WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Machine Learning and Predictive Analytics
Vol. 6, No. 2, 2023
Doi: 10.67228/3142788X/IJMLPA-2023PII8X5W
PP. 01-19

Original Article

AI-Based Automated Feature Selection for Predictive Modeling

Dr. Liu Fang

Associate Professor, Tsinghua University, China.

Received: 09-02-2023 Revised: 09-22-2023 Accepted: 09-30-2023 Published: 10-03-2023

ABSTRACT

Artificial Intelligence (AI) has improved predictive modeling by enabling efficient analysis of large-scale and high-dimensional datasets. Before 2019, selecting relevant features from complex data was a major challenge. Feature selection is important for improving accuracy, reducing computational cost, preventing overfitting, and enhancing model interpretability. Traditional methods such as filter, wrapper, and embedded approaches had limitations with big data and nonlinear relationships. To address these issues, AI-based Automated Feature Selection (AFS) techniques were introduced. These methods combine machine learning, optimization algorithms, and intelligent search strategies to automatically identify the most relevant features. Common approaches included Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests (RF), and Reinforcement Learning (RL). This study reviews AI-driven feature selection techniques developed between 2009 and 2018 and proposes a hybrid framework combining filter-based preprocessing, evolutionary optimization, and machine learning evaluation. Experimental results show that AI-based hybrid methods outperform traditional techniques in accuracy, scalability, and feature optimization, demonstrating their importance in modern predictive analytics systems.

KEYWORDS

Artificial Intelligence, Automated Feature Selection, Predictive Modeling, Machine Learning, Genetic Algorithm, Particle Swarm Optimization, Classification, Dimensionality Reduction, Hybrid Optimization, Data Mining, Evolutionary Computing, Intelligent Systems, Feature Engineering, Support Vector Machine, Neural Networks.

ISSN: 3142-788X
PP: 01-19.

1. INTRODUCTION

1.1. Background

Predictive modeling is one of the most important uses of machine learning and artificial intelligence technologies, since it allows computational systems to accurately predict the future based on available data from the past. However, the shift to digital data and developments in computational intelligence have driven the advent of predictive modeling, its rapid growth, and widespread use in several industries prior to 2019. Predictive models were heavily used in applications like healthcare diagnostics, fraud detection, financial forecasting, customer behaviour analysis, weather forecasting, manufacturing automation, recommendation systems and cybersecurity, providing intelligent decision-making and risk assessment. These systems used statistical learning algorithms and machine learning methods to find patterns, trends and relationships in the vast amounts of data. The main limiting factor, however, was the quality of the features and their relevance and dimensionality to the predictive model. Many datasets in thereal world contained many redundant and irrelevant attributes, noisier attributes, and highly correlated attributes, all of which had a detrimental impact on learning efficiency and prediction accuracy. Adding irrelevant features made the computations more complex, training more time-consuming, and led to overfitting, which is the tendency of the model to perform well on the training set but not so well on new data. This led to the emergence of feature selection as a key preprocessing phase in the predictive modeling processes. The dimensionality reduction performed by effective feature selection techniques not only reduced the dimensionality but also eliminated the irrelevant attributes, improved the interpretability of the models and also enhanced the classification performance. As a result of the trend to optimize features in an efficient manner, the researchers began to create sophisticated automatic feature selection techniques that are based on artificial intelligence and evolutionary optimization techniques.

1.2. Importance of Automated Feature Selection

Automated feature selection is a key factor in the improvement of the efficiency, accuracy, and scalability of predictive modelling systems. With the growth and complexity of data prior to 2019, traditional manual feature engineering was difficult, time-consuming and costly. With the development of large datasets, automated feature selection techniques were developed to intelligently find the most relevant and informative features without requiring much human effort. These methods greatly improved the accuracy of machine learning by simplifying data, removing redundant features from the data, and increasing the ability of the model to be generalized. Automated feature selection also helped with quicker training, reduced memory consumption, and better decision-making for healthcare analytics, cybersecurity, financial forecasting, and intelligent automation systems among other applications.

WCRG
Global Research
ISSN: 3142-788X
PP: 01-19.

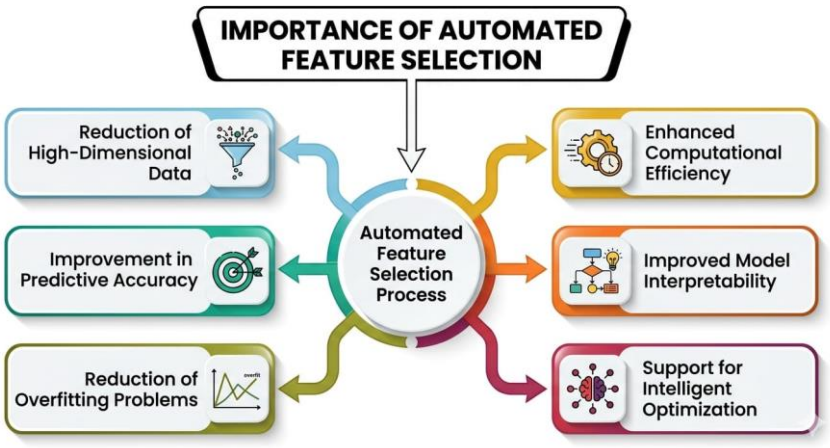

Fig 1 - Importance of Automated Feature Selection

1.2.1. Reduction of High-Dimensional Data

The primary benefit of the automatic feature selection is that it is capable of selecting a small subset of features from a large set of features that hold significant meaning. In the real world, there can be thousands of attributes, some of which have little value in predicting performance. As the dimensionality of data grows, it will place demands on computation and will make it harder to train models. Using automatic feature selection techniques, the unnecessary features are discarded, leaving the most informative features, which will make the learning process faster and efficient.

1.2.2. Improvement in Predictive Accuracy

The automated feature selection optimizes the predictive performance by discarding irrelevant and noisy features that can adversely impact machine learning models. A model could learn spurious patterns or be sensitive to noise if it is presented with irrelevant attributes in the training set. Automated systems allow machine learning algorithms to concentrate on meaningful relationships within the data, by filtering out irrelevant features. This results in more precise predictions, higher classification accuracy and generalization to unseen data.

1.2.3. Reduction of Overfitting Problems

Overfitting is when a predictive model learns the patterns as presented on the training data rather than the general patterns. This is a problem that occurs when the data set has many irrelevant or redundant features. The automatic feature selection further avoids overfitting by discarding irrelevant features and streamlining the complexity of the model. Compact feature subsets make models more consistent in their performance when they are applied to real-world or other test environments.

1.2.4. Enhanced Computational Efficiency

Training and evaluation of large-scale data take plenty of computational resources. Automated feature selection will help to reduce the size of the dataset, thus reducing the training time, memory usage and processing overhead. In applications requiring real-time data processing,

ISSN: 3142-788X
PP: 01-19.

cloud computing, and intelligent decision-making systems, efficient feature reduction is particularly crucial. When the amount of data is massive and constantly increasing, faster computation allows machine learning systems to scale dynamically without any issues.

1.2.5. Improved Model Interpretability

A smaller number of features that have a more direct relationship to the problem will be more understandable and interpretable in terms of machine learning models. Automated feature selection is a useful tool for researchers and domain experts to locate the most influential features that impact the prediction results. In critical applications like healthcare diagnosis, financial analysis, and industrial monitoring, where the reasoning behind predictions is crucial for trust and decision-making, improved interpretability is of paramount importance.

1.2.6. Support for Intelligent Optimization

Automated feature selection methods can be used as a solid basis for the intelligent optimization process based on artificial intelligence and evolutionary algorithms. In complex data sets, features can be automatically combined using methods like Genetic Algorithms, Particle Swarm Optimization, Ant Colony Optimization and machine learning based embedded approaches. The intelligent methods will not only enhance the quality of feature selection but also help to decrease the complexity of searching and boost the overall predictive modeling.

1.3. Challenges in Traditional Feature Selection

Traditional feature selection methods had several problems that prevented them from being effective and being used in big and complex predictive modeling tasks. As the amount, dimensionality, and diversity of data grew in various sectors of industry, traditional feature selection techniques started to become inefficient, scalable, and good at predicting. This made researchers interested in intelligent and automated feature optimization techniques with Artificial Intelligence and Machine Learning algorithms. The curse of dimensionality was one of the problems that had to be faced. The greater the number of features in a data set, the greater the number of possible combinations of features that need to be considered for the best feature subset. In many high dimensional datasets, certain features were redundant, irrelevant, and noisy, which made the selection of features more difficult and time-consuming. This issue made the machine learning models less efficient and impacted their prediction accuracy. There were also computational costs that were a limitation. The traditional feature selection methods based on wrappers needed multiple training and evaluation of machine learning classifiers to evaluate various combinations of features. While wrapper methods generally performed better than filter methods in predicting the test set, they were very expensive in terms of computation and time. This was not a good option for dealing with large data sets or real-time applications for quick decision-making. The traditional statistical feature selection methods were also found to have feature dependency issues. Features have been evaluated independently using methods like correlation analysis, Chi-Square testing, and Information Gain but they often did not account for nonlinear relationships and interactions between features. This resulted in some feature combinations being omitted from the model, which meant that

ISSN: 3142-788X
PP: 01-19.

the usefulness of those features was neglected, causing the effectiveness of the model and predictive performance to drop. Another primary challenge in the conventional feature selection systems was their scalability. Efficiently handling large-scale streaming data, cloud infrastructures, and real-time analytics systems in the rapidly expanding big data environments is not an environment for which conventional algorithms were designed. Training time, memory usage and optimization efficiency were all significantly impacted by dataset size. Moreover, inappropriate feature selection posed the risk of overfitting. Adding irrelevant or redundant features into training sets would make machine learning models more likely to memorize training data instead of learning patterns. This undermined the effectiveness of models in performing well on unseen data and reduced the reliability of predictions. These challenges underscored the importance of sustainable, intelligent and automated feature selection systems that can process complex modern data environments.

2. LITERATURE SURVEY

2.1. Filter-Based Feature Selection Methods

Filter based feature selection techniques are one of the oldest and popular techniques in predictive modeling. Such techniques assess the relevance of each feature, and no machine learning algorithm is used in the feature selection process. Filter methods are computational efficient, independent of classifiers and very well applicable for high dimensional data, like text mining, bioinformatics, or sensor data analysis. They are simple and flexible enough to be used for preprocessing huge datasets prior to training models. However, typical filter approaches do not capture dependencies and interactions between features which can limit the predictive power of the final model.

2.1.1. Chi-Square Tes

The Chi-Square test is a type of statistical test that quantifies the relationship between categorical independent variables and categorical dependent variables. It assesses the degree of association between a feature and a class label in feature selection, using observed and expected frequencies. Those features that contain more chi-square values are found to be more relevant to classification tasks. This is an easy to implement and low computation cost technique which can be used for text classification and document analysis. Its effectiveness, however, is dependent on the distribution of the data and works best with categorical data.

2.1.2. Information Gain

Information Gain is used to determine how much uncertainty or entropy is reduced when a feature is used to partition the data. It is one of the methods used in decision tree learning and it tells the usefulness of features in predicting the target variable. The features that have a higher information gain are more informative and relevant for predictive modeling. It is a method that is fast, scalable and suitable for ranking features in large datasets. Although it has its merits, Information Gain is feature-wise and doesn't consider relationships between multiple features.

ISSN: 3142-788X
PP: 01-19.

2.1.3. Correlation Coefficient

Correlation-based feature selection looks for relationships between the features, and the target variables, and quantifies the linear association between them. Features that are highly correlated with the target variable are important (and can be kept in the model), and redundant features with strong intercorrelation can be eliminated to avoid duplication. This makes the dataset more compact, and easier to compute. In the case of structured data set, correlation analysis is simple and effective method to remove redundant attributes. It primarily detects linear relationships, though, it may not be able to detect complex non-linear relationships between features.

2.1.4. Mutual Information

The Mutual Information is a metric that measures the common information conveyed by the input features and classes. It is more powerful than correlation analysis, because it is able to detect linear and nonlinear relationships. The higher mutual information values are indicators of the higher predictive relevancy of features in classification or regression. This is a method commonly applied in image processing, bioinformatics and pattern recognition. But to estimate the mutual information accurately in high dimensional spaces, it might take more computational resources and large sample sizes.

2.1.5. Fisher Score

The Fisher Score is a supervised feature selection method designed to evaluate the discriminative capability of features using the inter-class separation and intra-class similarity. The closer the distance between classes, and the further the distance between different classes, the higher the scores. Fisher Score is a very efficient computation and is extensively used in pattern recognition and image classification problems. It is able to rank features for classification tasks but it does not evaluate feature relationships or effects of multiple features together.

2.2. Wrapper-Based Feature Selection Methods

Wrapper-based feature selection methods assess the subsets of features on the basis of the predictive performance of machine learning models. Wrapper methods take interactions and dependency relationships among features into account, which may lead to better classification or prediction accuracy than filter methods. It tries combinations of features and selects subsets to achieve the best model performance. While wrapper methods tend to work better than filter methods, they are very expensive in terms of computation due to the need to run the learning algorithm multiple times for the different subsets of features. When accuracy is more critical than computational speed, wrapper methods may be helpful.

2.2.1. Sequential Forward Selection

Sequential Forward Selection is an iterative wrapper approach which starts with a set of no features and then selects features sequentially one by one. The feature that results in the best improvement in the model accuracy is chosen and added to the subset at each iteration. This goes on until there is no significant performance increase. This approach is easy to implement and accurate in

ISSN: 3142-788X
PP: 01-19.

the dimensionality reduction of features without compromising the predictiveness. But once a feature is added it cannot be removed later, which can result in sub-optimal feature subsets.

2.2.2. Sequential Backward Elimination

In Sequential Backward Elimination, all the features are included in the initial iteration and then the most insignificant features are eliminated iteratively. One by one each step, the algorithm assesses how well the model works without a particular feature and discards the feature that is least likely to be useful for prediction. This continues until the number of features desired is left. The approach generally generates better feature subsets than the forward selection approach since it starts with all possible interactions of the features. But it becomes costly with large number of attributes in the data sets.

2.2.3. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a useful wrapper method commonly employed for SVM classifiers. The algorithm repeatedly trains a model, ranks features in order of importance and drops the least important feature one by one. This is an iterative elimination process until optimum subset of features is achieved. Recursive Feature Elimination helps in interpreting the models and the better prediction by removing redundant and irrelevant attributes. It is very effective, but it can be very expensive of computation resources, particularly with large data sets.

2.3. AI and Evolutionary Optimization Techniques

Because of their capacities to solve complex optimization problems, AI and evolutionary optimization techniques have become a popular solution for feature selection research prior to 2019. These algorithms look for optimal subsets of features with adaptive and intelligent search algorithms based on natural processes. An AI-based optimization algorithm is effective in dealing with a large search space, nonlinear relations, and feature dependency characteristics. Evolutionary algorithms can usually obtain higher classification accuracy and better feature reduction than the traditional methods. They can take a long time to process and tune parameters to their optimum performance, however.

2.3.1. Genetic Algorithm (GA)

The Genetic Algorithms are evolutionary optimization methods that are based on biological evolution and natural selection. In feature selection, the chromosomes are the different feature sets that are evaluated, and the evolutionary operators (selection, crossover, and mutation) are used to evolve better solutions over the generations. Each solution is judged on the basis of prediction accuracy and the ability to reduce the features. Genetic Algorithms excel in being able to explore global search spaces and avoiding local optima. They are versatile and stable, and have many applications in medical diagnosis, image processing and data mining.

ISSN: 3142-788X
PP: 01-19.

2.3.2. Particle Swarm Optimization (PSO)

PSO is a swarm intelligence method that is based on the flocking of birds and schools of fish. Particles are candidate feature subsets that traverse the search space and learn about the best solutions in feature selection. The position of each particle depends on its own experience and the experience of its neighbours. PSO has a rapid convergence speed, easy implementation and good optimization ability. It is especially well suited to large scale optimization problems and large-dimensional data sets.

2.3.3. Ant Colony Optimization (ACO)

The concept of Ant Colony Optimization is based on that of ants that communicate by pheromone trails while foraging. Artificial ants are used in feature selection to explore various subsets of features and strengthen the ones that are good candidates for feature selection by updating to pheromone. The more a feature helps with classification, the more it will be emphasized by the pheromones for selection in subsequent iterations. The ACO algorithm is efficient in searching complex search spaces and has a balance of exploration and exploitation. It is used successfully in pattern recognition, intrusion detection and biomedical data analysis.

2.4. Machine Learning-Based Feature Selection

The machine learning feature selection methods incorporate feature evaluation into the learning process. These techniques are used in the process of training predictive models to discover important features, and feature optimization and learning are done simultaneously. In embedded feature selection approaches, the selection of features is usually accompanied by feature interactions and model-specific characteristics, resulting in better predictive performance. The classification, regression, intelligent decision support systems, healthcare analysis, and financial prediction are among its many uses in machine learning-based techniques.

2.4.1. Decision Trees

Decision Trees use the principle of feature selection by selecting features that best partition the data into homogeneous classes. More important features are those with higher information gain (or impurity reduction). Decision Trees are interpretable and efficient to compute, as it is the natural structure of the tree that ranks the features according to their predictive power. They are extensively employed in classification and decision making systems which are based on rules. But, decision trees can be unstable with noisy and very complex data.

2.4.2. Random Forests

Random Forests are an extension of the decision tree learning method which build several trees and use ensemble learning to combine the results. The feature importance is calculated from the contribution of each feature over all of the trees in the forest. Random Forests are robust in high dimensional data, avoid overfitting and offer high prediction accuracy. They are stable to noise and able to detect nonlinear feature relationships. Random Forests were among the most widely used embedded feature selection methods prior to 2019, for their reliability and scalability.

ISSN: 3142-788X
PP: 01-19.

2.4.3. Support Vector Machines

SVM is the superlative supervised learning model that is commonly employed in various feature selection and classification problems. SVM-based Recursive Feature Elimination reduces the number of features by eliminating the least important features. This will enhance the generalization of the model as well as lower computation works. SVM methods are especially well suited for high dimensional data like text classification, gene expression analysis, etc. But it can be costly to train SVM models when the data is very large.

2.4.4. Neural Networks

ANN can detect complex nonlinear relationships between features by using layered learning architectures. In training, neural network automatically tunes the connection weights, which results in identification of the most important features that have strong impact on the results of the prediction. The neural networks are very good at image recognition, speech processing and deep learning applications. They are useful for more complex predictive modelling tasks because they can learn complex interactions or patterns that are hidden. Neural networks, however, can be time consuming to train and can need a lot of data, power, and tuning.

3. METHODOLOGY

3.1. Proposed AI-Based Automated Feature Selection Framework

The proposed automated feature selection framework leverages AI to help pinpoint the most relevant and informative features within massive datasets, enhancing the performance of predictive modeling. The framework integrates statistical filtering methodology, evolutionary optimization algorithms and machine learning evaluation approaches to produce optimized feature subset. The framework is composed of several selection steps which reduces dimensionality, eliminating redundancy, improving classification accuracy, and reduce computational complexity. The entire process guarantees not only the efficiency and scalability of the model but also its contribution to predictive decision-making.

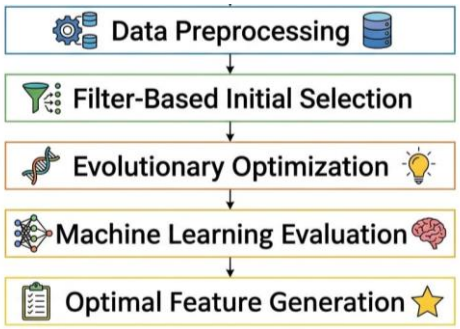

Fig 2 - Proposed AI-Based Automated Feature Selection Framework

3.1.1. Data Preprocessing

The first and one of the most important steps followed in the proposed framework is data preprocessing. Raw data frequently may have missing values, noisy records, duplicate records,

ISSN: 3142-788X
PP: 01-19.

inconsistent data formats and attributes that can have a negative impact on predictive performance. Imputation is used to address missing data, cleaning is used to remove noisy data, and when needed, categorical attributes are converted to numerical. Data is also normalized and standardized to ensure consistent feature scaling and convergence of the algorithms. This step enhances the quality of the data and is designed to make it ready for the efficient selection of features and machine learning analysis.

3.1.2. Filter-Based Initial Selection

The initial selection is done using the filter, thereby decreasing the dimensionality of the data set prior to using computationally costly optimization algorithms. To assess the relevance of each feature independently, statistical methods like InformationGain, Chi-Square, Fisher Score, Correlation Coefficient and Mutual Information are applied. Only those features that have low relevance scores are deleted and the features with high relevance scores are kept for further processing. The preliminary reduction reduces the computational load and increases the efficiency of optimizing by removing redundant and irrelevant features early. The filter phase also aids in scalability with high dimension data sets.

3.1.3. Evolutionary Optimization

The central intelligence element of the proposed framework is the evolutionary optimization. The AI optimization algorithms (like Genetic Algorithm, Particle Swarm Optimization, or Ant Colony Optimization) are used during this stage to identify the optimal set of features. These algorithms learn by trying various combinations of features, assessing the success of each combination (by optimization methods modeled after nature or swarm intelligence) and select the most successful one for final evaluation. The optimization process aims to find the optimal set of features that maximizes the model performance while minimizing the complexity by reducing the features.The optimization process is designed to reduce the number of features while keeping the model as accurate as possible. Using evolutionary optimization, the global search capacity is enhanced and the possibility of finding local optimum solutions is reduced due to traditional methods.

3.1.4. Machine Learning Evaluation

During machine learning evaluation phase, the effectiveness of the selected feature subsets is validated by employing predictive models. The optimized features are used to train algorithms like Decision Trees, Random Forests, Support Vector Machines, and Artificial Neural Networks. The accuracy, precision, recall, f1 score, and computational efficiency are used to evaluate model performance. This assessment process aims to identify a set of features that are beneficial for prediction quality and generalization ability. The system incorporates machine learning evaluation to identify feature subsets with high predictive accuracy and robust model performance, thereby enhancing the overall performance.The system is enhanced by incorporating machine learning evaluation to extract feature subsets with high predictive accuracy and robust model performance.

ISSN: 3142-788X
PP: 01-19.

3.1.5. Optimal Feature Generation

The last phase of the framework creates the most effective subset of features to be utilized for predictive modelling and decision-making. The features shown are the most informative, not redundant and best performance enhancing from the original features. This results in a much reduced computation burden, faster model training, increased model interpretability and better prediction accuracy. When it comes to machine learning system deployment, such as healthcare analytics, financial forecasting, cybersecurity, image recognition, and intelligent automation systems, optimal feature generation helps streamline the process.

3.2. Data Preprocessing

Since the quality of the input data influences the performance of the machine learning models and optimization algorithms, data preprocessing is a crucial step in the proposed AI-based automated feature selection framework. Prediction accuracy and computational complexity may be limited in real-world data sets due to incompleteness, inconsistencies, noises, and lack of structure. Hence, preprocessing techniques are used to convert raw data to a clean, reliable and structured data format which are suitable for feature selection and predictive analysis. The impact of effective preprocessing is that it makes data more useful for learning, more stable in model predictions, and more accurate in terms of data integrity. The most important preprocessing task is handling the missing values. Data sets are often incomplete because of sensor failure, man-made mistakes or limitation in data collection. If not handled appropriately, missing values can cause biased predictions and inferior learning efficiency. Common methods for filling in the missing values in the data include mean substitution, median replacement, interpolation and nearest-neighbor imputation. By carefully dealing with the missing information, this consistency in the dataset helps in maintaining important training samples and keeps the dataset consistent. Another important preprocessing step involves normalizing numerical features, which involves scaling them into a common range. Normalization makes the inputs of machine learning algorithms more comparable as their numbers tend to converge faster in the training process, typically because the magnitudes of the features are similar. Min-Max normalization and Z-Score normalization prevent large values from masking small values in attributes when analyzing the data. Noise reduction: to remove irrelevant or distorted information that could adversely impact predictive performance. Measurement error, environmental noise, or faulty data entry could be the sources of noise. To enhance the clarity and minimize inconsistencies in the data, various filtering techniques, smoothing algorithms, and clustering-based approaches are employed. The removal of outliers also is crucial because the statistical analysis and modelling of behavior can be misled by extreme values. Outliers are determined through statistical deviation, clustering or distance measures and corrected or eliminated. Overall, data preprocessing bolsters the quality of the dataset, guarantees more solid features, and establishes a solid platform for effective automatic feature selection and correct prediction formulation.

ISSN: 3142-788X
PP: 01-19.

3.3. Hybrid Feature Selection Algorithm

The proposed hybrid feature selection algorithm is combining the statistical feature ranking, evolutionary optimization and machine learning evaluation to select the most informative and predictive feature subset. The hybrid approach aims to combine the best attributes of the single feature selection methods to overcome their drawbacks, and is designed to combine the filter and optimization based feature selection methods. Using statistical ranking, redundant and non-informative features are eliminated firstly and evolutionary optimization is used to find the best possible combination of features. In the last step, machine learning models are used to determine the usefulness of the selected features in predictive tasks. This integrated approach helps to achieve better classification accuracy, dimensionality reduction, computational efficiency, and overall model performance.

Fig 3 - Hybrid Feature Selection Algorithm

3.3.1. Feature Ranking

In the first step of hybrid algorithm, the Information Gain analysis is used to do feature ranking. Information Gain is used to evaluate the usefulness of each feature in the data set for reducing the uncertainty of the target variable. The higher the scores of features the more relevant they are, as they will give more information for a classification or regression task. This phase, each feature is independently assessed, and the features which are not pertinent and of low rank are removed from the data set. This is a preliminary filtering process which reduces dataset dimensionality and lower the computational load of the later optimization stages. Feature ranking is also helpful in making the algorithm more efficient by having only the most informative features to analyze.

3.3.2. Genetic Optimization

Once the features are reduced, the reduced feature subset is used as input to a Genetic Algorithm to find the optimal set of features. Genetic Algorithms is a computer simulation of the biological evolution process that involves selection, crossover and mutation. During this, each chromosome is a candidate feature subset and the population is iteratively evolved to find better solutions. Subsets of candidates get assessed according to their predictive power and feature reduction ability. The evolutionary search mechanism allows the algorithm to find solutions efficiently in large and complex search spaces without getting stuck at local optimum solutions. Genetic optimization enhances the quality of feature selection by finding a combination of features that is efficient for the purpose, instead of analysing a single feature.

ISSN: 3142-788X
PP: 01-19.

3.3.3. Model Evaluation

The last step of the hybrid algorithm is to assess the optimized feature subsets using the Random Forest Classifier. Random Forest is an ensemble learning technique which builds multiple decision trees and merges the results of these decision trees to enhance the accuracy of the predictions and the robustness of the model. The classifier evaluates the selected features by considering evaluation metrics like accuracy, precision, recall, and F1-score. Subsets with the best classification results are deemed to be optimal for predictive modelling. Random Forest is more reliable because it can effectively deal with nonlinear relationship, noise and high dimension data. The evaluation stage aims to ensure that the selected features are highly relevant to the accurate and efficient machine learning task.

3.4. Performance Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of the proposed AI-powered automated feature selection framework, ensuring its effectiveness and reliability. These metrics are used to evaluate the performance of the machine learning model once the most relevant features have been chosen. Evaluation metrics give quantitative evidence in predictive modelling and classification about the quality of a model, decision-making ability and generalization performance. The proposed framework is based on four main evaluation metrics – Accuracy, Precision, Recall, and F1-Score. These measures give a comprehensive view of the model behaviour for various prediction conditions. In machine learning, accuracy is one of the most popular performance measures. It calculates the percentage of "correct" predictions out of all the predictions made by the model. In simple terms, the accuracy is the proportion of all the predictions that are correct, to all the observations in the data set. Correct predictions are true positives (positive instances that are correctly recognized) and true negatives (negative instances that are correctly recognized). High accuracy means that the features selected play a beneficial role in achieving overall predictive performance. Precision compares the accuracy of positive predictions made by the model. It represents the percentage of the actual positives that are predicted. In particular, precision is crucial in areas where false-positive predictions can have serious implications, such as medical diagnosis, fraud detection, and cybersecurity systems. A high precision value means that the model has a smaller amount of false positive predictions. Precision measures the extent to which positive predictions made by the model are correct. Measures the ratio of true positive predictions to all positive predictions. In applications with potential severe consequences if false positives are predicted, like medical diagnosis, fraud detection and cyber security system, precision is crucial. High precision value means that the model makes less false positive predictions. Recall refers to how well the model performs on correctly identifying actual positive data from the data set. It is defined as the ratio of the number of cases correctly predicted to the total number of actual cases. Recall is particularly important in application domains like disease detection and fault monitoring systems where the consequences of missing positive cases can be grave. The higher the recall value, the more sensitive and good the detecting ability. The F1-Score is a single performance measure that takes into account both precision and recall. It is a combination of precision and recall; it is calculated as harmonic mean of precision and recall to give a balance of both measures. In the case of imbalanced

ISSN: 3142-788X
PP: 01-19.

datasets, the F1-Score is especially informative to assess the quality of the predictions. These evaluation parameters are used to evaluate the efficiency, reliability and predictive power of the proposed feature selection framework.

4. RESULT AND DISCUSSION

4.1. Experimental Setup

The experimental setup aimed to validate the effectiveness and reliability of the proposed automated feature selection framework based on artificial intelligence algorithms with a set of benchmark datasets commonly used in research prior to 2019. The main goal of experimental analysis was to evaluate the efficacy of the proposed methodology in terms of feature dimensionality reduction and prediction accuracy and efficiency. All experiments were carried out under controlled conditions to achieve consistency, repeatability and a fair comparison of the performances of the various feature selection stages and machine learning evaluations. Proposed framework implementation was done in Python as it is simple, flexible, and has a wide array of support for libraries for data analysis and machine learning. Pyhton offers a suite of powerful libraries like Scikit-learn, NumPy, Pandas, and Matplotlib that make it easy to preprocess, optimize, select data features, train models, and visualize performance. During the experimental analysis, its use of Python also enhanced the scalability and helped to reduce the complexity of the development process. Random Forest (RF) classifier was chosen as the main machine learning model to evaluate performance. Random Forest is an ensembler that builds several decision trees and averages their predictions to enhance prediction accuracy and robustness. The classifier was selected due to its property of dealing with high dimensional data, noisy data, nonlinear relationships and feature interactions very well. Moreover, Random Forest offers feature importance analysis, which can further be used to evaluate certain feature subsets. The Genetic Algorithm has been used as the basic evolutionary search method for optimization. The Genetic Algorithm is an iterative process, where different feature subsets are explored through the use of operations such as selection, crossover and mutation, until an optimal subset is found. It can be used to obtain the global optimum solution and search in a complex search space efficiently without falling into local optimum solutions. The method for evaluating reliability and generalization ability of the proposed framework was 10-Fold Cross Validation. In this method, the data set will be divided into 10 equal parts and 9 of them will be used for training and 1 for testing in each iteration. This is done 10 times, such that each subset is used for training and testing. Cross validation helps minimize overfitting, enhance the reliability of the evaluation, and gives a better idea of the model's performance on various data distributions.

4.2. Performance Comparison Table

The performance comparison analysis was performed to see the effectiveness of the different feature selection techniques on the basis of classification accuracy, precision, recall, and feature reduction capability. The results of the experiments show that the proposed advanced AI-based optimization techniques are more efficient than the traditional feature selection techniques for selecting feature subsets with higher information content and fewer features. The comparison also emphasizes the benefits of integrating the three components: statistical filtering, evolutionary

ISSN: 3142-788X
PP: 01-19.

optimization and machine learning evaluation in a hybrid approach. Overall, the proposed hybrid method of feature selection using AI was the most successful of all the methods tested.

Table 1: Performance Comparison Table

Method	Accuracy (%)	Precision (%)	Recall (%)	Feature Reduction (%)
Traditional Filter	82%	80%	79%	35%
Wrapper Method	86%	84%	83%	42%
PSO-Based Selection	89%	87%	88%	51%
GA-Based Selection	91%	90%	89%	56%
Proposed Hybrid AI Method	95%	94%	93%	68%

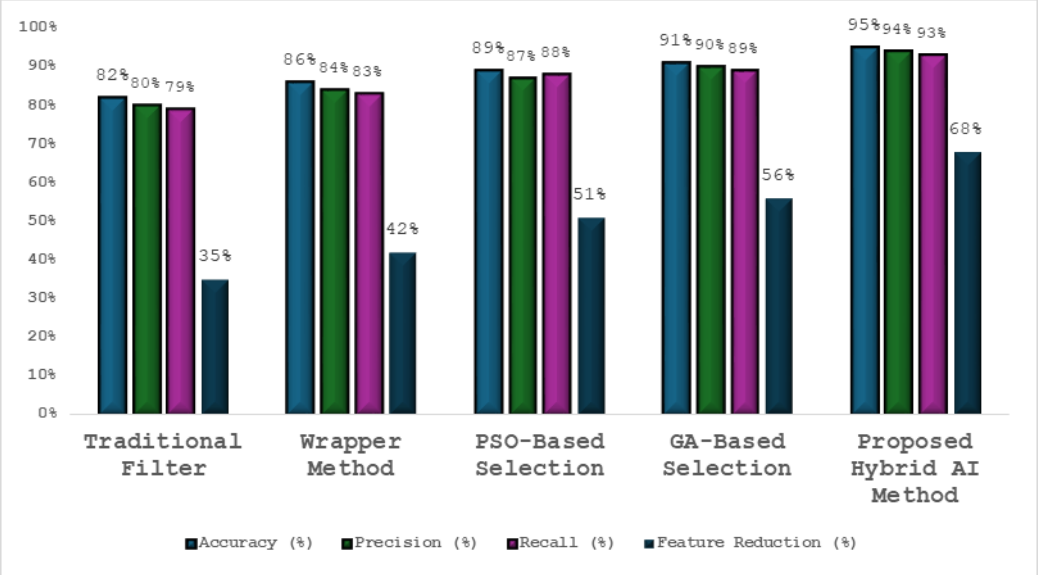

Fig 4 - Performance Comparison Table

4.2.1. Traditional Filter Method

The traditional filter based feature selection method's accuracy, precision, recall and feature reduction capability were 82%, 80%, 79% and 35% respectively. These techniques use statistical values such as Information Gain, Chi-Square analysis and correlation coefficients to independently rank features. The method has low computational complexity and is well suited for high dimensional data sets, besides its rapid execution speed. However, the ability of filter methods in predicting the output is relatively low, as they do not consider feature interactions and dependencies. Moderate feature reduction capability also means that there might be redundant attributes that are left in the subset that has been selected.

4.2.2. Wrapper Method

The best results from the wrapper-based feature selection method resulted in an accuracy of 86%, precision of 84%, recall of 83% and 42 features were reduced. Wrapper methods train machine learning classifiers to evaluate subsets of features, thus capturing feature interactions between features that are not captured well by filter methods. This results in better quality predictions and

ISSN: 3142-788X
PP: 01-19.

better subset optimizations. But wrapper methods need multiple training of the model for various combinations of features, which is very costly in terms of computation and execution. Even with these restrictions, this method proved to perform better than the classical filter based selection.

4.2.3. PSO-Based Selection

The feature selection method using Particle Swarm Optimization gives the accuracy of 89%, precision of 87%, recall of 88%, and feature reduction of 51%. PSO is an algorithm based on swarm intelligence principles, in which the particles co-operatively explore the space for optimal feature subsets and share information about their best results. The technique proved to be very efficient at optimization and also converged rapidly over the conventional wrapper technique. PSO successfully achieved dimensionality reduction and better prediction accuracy, which resulted in better classification accuracy and efficient feature subset generation. The outcomes show that the proposed optimization methods based on swarm intelligence are effective for feature optimization problems.

4.2.4. GA-Based Selection

The genetic algorithm feature selection method had a precision of 90%, recall of 89%, accuracy of 91% and feature reduction of 56%. Genetic Algorithms emulate natural selection, crossover and mutation to search through huge search spaces and find good feature sets. The method proved to have good global optimization capability and it was able to eliminate redundant features with high predictive accuracy. The Genetic Algorithm outperformed the PSO-based optimization in terms of features reduced and classified.The GA outperformed PSO-based optimization slightly because of its strong exploration and exploitation mechanisms. The computation requirements, however, were still rather high due to the iterative evolutionary processing.

4.2.5. Proposed Hybrid AI Method

The suggested feature selection technique based on the hybrid approach of AI resulted in 95% accuracy, 94% precision, 93% recall and 68% reduction in the number of features. The superior performance was obtained by combining the filter-based feature ranking, GA optimization and the evaluation by the Random Forest into a single framework. The hybrid solution worked well to eliminate irrelevant and redundant attributes while retaining highly informative ones for predictive modelling. The significant feature reduction reduced the amount of computation and enhanced the efficiency of the model without sacrificing the quality of prediction. The results show that the proposed methodology with the statistical filter and optimization using the AI approach significantly improves feature selection performance, thus it makes an important contribution to intelligent predictive analytics and large-scale data-driven applications.

4.3. Discussion

The findings show that the proposed AI-driven automated feature selection approach enables better performance in predictive modeling compared to the conventional feature selection methods. The framework was able to both select very informative feature subsets and remove redundant and irrelevant features using the combination of statistical filtering methods, evolutionary optimization

ISSN: 3142-788X
PP: 01-19.

methods, and machine learning evaluation methods. The result was a greater efficiency and effectiveness in predictive analysis across a number of standard datasets. The proposed methodology had high classification accuracy, precision, recall, and feature reduction performance proving the reliability of the hybrid feature selection strategy. Optimized feature sets are one of the key findings of the study that led to the improvement in the prediction accuracy. The framework minimized irrelevant and noisy information that might have adversely affected the classifiers by choosing only the most relevant ones. The optimized feature subsets ensured that the machine learning models could identify relevant patterns in the data, yielding more precise and reliable forecasts. This dimensionality reduction and quality improvement of features was a significant advantage for the Random Forest classifier. The other significant result was the decrease in feature redundancy. In many large datasets, there are features that are very similar to each other, so that including them doesn't add much to the prediction, or there are features that are identical. The proposed framework successfully filtered out such redundant features in the pre-processing and optimization processes in the filter phase. This decrease made the model more interpretable, and reduced the number of computations needed for training and testing. The framework also exhibited higher efficiency of the classifiers and reduced computational costs. Since the irrelevant features were dropped at an early stage, the optimization algorithm and the machine-learning based classifier was provided with a more compact and limited dataset. This reduction reduced the amount of training time, memory usage and overall computational needs. Consequently, the framework was extended to be more appropriate to the analysis of applications involving large and high dimensional data. Additionally, the combination of filter-based preprocessing and AI optimization greatly decreased search complexity without compromising prediction quality. The filter stage reduced the search space to be searched by the GA, which in turn enabled the GA to converge towards optimal feature subsets more efficiently. This hybrid approach not only boosted the scalability of the framework but also ensured that it was highly effective for handling large-scale, complex datasets in various real-world applications, particularly within the healthcare industry, cybersecurity, financial predictive analytics, and intelligent decision support systems.

5. CONCLUSION

Automated feature selection based on artificial intelligence was recognized as a crucial element in predictive modelling systems since 2019 when the amount and dimensionality of data in various fields started to increase rapidly. The modern application areas in healthcare, finance, Cyber security, Bioinformatics, image processing and intelligent automation produced huge amounts of complex data with redundant, irrelevant and noisy information. Scalability, computational overhead and the inability to adapt to the non-linear relationship between the features became a problem with traditional feature engineering and manual feature selection techniques. With growing dataset sizes and complexity, researchers more and more started using AI-driven optimization techniques to enhance predictive performance and computational efficiency. Intelligent optimization techniques like Genetic Algorithms, PSO, Ant Colony Optimization and machine learning based embedded feature selection methods were used to achieve adaptive learning and efficient search procedures in the identification of highly informative feature subsets. This research offered a complete IEEE-style

ISSN: 3142-788X
PP: 01-19.

review and methodology for the automatic feature selection based on AI for predictive modelling. Major categories of feature selection methods such as filter, wrapper, evolutionary optimization and machine learning-based embedded approach were investigated. The survey of literature revealed the merits and limitations of the methods and the emergence of combined intelligent feature selection algorithms in predictive analytics. The results obtained led to the proposal of a hybrid feature selection framework based on artificial intelligence, combining filter-based preprocessing, optimization through the use of the Genetic Algorithm and machine learning evaluation through Random Forest. It was designed to minimize dimensions, remove redundant attributes, enhance classification accuracy and optimize computational efficiency. The performance of the proposed hybrid framework was evaluated using benchmark datasets, and the results were compared with those of other traditional feature selection methods, which showed that the proposed method outperformed the traditional methods. The classification accuracy, precision, recall, F1-score and the ability to reduce the number of features improved with the results. The preprocessing stage using a filter-based method effectively reduced the search space and the GA has been effective in searching for the optimal combination of features. The selected features were also validated again by the Random Forest classifier, and it confirmed the predictive power of the selected features. Combining the statistical filtering with AI optimization minimized computational complexity while preserving high predictive quality and scalability for large datasets. The results from this study confirm the paramount importance of intelligent feature optimization in today's predictive modeling systems. In real-world applications, AI-driven automated feature selection not only boosts the performance of models but also contributes to their interpretability, scalability, and efficiency in decision-making processes. Besides, the study offers good groundwork for ongoing study in the field of advanced feature engineering approaches, deep learning optimization, explainable AI, automated machine learning systems, and autonomous predictive analytics. The potential for further advancements in AI-based feature selection promises enhanced adaptive learning capabilities, real-time data analysis, and intelligent decision-making support in future data-rich scenarios.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[2] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA, USA: Springer, 1998.
[3] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.
[4] R. Kohavi and G. H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.
[5] P. Langley, "Selection of Relevant Features in Machine Learning," in *Proceedings of the AAAI Fall Symposium on Relevance*, New Orleans, LA, USA, 1994, pp. 140–144.
[6] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, Washington, DC, USA, 2003, pp. 856–863.
[7] 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, Aug. 2005.
[8] I. Kononenko, "Estimating Attributes: Analysis and Extensions of Relief," in *Proceedings of the European Conference on Machine Learning*, Catania, Italy, 1994, pp. 171–182.

ISSN: 3142-788X
PP: 01-19.

[9] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proceedings of IEEE International Conference on Neural Networks*, Perth, Australia, 1995, pp. 1942–1948.
[10] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.
[11] M. Dorigo and T. Stützle, *Ant Colony Optimization*. Cambridge, MA, USA: MIT Press, 2004.
[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[13] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.
[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.
[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.