Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

High Dimensional Datasets Optimization handling by Wrapper Sequential Feature Selection in Forward Mode - A Comparative Survey

Ravi Shankar Mishra

Professor and Head, Department of Electronics and Communication Engineering, Sagar Institute of Science &Technology (SISTec), Gandhi Nagar Campus, Bhopal, India

E-mail: ravims3@gmail.com

Abstract

High-quality data might be difficult to be produced when there is a large quantity of information in a single educational dataset. Researchers in the field of educational data mining have recently begun to rely more and more on data mining methodologies in their investigations. However, instead of undertaking feature selection methods, many research investigations have focused on picking appropriate learning algorithms. Since these datasets are computationally complicated, they need a lot of computing time for categorization. This article examines the use of wrapper approaches for the purpose of managing high-dimensional datasets in order to pick appropriate features for a machine learning approach. This study then suggests a strategy for improving the quality of student or educational datasets. For future investigations, the suggested framework that utilizes filter and wrapper-based approaches may be used for many medical and industrial datasets.

Keywords: Feature selection, wrapper technique, data mining, high dimensional dataset, Forward domain

1. Introduction

Many sensors' data may now be accessed through the Internet of Things (IoT). Various applications take advantage of the massive volumes of data that are now readily accessible. Large datasets containing characteristics gathered from a variety of sensors may have varying degrees of correlation with the objective, depending on the feature combinations or several perspectives used. As an example, a broad variety of sensor types such as pressure sensors, thermistors as well as potentiometers are often employed to collect

DOI: https://doi.org/10.36548/jucct.2022.3.006
Received: 30.07.2022, received in revised form: 21.08.2022, accepted: 06.09.2022, published: 16.09.2022

data in the automobile industry. Various feature subsets interact with the target in different ways because of the heterogeneity of the sources. To put it in another way, each feature subset has a unique relationship to the objective. As an example, the attributes of an automobile's air system correspond differently to the mark than those of its fuel system. When opposed to using a single big feature subset, considering numerous perspectives of the dataset may enhance prediction quality [1]. As a result, an algorithm is required to take advantage of the correlations that exist among the different types of data. By using such a framework, it is possible to better understand and forecast the complex relationships in high-dimensional feature space. These interactions can only be captured if various perspectives of the feature space are compared.

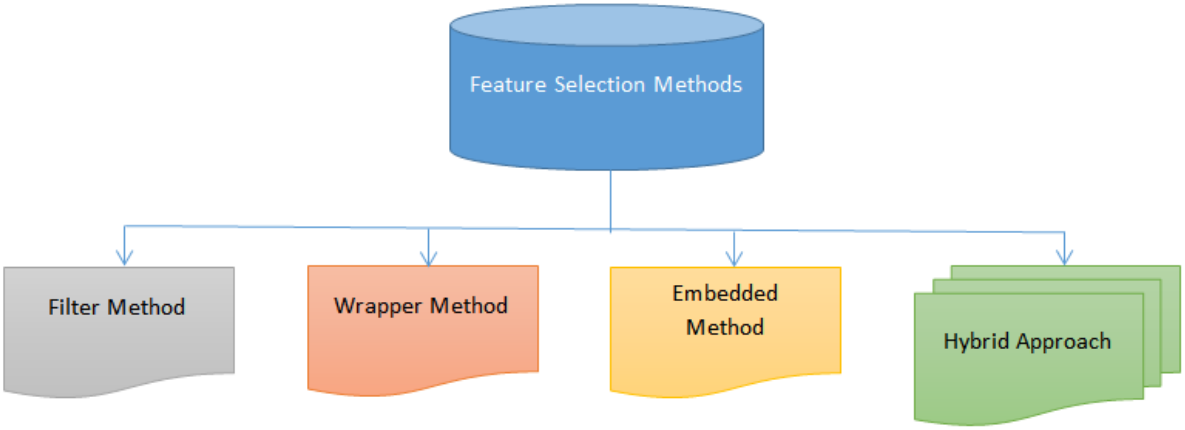

Figure 1. Overview of Feature selection Methods

With multivariate correlation datasets, features may have a low relevance to the aim, even though they are correlated with a large number of variables. In conjunction with other dataset aspects, these unique traits may have a profound impact on the target prediction. Higher-order interactions between components may be seen here. When doing a bivariate correlation analysis in certain instances, the results might be misinterpreted. As a result, improving prediction accuracy necessitates assessing its importance based on higher-order interactions. Suppose, considering the responsibility of determining the health of a certain component in an airplane, higher-order interactions occur in such a system because of the presence of characteristics that reflect the individual components and environmental variables. An evaluation of an element's significance without taking into account its relationships with other features is a faulty approach [2, 3]. The goal of this work is to describe an algorithm that measures a feature's value based on how it interacts with other features.

1.1 Mixed Dataset

Each state or category represents a qualitative quality i.e., the categories cannot be used as numerical values. This means that they must be assessed differently from the continuous feature values in terms of their ability to forecast the target. It might be difficult to compare continuous and categorical characteristics with various correlation functions since they are not immediately comparable. A single criteria function that is independent of the data type is considered in order to establish the importance of certain mixed characteristics.

1.2 Feature Optimization Approach

The feature optimization techniques come under data mining, choosing relevant subsets of characteristics from the original sets [4]. Data that is uninteresting, redundant, or noisy are omitted from this feature selection. The ultimate performance will be tuned up for the better prediction with many data mining algorithms when the number of components is decreased [5]. Underlying features are chosen by applying a stopping condition, and then verifying the results [6-8].

To begin, a search approach is used to generate new candidates for feature subsets during subset creation. The potential subsets are then examined and compared to prior subsets using the evaluation criteria. If the newer subset has a superior feature, the older can be changed; that means, it is updating repeatedly. Finally, a basic data validation process is required to ensure that the best-selected feature subsets are accurate. When it comes to feature selection, Fig. 1 shows the progression.

1.3 Research Motivation

The flaws of research paper become apparent via a close examination of its output. Exploration is the driving force behind feature selection. These characteristics are used for pre-processing, computational reasons, and storage and transfer in this research. More training samples are required to attain a higher classifier performance. In order to deal with classification issues, feature selection is critical; it decreases the data's noise and features. These procedures benefit from the improved speed, simplicity, inventiveness, conceptual clarity, and classification accuracy that come from the use of basic rules. To summarize, the concept of dimensionality may aid in several aspects of data mining, machine learning, pattern recognition, and so on.

2. Literature Survey

The proposed new concepts of Improved Dependency Classes (IDC) were framed by Raza and co-workers [9]. Consequently, IDC replaced dependency measurement and increased classification performance by decreasing the time it takes to run and the amount of runtime RAM it uses.

RLR using a Support Vector Machine (SVM) as the selection mechanism has been suggested by Guo and colleagues [10]. Because it provides a globally optimum solution with linear complexity, this feature selection technique outperformed others.

The approach for selecting Wrapper features takes advantage of an evolutionary search strategy. It comes to algorithms approach that may be used in many optimization algorithms.

Using the Continuous Feature Selection technique, Sharma et al. [11] have been able to process features with a minimum size of 10 at a time, while also boosting the level given by the components. Parts were removed from the system in stages until the stop condition has been achieved at the end of each step.

According to Kang et al. [12], random forward search is used with pertinent characteristics to achieve global optimization. SFS and SBE are two further sequential selection procedures that were used to arrive at the final model, which is called a "bold sequential selection."

2.1 Mixed datasets provide a challenge

This kind of data is often seen in current datasets. By discretizing the consistent characteristics, traditional techniques turn a dataset with heterogeneous data types into a homogenous data type.

These methods of data transformation eliminate the need to deal with various data types individually. The discretization technique used has a significant impact on the usefulness of the chosen characteristics, but it also causes information to be lost. Pre-processing by encoding the categories with numeric values is another basic approach. As the categories indicate a qualitative quality, such encodings are meaningless, and assigning random numbers may lead to inaccurate results, when measuring the distance between two category features that have been differentially coded in relation to the target [13]. It is thus

essential to conduct multivariate relevance and redundancy estimates on big datasets without the need of extra pre-processing methods (such as discretization and encoding).

2.2 Summary of the Problem Statement

An exploration of the search space and an exploitation of the best answer are the two factors that hybrid algorithms use to make their decisions. A flaw in the exploration approach of the native sine cosine algorithm results in poor performance in the search space. Hybrid algorithms, on the other hand, may be improved or modified using hybridization approaches to create new hybrid algorithms that balance exploration and exploitation of the search area. The attempts to develop a model based on hybridization techniques to solve feature selection difficulties by decreasing the number of features and weak relevant and irrelevant characteristics, are motivated by this stimulus.

3. Handling of High dimensional Datasets

The selection of descriptors that best characterize a particular domain is referred to as "feature selection". Many studies in domains where datasets with hundreds of thousands of variables are available, focus on feature selection. The useful feature subset may be mined, as well as the classification accuracy and speed that feature selection can provide for high-dimensional data selection.

3.1 Wrapper Approach

Classification of feature selection techniques has traditionally been done in two ways. Statistical features of the variables are used to filter out those that aren't very useful, which is the first strategy [14]. This is performed before any categorization method is used. Wrapper approaches, which are computationally intensive yet frequently based on the evaluation strategy, are the second solution for this problem. Many of the following paradigms are wrapper, filter, hybrid, embedded, and unsupervised selection for high dimensional dataset. Features may be returned as weights, ranks, or subsets as an output from feature selection algorithms. Each paradigm has its own set of principles that are discussed below.

3.2 Wrapper Sequential Forward Selection (WSFS)

The feature subset will be reduced in dimensionality using the wrapper technique. In the wrapper technique, a search strategy is used to find the most accurate feature collection. It is possible to categorize these search methods into three groups: sequential, complete, and

random. To conduct this research, a sequential search approach has been used. According to the argument, this method is easier to apply since it is less computationally intensive. Furthermore, it is capable of swiftly locating characteristics and returning relevant results in the search space [15]. There are many prevalent algorithms available in sequential search approach [16]. However, this research will solely focus on the wrapper approach's Sequential Forward Selection (SFS). Starting with a blank slate, an SFS aims to find the most relevant results. Then, the finest bits will be picked for each iteration and integrated with the current characteristics that have been selected in a blank set. Finally, the iteration will be halted whenever it reaches a point where the components exhibit no progress [17].

3.2.1 Forward Domain Approach

The filter method has a number of benefits, including being quick, scalable, and independent of categorization. This has the drawback of having a low degree of categorization accuracy [18]. Because of the Wrapper approach, the search process for a feature subset is interactive, and the reliance of features is considered [19]. However, its drawback is that it requires a lot of time to build a classifier. The hybrid filter algorithm, which combines the first ‘n’ characteristics of several filter algorithms, is the most often used filter algorithm.

3.2.2 Wrapper Method

Learning is used as a black box in the wrapper to score subset features. An induction approach is used to assess the value of feature subsets in wrapper schemes. Wrapper techniques are justified by the fact that the induction method that uses the feature subset in the end should produce a higher accuracy estimate than a separate measure with an altogether different inductive bias. When a feature's value depends on the relevance of another feature, it is known as a "interacting feature". Components, on the other hand, become redundant when the values they represent are contingent on the values of other characteristics.

It may see a shift toward a wrapper-based strategy for choosing important characteristics in the educational data mining research. The hybrid technique is a blend of filter- and wrapper- based approaches. Some researches in the medical field have used the embedded approach for feature selection. However, there is a dearth of study in this field at the university level. As a result, researchers would have a great opportunity to teach students about the hybrid approach of feature selection.

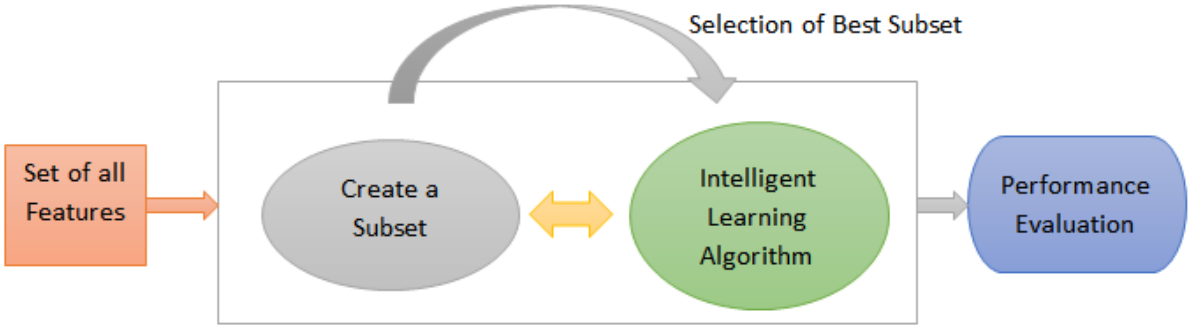

Figure 2. Steps of Feature Selection Process

3.2.3 Extensive Wrapper Paradigm

Wrapper-based techniques estimate the prediction error to determine the usefulness of a feature. In other words, a classification or regression technique is utilized as the cost function for the feature selection job. The wrapper-based processes are computationally intensive, yet they surpass filter-based methods when it comes to quality predictions. Over-fitting is a common problem with wrapper techniques, which makes them difficult to generalize. As a result, different classification or regression procedure is used to examine a subset of characteristics results in reduced prediction accuracy [20, 21].

4. Comparative Observations

This comparison of algorithm's efficiency was described on a variety of publicly accessible datasets, and the results were quite promising. There are many lists on many datasets that were used to compile this report. Machine learning researchers often utilize these datasets to test their algorithms against one other in the UCI Repository of Machine Learning Databases. There are more than 2000 characteristics and 800 occurrences in the dataset utilized in this article.

Table 1. Number of Datasets for Evaluation

S.No	Datasets	Filters	Wrapper	Embedded
1	Medical	1	1	1
2	Educational	3	3	3
3	Cylinder Bands	2	2	2
4	Balance Scale	1	1	1

Comparing the suggested technique to other feature selection methods, such as filter, wrapper, embedded and hybrid, reveals how much more accurate it is. Finally, the comparison of different feature selection approaches is shown in the below table.

Table 2. Performance Measures Comparison

S.No	Performance Measures	Filter Method	Wrapper Method	Embedded Method	Hybrid Approach
1	Combination of Algorithm	Cannot combine with ML	Combine with ML	Useful for redundancy analysis	Incorporation with convenience
2	Computation Time	Fastest	Faster	Moderate Speed	Slow
3	Over-fitting Problem	Less Prone	High Chance	Moderate	Random
4	Prediction Rate	Higher	High	Low	High
5	Accuracy	Higher	High	Less	High
6	Overall Efficiency	Good	Good	Moderate	Random

Comparing these approaches with the existing methods on real-world data with a high dimensionality, shows that it is efficient and effective. For predicting or analysing class labels that aren't known, this classifier may be utilized. The dataset, which comprises characteristics and examples, is the most important factor in determining how well the classifier performs. In addition, a correlation-based feature selection approach, a novel framework has been introduced for efficient optimized feature selection based on significance and redundancy analysis.

5. Conclusion

Data preparation is an efficient approach for resolving small sample classification issues. Feature selection, a crucial pre-processing tool, may remove duplicate features to enhance classification performance and discover significant aspects associated with classification issues. It is also possible to increase the performance of feature selection algorithms by combining their advantages and disadvantages, since each approach has its advantages and disadvantages. An experiment will be carried out in the future to show that the suggested framework can enhance the quality of students' datasets and boost their

prediction accuracy. For classification problems involving high-dimensional datasets, additional hybrid and intelligent feature selection strategies would be investigated in the future.

References

[1] Yi Yang, Wei Liu, Tingting Zeng, Linhan Guo, Yong Qin, Xue Wang, "An Improved Stacking Model for Equipment Spare Parts Demand Forecasting Based on Scenario Analysis", Scientific Programming, vol.2022, pp.1, 2022.

[2] Shiwei Liu, Decebal Constantin Mocanu, Amarsagar Reddy Ramapuram Matavalam, Yulong Pei, Mykola Pechenizkiy, "Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware", Neural Computing and Applications, vol.33, no.7, pp.2589, 2021.

[3] Edmundo Bonilla-Huerta, Alberto Hernandez-Montiel, Roberto Morales-Caporal, and Marco Arjona-Lopez, "Hybrid Framework Using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data," IEEE/ACM Transactions on Computational Biology And Bioinformatics January/February, vol.13(1), pp. 12-26, 2016.

[4] Kung-Jeng Wang, Angelia Melani Adrian, Kun-Huang Chen, Kung-Min Wang, "An improved electromagnetismlike mechanism algorithm and its application to the prediction of diabetes mellitus," Journal of Biomedical Informatics, vol. 54, pp. 220–229, 2015.

[5] Joaquin Abellan, Carlos J. Mantas, Javier G. Castellano, Serafin Moral-Garcia, "Increasing diversity in random forest learning algorithm via imprecise probabilities", Expert Systems With Applications, vol. 97, pp. 228–243, 2018.

[6] Messaouda Nekkaa, and Dalila Boughaci, "A memetic algorithm with support vector machine for feature selection and classification," Memetic Comput., vol. 7, pp. 59–73, 2015.

[7] S.Sasikala, S. Appavu alias Balamurugan, and S. Geetha, "A novel adaptive feature selector for supervised classification," Information Processing Letters, vol. 117, pp. 25 – 34, 2017.

[8] Aiguo Wang, Ning An, Guilin Chen, Lian Li, and Gil Alterovitz, "Accelerating wrapper-based feature selection with K-nearest-neighbour," Knowl.-Based Syst., vol. 83, pp. 81–91, 2015.

[9] Muhammad Summair Raza and Usman Qamar. An incremental dependency calculation technique for feature selection using rough sets. Information Sciences, 343-344:41–65, 2016.

[10] Shun Guo, Donghui Guo, Lifei Chen, and Qingshan Jiang. A centroid-based gene selection method for microarray data classification. Journal of Theoretical Biology, 400:32–41, 2016.

[11] A Sharma, S Imoto, and S Miyano. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 9(3):754–764, 2012.

[12] Seokho Kang, Dongil Kim, and Sungzoon Cho. Efficient feature selection-based on random forward search for virtual metrology modeling. IEEE Transactions on Semiconductor Manufacturing, PP(99):1–1, 2016.

[13] E. Emary, Hossam M. Zawbaa, and Aboul Ella Hassanien. Binary ant lion approaches for feature selection. Neurocomputing, 213:54–65, 2016.

[14] Mary Walowe Mwadulo. A review on feature selection methods for classification tasks. International Journal of Computer Applications Technology and Research, 5(6):395–402, 2016.

[15] Muhammad Summair Raza and Usman Qamar. An incremental dependency calculation technique for feature selection using rough sets. Information Sciences, 343-344:41–65, 2016.

[16] Shun Guo, Donghui Guo, Lifei Chen, and Qingshan Jiang. A centroid-based gene selection method for microarray data classification. Journal of Theoretical Biology, 400:32–41, 2016.

[17] Jain A, Jain V (2022) Sentiment classification using hybrid feature selection and ensemble classifier. J Intell Fuzzy Syst 42(2):659–668.

[18] Abasabadi S, Nematzadeh H, Motameni H, Akbari E (2021) Automatic ensemble feature selection using fast non-dominated sorting. Inform Syst 100:101760

[19] Maleki N, Zeinali Y, Niaki ST (2021) A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. Expert Syst Appl 164:113981

[20] Seijo-Pardo B, Bolon-Canedo V, Alonso-Betanzos A (2017) Testing different ensemble configurations for feature selection. Neural Process Lett 46(3):857–880.

[21] Li M, Vanberkel P, Zhong X (2022) Predicting ambulance offload delay using a hybrid decision tree model. Socioecon Plann Sci 1(80):101146.

Author's biography

Ravi Shankar Mishra is currently working as a Professor and Head in the Department of Electronics and Communication Engineering, Sagar Institute of Science &Technology (SISTec), Gandhi Nagar Campus, Bhopal, India. His area of research includes VLSI and antenna design.