, pp. 6241-6252, 2013.

[5] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen, and X. Liu, "Feature evaluation and selection with cooperative game theory," *Pattern recognition,* vol. 45, no. 8, pp. 2992-3002, 2012.

[6] S. M. H. Fard, A. Hamzeh, and S. Hashemi, "Using reinforcement learning to find an optimal set of features," *Computers & Mathematics with Applications,* vol. 66, no. 10, pp. 1892-1904, 2013.

[7] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems,* vol. 24, no. 7, pp. 1024-1032, 2011.

[8] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert systems with applications,* vol. 36, no. 3, pp. 6843-6853, 2009.

[9] D. Mladenić, "Feature selection for dimensionality reduction," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, 2005: Springer, pp. 84-102.

[10] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering,* vol. 17, no. 4, pp. 491-502, 2005.

[11] B. Chen, L. Chen, and Y. Chen, "Efficient ant colony optimization for image feature selection," *Signal processing,* vol. 93, no. 6, pp. 1566-1576, 2013.

[12] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[13] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[14] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Transactions on Multimedia,* vol. 15, no. 3, pp. 661-669, 2012.

[15] I. Guyon, "A practical guide to model selection," *Proc. Mach. Learn. Summer School Springer Text Stat,* pp. 1-37, 2009.

[16] X. Zhao, W. Deng, and Y. Shi, "Feature selection with attributes clustering by maximal information coefficient," *Procedia Computer Science,* vol. 17, pp. 70-79, 2013.

[17] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition,* vol. 42, no. 7, pp. 1330-1339, 2009.

[18] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," *IEEE transactions on systems, man, and cybernetics, part b (cybernetics),* vol. 26, no. 1, pp. 29-41, 1996.

[19] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[20] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, 1995, vol. 4: IEEE, pp. 1942-1948.

[21] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence,* vol. 41, pp. 77-93, 2004.

[22] T. M. Mitchell, "Machine learning," ed: McGraw-hill, 1997.

[23] J. Biesiada and W. Duch, "Feature selection for high-dimensional data—a Pearson redundancy based filter," in *Computer recognition systems 2*: Springer, 2008, pp. 242-249.

[24] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725,* 2012.

[25] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to pattern recognition: a matlab approach*. Academic Press, 2010.

[26] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in neural information processing systems,* vol. 18, 2005.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[28] A. J. Ferreira and M. A. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition,* vol. 45, no. 9, pp. 3048-3060, 2012.

[29] M. Haindl, P. Somol, D. Ververidis, and C. Kotropoulos, "Feature selection based on mutual correlation," in *Progress in Pattern Recognition, Image Analysis and Applications: 11th Iberoamerican Congress in Pattern Recognition, CIARP 2006 Cancun, Mexico, November 14-17, 2006 Proceedings 11*, 2006: Springer, pp. 569-577.