Information gain-based multi-objective evolutionary algorithm for feature selection Research Technical Report

Chen haitao

2025.1.10

Abstract

Feature selection (FS) has garnered significant attention because of its pivotal role in enhancing the efficiency and effectiveness of various machine learning and data mining algorithms. Concurrently, multi-objective feature selection (MOFS) algorithms strive to balance the complexity of multiple optimization objectives during the FS process. These include minimizing the number of selected features while maximizing classification performance. Nonetheless, managing the complexity of feature combinations presents a formidable challenge, particularly in high dimensional datasets. Evolutionary algorithms (EAs) are increasingly adopted in MOFS owing to their exceptional global search capabilities and robustness. Despite their strengths, EAs face difficulties in navigating expansive solution spaces and achieving a balance between exploration and exploitation. To address these challenges, IGEA is designated, a novel information gain based EA for MOFS, designated as IGEA. This approach utilizes a clustering method for selecting adverse parent population, thereby enhancing individual variability and maintaining a high quality population. Considerably, IGEA employs information gain as a metric to evaluate the contribution of features to classification tasks. This metric informs crucial operations such as crossover and mutation. Experimental results on 6 widely used classification datasets to show IGEA's superiority.

Keywords: Feature selection, Classification, Evolutionary algorithm, Information gain

1 Introduction

Classification has emerged as a major focus in machine learning research in recent years due to its wide - ranging applications in pattern recognition , natural language processing , and the medical field . However , the advent of large - scale data have exacerbated the " curse of dimensionality " , presenting additional challenges in classification tasks . High - dimensional data not only increases computational and storage burdens but may also lead to decreased model performance . Therefore , effectively processing high - dimensional data , selecting appropriate dimensionality reduction techniques , and preventing overfitting remain challenging in classification problems .

In dimensionality reduction in data mining , two primary methodologies are prominently utilized : feature extraction and feature selection (FS) . Feature extraction techniques , such as principal component analysis and linear discriminant analysis , transform data into a lower - dimensional space , simplifying the dataset

while attempting to preserve as much information as possible . However , these methods can alter the original representation of the data . In contrast , FS focuses on identifying and retaining a subset of relevant features from the original dataset , thus preserving the original context and interpretability of the features . Despite the clear benefits of FS in maintaining data integrity , it presents its own set of challenges , particularly regarding the combinatorial complexity and the quality of the selected features , which are critical for effective model performance .

One challenge of FS is its combinatorial nature , making it an NP - hard problem . It involves selecting the optimal subset of features from a potentially vast set , a process that becomes increasingly complex with more features . For instance , in the context of classification problems with n original features , the total number of possible feature subsets escalates to 2 n - 1 , excluding the empty set . This exponential growth in potential combinations results in increased spatial and temporal complexity , making the process of identifying the most effective feature combination both computationally intensive and challenging . Despite extensive research and development in FS techniques , there is still a pressing need for further advancement in more efficient and sophisticated methods to keep pace with the growing complexity of data in modern machine learning tasks .

Another challenge is the quality of features within a subset , which significantly influences the performance of FS methods . Features are categorized based on their correlation with labels into irrelevant , inadequately relevant , and strongly relevant categories . Strongly relevant features , which show significant correlation with class labels , are essential for effective prediction models and are prioritized in FS .

In practical applications , FS typically necessitates optimizing multiple objectives concurrently , such as maximizing accuracy and minimizing feature dimensionality . Consequently , FS is often conceptualized as a multi - objective optimization problem , which is referred to as multi - objective optimization FS (MOFS) . Given the complexities inherent in MOFS , various methodologies have been proposed , with multi - objective evolutionary algorithms (MOEAs) noted for their efficacy . MOEAs are distinguished by their superior global search capabilities , the absence of a need for specific domain knowledge , and robustness . Benefiting from a population - based search mechanism , MOEAs excel at addressing varied and often conflicting objectives within a single framework , resulting in a set of nondominated solutions . These solutions reflect the optimal balance between different objectives , making them particularly effective in addressing MOFS .

MOEAs can be categorized into three types based on their environmental selection strategies : 1) decomposition - based ; 2) indicator - based ; and 3) Pareto - dominance - based . The efficacy of decomposition - based methods in MOEAs largely depends on the chosen decomposition techniques and accompanying weight distribution vectors . However , adjusting and optimizing these methods and weight parameters present significant challenges , particularly when Pareto fronts (PFs) are unknown or highly irregular . Indicator - based methods also face challenges in selecting appropriate performance metrics and the additional computational burden this entails . Moreover , the choice of these performance indicators is crucial for the effective functioning of the algorithm , yet simultaneously increases the algorithm 's computational complexity . In contrast , Pareto - dominance - based MOEAs , free from the complexities of selecting decomposition methods , metrics , and optimizing weight parameters , typically require fewer computational resources . Their primary goal is to guide the solution population toward the PFs , representing all nondominated solutions . However , these algorithms often lack sufficient search capability . In addition , managing high - dimensional data , ensuring computational efficiency , and maintaining a balance between exploration and exploitation are focal points in

contemporary research . These limitations highlight the urgent need to develop more effective and advanced algorithms capable of addressing these challenges .

2 Related works

Extensive research has been carried out on feature selection (FS), which is mainly divided into three categories: filter, embedded, and wrapper. Each method provides unique strategies to cope with the complexity of modern datasets. Filter methods assess the importance of features using criteria such as information theory, feature correlation, and distance measures, selecting the top n features from the ranking. Methods that integrate information theory are particularly favored for their ability to evaluate the intrinsic value of features by quantifying "information content," such as symmetric uncertainty, information gain (IG), and the ReliefF function. It is worth noting that the concept of the minimum-redundancy-maximum-relevance (mRMR) criterion has been introduced in some studies. In this approach, features are selected in descending order of their dependence, providing a foundation for the development of many subsequent methods. Filter methods are characterized by high computational efficiency, requiring only a single classifier evaluation. However, determining the number of features n to select without experimental trials remains a challenge. These methods may also overlook the complex interdependencies among features, potentially leading to redundant feature selection and reduced model efficiency. The dependence on specific dataset features may also reduce the generalization ability of classification models derived from filter methods, especially when applied to new datasets. In contrast to filter methods, embedded methods consider the interactions between feature subsets and learning models, aligning particularly well with deep learning models. A common approach is to minimize the loss value and feature coefficients simultaneously. However, these methods have some key issues, including the risk of overfitting, higher computational complexity, and potential ineffectiveness with non-linear feature relationships. The interpretability of the results from embedded methods is also limited, which is a significant drawback in contexts where a clear understanding of feature influence is required.

In feature selection, wrapper methods utilize classifiers, such as k-nearest neighbor (KNN) and support vector machine (SVM), to evaluate subsets and identify the optimal set within an acceptable level of computational cost. These methods directly output an optimal feature subset, effectively capturing complex feature interactions and providing practical, performance-based validation in real-world applications. Among the various wrapper-based methods studied, evolutionary algorithms (EAs) stand out for their adaptability and ability to operate without prior domain knowledge or assumptions about the search space. The population-based search mechanism in EAs enables the generation of multiple solutions and the balancing of competing objectives within a single run, making them highly effective for multi-objective optimization feature selection (MOFS). An important approach involves enhancing the mating selection and environmental selection processes by iteratively modeling Pareto non-dominated fronts using a generalized simplex model. Another method employs an evolvable mask to guide the direction of mutation for decision variables, thereby ensuring solution sparsity. In addition, some studies have introduced a multi-objective evolutionary algorithm that integrates unsupervised neural networks. It combines sparse distributions and compact representations learned from restricted Boltzmann machines and denoising autoencoders, providing an optimal search space for genetic operators. From a different perspective, some studies use the distance between solutions in the decision space as a similarity crite-

rion. After eliminating highly similar solutions, these algorithms employ diversity-based selection methods to maintain Pareto dominance. Furthermore, some studies have introduced a new mutation operator that combines local and global information to locate promising feature subsets and maintain overall search efficiency.

Although these methods have proven effective for MOFS, the increasing dimensionality of data presents significant challenges. Existing algorithms often become trapped in local optima when dealing with large-scale feature selection problems. Moreover, the presence of redundant and irrelevant features increases computational complexity and reduces search efficiency. As the search space expands, it becomes increasingly difficult for search operators to find optimal solutions that effectively balance dimensionality reduction with classification accuracy. In addition, the limited capacity of the algorithm to balance exploration and exploitation hinders its ability to traverse the solution space and effectively identify the most valuable and informative feature subsets. Therefore, as mentioned earlier, there is a continued need for a more precise method to assess the importance of features. To address these challenges, we propose an innovative approach utilizing principles from information theory.

2.1 Information gain

IG is a measure originating from information theory that quantifies the expected amount of information gained by partitioning data based on specific features. In IG, this information amount is primarily measured using the concepts of entropy and conditional entropy. Firstly, the entropy of a discrete random variable X is defined as follows:

$$H(X) = -\sum_{x_i \in X} P(x_i) \log(P(x_i))$$

Where x_i denotes a particular outcome of the random variable X, and $P(x_i)$ signifies the likelihood of occurrence of x_i across the entire value range of X. Next, the conditional entropy of X in the context of a distinct discrete random variable Y is described as follows:

$$H(X|Y) = -\sum_{y_j \in Y} P(y_j) \sum_{x_i \in X} P(x_i|y_j) \log(P(x_i|y_j))$$

Where $P(y_j)$ is the probability of observing y_j from Y, $P(x_i|y_j)$ is the conditional probability of observing x_i from X given that y_j from Y has occurred. Finally, the Information Gain (IG) between X and Y is used to measure the amount of information shared by X and Y together:

$$IG(X,Y) = H(X) - H(X|Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P(x_i, y_j) \log\left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right)$$

Fundamentally, IG indicates how effectively a feature segregates data into more homogeneous, or "pure" groups in relation to the target outcome. Compared to the entire dataset, features with high IG create subsets that are more predictable and less random when utilized for data segmentation. Thus, IG signifies the value of features in diminishing the uncertainty of target variables. This reduction in uncertainty is intrinsically linked to the efficacy of features in enhancing the accuracy of model predictions.

3 Experiment design

3.1 Classification datasets

To ensure a comprehensive assessment of generalization ability and effectiveness in FS tasks, this study meticulously selected a series of challenging and representative datasets. These datasets are deeply rooted in real-world application scenarios, ensuring the practical relevance and applicability of our experiments. They also cover a wide range of features, categories, and instance number, providing a holistic evaluation of the performance of the proposed IGEA.

No.	Name	Features	Instances	Classes	No.	Name	Features	Instances	Classes
1	Zoo	16	101	7	13	HAR	561	2947	6
2	Climate	18	540	2	14	Isolet5	617	1559	26
3	Wine	22	267	2	15	Multiplefeatures	649	2000	10
4	Spect	44	267	2	16	CNAE-9	856	1080	9
5	Spambase	57	4601	2	17	Yale32	1024	165	15
6	Sonar	60	208	2	18	ORL32	1024	400	40
7	Hillvalley	100	1212	2	19	QSAR-androgen-receptor	1024	1687	2
8	Urban Land Cover	147	675	9	20	SRBCT	2308	83	4
9	Semeion-handwritten-digit	256	1593	10	21	DriveFace	6400	606	3
10	SCADI	205	70	7	22	Duke-breast-cancer	7129	44	2
11	LSVT-voice-rehabilitation	310	126	2	23	Orlraws10P	10304	100	10
12	Madelon	500	2600	2					

Figure 1. datasets in Ascending Order of the Feature Number

Six datasets were selected, namely Zoo, Spambase, Hillvalley, Semeion, CNAE-9, and SRBCT. These datasets have varying numbers of features, ranging from dozens to thousands, and also differ in the number of instances.

3.2 Performance indicators

I initiated the performance analysis of various MOEAs in FS tasks by first evaluating the minimum classification error (MCE). MCE measures the precision of classification by the solutions obtained in each algorithmic run. Following this, hypervolume(HV) is also selected to evaluate the IGEA, HV is a robust metric for evaluating multiobjective optimization solutions, quantifying the extent of the objective space covered by the solution set. A higher HV value suggests more comprehensive attainment of multiple objectives, indicative of superior algorithm efficacy. The HV metric measures the volume of the area enclosed by the PF in the target space, typically relative to a reference point or set of reference points. In two-dimensional space, this is represented as the area below the PF; in higher dimensions, it extends to volume. For the bi-objective optimization task in this study, with feature proportion and classification error rate as objectives scaled between 0 and 1, we set the reference point for HV at (1,1). Thus, a greater HV corresponds to enhanced algorithm performance.

3.3 Parameter setting

In this study, the experiments were conducted using MATLAB on the open-source platform PlatEMO. To ensure the reliability of the experimental outcomes, each competing algorithm was run 30 times independently on each dataset. The parameters common to all algorithms included an adaptive population size N, which was determined based on the number of features in the dataset, with a maximum limit of 200. Specifically, N was

set at 100 for datasets with fewer than 100 features, capped at 200 for those with more than 200 features, and matched to the feature count for all other cases. The termination criterion for each algorithm was set at 100 times the population size, based on the maximum number of evaluations of the objective function, implying a maximum evolutionary generation count of 100. For classification tasks, each dataset was randomly split into training and test subsets, approximately in a 7:3 ratio. To maintain statistical credibility, the division of datasets was consistent across all 30 runs for each algorithm. During training, the KNN algorithm, with 10-fold cross-validation, was used to compute the classification error on the training set. The number of nearest neighbors in KNN was set to 5.

4 Experimental results and discussion

	Mean	Std	论文中结果
CNAE9	0.149485597	0.010743265	0.087
HillValley	0.393040293	0.015320424	0.395
semeion	0.15520615	0.007272347	0.089
Spambase	0.154371981	0.015085917	0.153
SRBCT	0.034722222	0.029122161	0.018
Zoo	0.164444444	0.019443076	0.153

Figure 2. MCE results

	Mean	Std	论文中结果
CNAE9	0.8487419	0.010017015	0.903
semeion	0.6233872	0.009879267	0.878
SRBCT	0.9671318	0.026377143	0.982
HillValley	0.5358018	0.012166798	0.63
spambase	0.7158928	0.024376702	0.808
Zoo	0.7306129	0.012383873	0.768

Figure 3. HV results

From the experimental results, the deviation from the original in the paper is not significant, but there are still areas for improvement.

5 References

[1] Zhang, B., Wang, Z., Li, H., Lei, Z., Cheng, J., et al. (2024). Information gain-based multi-objective evolutionary algorithm for feature selection. Information Sciences, 677, 120901.