A META-ANALYTICAL REVIEW OF META-HEURISTIC ALGORITHMS FOR FEATURE SELECTION IN HIGH-DIMENSIONAL DATASETS

THATIKONDA SOMASHEKAR
Research Scholar
University College of Engineering,
Department of Computer science and Engineering, Osmanai University.
soma.ts@gmail.com

Prof P. PREMCHAND
Dept computer science and engineering.
Osmania university.
ppremchand@gmail.com

ABSTRACT

As data mining continues to progress, feature selection on high-dimensional datasets (HDD) utilizing meta-heuristic methods is becoming more common. However, because there are differing opinions on the subject, it is still challenging to determine the threshold of features in a dataset that should be classified as HDD. Feature selection techniques effectively produce highly computationally efficient optimal solutions. Large feature sets of data are very prevalent, particularly in the field of bioinformatics and other sophisticated technical processes. The goal was to ascertain the threshold for a number of features to be HDD, after which the trend or possible FS technique for HDD would be identified, along with the most favored classifiers and meta-heuristic algorithms for both wrapper-based and filter-based FS methods to analyze HDD. Despite being used for classification tasks in recent decades, traditional FS approaches are unable to properly decrease the high dimensionality of the text feature space, which results in ineffective prediction models. The literature is still filled with both clear and ambiguous results about the employment of efficient techniques, which, if not carried out correctly, change precision, the viability of real-world application, and the overall effectiveness of the predictive model. These features reduce the quality of the entire feature set by obscuring the important information that important features convey.

Keywords: *Meta-heuristic algorithms, Feature selection, high-dimensional datasets (HDD), wrapper-based and filter-based, FS techniques.*

INTRODUCTION

Feature selection techniques effectively produce highly computationally efficient optimal solutions. Because all characteristics might not be useful, the traditional approaches become wasteful. Instead, a "subset of features" may need to be extracted using feature selection techniques that can improve accuracy. Unquestionably, the growth of datasets is no longer a novel occurrence given the speed at which science and technology are developing. Over time, datasets are growing in size and dimensionality. The form that a dataset is often represented in is necessary to further understand the meaning of high-dimensional datasets (HDDs). Usually, datasets are viewed as matrices, where the column denotes features and the row denotes instances. HDDs are datasets that contain a large number of characteristics. Low categorization performance could be caused by these unimportant features. Therefore, feature selection is necessary to reduce dimensionality and improve HDD classification performance. The process of choosing the most significant features is known as feature selection. Three ways can be used to further classify feature selection: wrapper-based, embedded-based, and filter-based. While embedded-based feature selection occurs during model training in the machine learning process, wrapper-based feature selection leverages the strength of the base classifiers to identify

the best features in a dataset. Because the classifiers interfere with the feature selection process, both wrapper-based and embedded-based approaches have longer execution times. While filter-based approaches are quick to calculate HDDs and can be used with any type of predictive model, embedded-based approaches require specific predictive models. It is evident from these three approaches that filter-based feature selection is quicker than wrapper-based methods and more applicable to HDDs since it chooses a subset of features without the need for a learning mechanism. Furthermore, of all feature selection techniques, filter-based approaches are the least difficult and work with a variety of datasets, including HDDs.

LITERATURE REVIEW

Amparo Alonso-Betanzos (2019) Because it is predicated on the idea that combining the output of numerous models is preferable to employing a single model and typically yields positive results, ensemble learning is a popular area of machine learning. Although it has typically been used for classification, it can also be applied to enhance other fields, such feature selection. The primary objective of feature selection is to increase classification accuracy by choosing the elements that are pertinent to an issue and eliminating those that are superfluous or unnecessary. In addition to evaluating recent developments and offering commentary on upcoming trends that still need to be addressed, we also give the reader an overview of the fundamental ideas required to construct an ensemble for feature selection.

Marco F. Duarte (2018) By removing duplicate and unnecessary features from high-dimensional data, feature selection is a dimensionality reduction strategy that chooses a subset of representative features. Due to its exceptional performance when compared to conventional feature selection techniques that disregard feature correlation, feature selection in conjunction with sparse learning has garnered a lot of interest lately. By applying a sparsity constraint to the transformation matrix, these works first map data onto a low-dimensional subspace before choosing features. However, because the underlying correlation structures of data are frequently non-linear, their design limitation to linear data transformation may be a disadvantage. We suggest an autoencoder-based unsupervised feature selection method that uses a single-layer autoencoder for a joint framework of feature selection and manifold learning in order to take advantage of a more complex embedding. More precisely, as in earlier work, we impose column sparsity on the weight matrix that connects the input layer with the hidden layer. In order to preserve local data geometry from the original data space to the low-dimensional feature space, we also use spectral graph analysis on the projected data into the learning procedure.

Raca Todosijević (2017) Maximizing the minimal accumulative dispersion among the selected elements is the goal of the maximum min-sum dispersion issue. The problem is known to be strongly NP-hard. We provide a heuristic in this paper that shifts the objective functions of two distinct issues inside a variable neighborhood search framework. The key distinction is that this heuristic permits the use of various formulations of multiple optimization problems, even though it can be viewed as an expanded version of the variable formulation search technique that considers alternative formulations of a single problem. Here, we employ a different formulation of the max-sum type of the maximum diversity issue, which was

initially of the max-min type. The proposed method improves the best-known findings for the majority of examples in a shorter computing time, according to computational experiments conducted on the benchmark instances used in the literature.

Gang Chen (2017) Enzymatic digestion was used to separate cells from meniscal debris from individuals who had meniscal injury. The cells were then grown in vitro to the third passage and examined for growth and appearance. The immunophenotype and tri-lineage differentiation potential of third-passage cultures were also examined. Following four to six days of development, the meniscal debris cells took on a lengthy, fusiform shape and stuck to the culture dish's plastic walls. Colonies and clusters of cells were seen after 8–10 days. Third-passage cells proliferated well and had a consistent shape. Cells exhibited positive staining for Alizarin Red, alkaline phosphatase activity for the formation of mineralization nodes and early osteogenic marker, Oil Red O for lipid vacuoles, Toluidine blue, and Col II immunohistochemical staining for the cartilage-specific matrix after cultures were induced to differentiate into bone, adipose, and cartilage. The mesenchymal stem cells that were extracted and cultivated from meniscal debris were discovered to be able to differentiate into three distinct lineages. MSCs were found in meniscal debris and demonstrated their capacity for culture and differentiation, which paved the way for research into their potential for meniscal regeneration.

R Cheng (2016) Many machine learning problems, like classification, require a high number of input features to be solved. But not every feature is necessary to solve the issue, and occasionally adding features that are not important can make learning less effective. Please review the article title revision. As a result, feature selection—the process of choosing the most pertinent features—is crucial. To identify a subset of the most crucial features for completing a given machine learning job, numerous feature selection techniques have been created, such as particle swarm optimization (PSO) algorithms and evolutionary algorithms. However, as the number of features significantly rises, the usefulness of PSO for feature selection is diminished because the traditional PSO performs poorly for large-scale optimization issues. In order to solve high-dimensional feature selection problems, we suggest using a relatively new PSO variant called the competitive swarm optimizer (CSO), which was created specifically for large-scale optimization. Furthermore, feature selection—which may be viewed as a combinatorial optimization problem—is carried out by the CSO, which was initially designed for continuous optimization.

Feature Selection (FS)

The process of choosing a relevant subset of features or qualities that are essential for completing the classification or clustering task is referred to as "feature selection." Working with the entire collection of features becomes practically challenging as the number of features increases, and the features that are available may have noise and redundancy. Consequently, feature selection becomes essential for carrying out highly significant data analysis. Although feature selection can help reduce the amount of features, it is a laborious process, particularly when there are a lot of features. Selecting essential features that are helpful for improving model performance is crucial since redundant and irrelevant features may negatively affect the model's performance. This will increase the prediction system's

speed and performance. The characteristics are often divided into:

i. **Relevant features:** The features that have significant influence on output and their role can't be assumed by rest of features
ii. **Irrelevant features:** The features that don't have any significant influence on output
iii. **Redundant features:** The features that can take the role of other features.

Selecting valuable features from a dataset is a challenging task, especially when we have to select a valuable subset from hundreds or thousands of features. This research aims to select a subset of valuable features for which the classifiers perform the best.

Trend in meta-heuristics

The meta-heuristics work very well when compared to exact search mechanisms since they are most appropriate for issues with high computing complexity and do not require searching the full search space. Furthermore, hybrid techniques are favored because they effectively strike a balance between exploration and exploitation, improving search algorithm performance. We proposed the hybrid ACOSA algorithm as a result of this. Effective and efficient search space exploration is a crucial component of the successful creation of a new meta-heuristic. Additionally, the search method needs to be smart enough to swiftly travel to unknown places and thoroughly investigate search space that produces high-quality solutions. This entails finding search locations that produce high-quality results fast and avoiding wasting time on areas that have already been thoroughly investigated or that produce subpar results.

High-dimensional data

There have been extensive studies in improving the performance of LDP in applications with high-dimensional data which include the use of various techniques including user partition, binary vector encoding, and other dimension reduction techniques such as the use of hash functions or matrix transformation. Although such techniques may reduce the impact of the high dimension of the data, they cause other problems such as an increase of error in construction and an increase in decoding complexity. this introduces another challenge in the design of LDP schemes, namely finding a good balance in the performance of such schemes in these metrics. Furthermore, solutions for statistical queries on multiple attributes are typically solved by the use of sampling since the majority of LDP schemes focus on single attribute queries. Sampling requires the users to be partitioned where each group of users may only report a part of the attributes. This causes the possibility that there is an insufficient number of reports in some of those attributes, causing the estimates for such attributes to have low statistical accuracy. How to effectively tackle all these issues and produce a more accurate estimate while preserving local differential privacy remains an open problem.

Data dimensionality

New scalable DL methods must be developed in order to process and analyze high-dimensional data. Existing techniques for data mining and machine learning are either computationally inefficient or do not scale well to high-dimensional data types like unstructured data and images. The amount of data dimensions tends to increase exponentially in terms of time and storage needs [228]. Analytics would become more challenging because to the

large datasets' high degree of dimensionality. Therefore, to solve this problem, sophisticated DL techniques are required.

Metaheuristic Algorithms

Optimization techniques that find the best (near-best) answer to optimization problems are known as metaheuristic algorithms. These algorithms are derivative-free approaches and, feature simplicity, flexibility and potential to avoid local optima. Metaheuristic algorithms behave stochastically, producing random solutions at the beginning of their optimization process. Unlike gradient search methods, it does not necessitate calculating the derivative of the search space. Because of their uncomplicated implementation and straightforward notion, metaheuristic algorithms are both adaptable and easy to use. The algorithms are easily adaptable to the specific scenario at hand. The primary characteristic of metaheuristic algorithms is their exceptional capacity to avoid premature convergence. Because algorithms behave stochastically, the methods function as a black box, avoiding local optima and effectively and efficiently exploring the search space. The two primary and crucial components of the algorithms—exploration and exploitation—are traded off. The algorithms extensively examine the promising search space during the exploration phase, and the local search of the promising area or areas discovered during the exploration phase is the source of exploitation. Electrical engineering (to determine the best way to generate power), industrial fields (job scheduling, transportation, vehicle routing, facility location), civil engineering (to design bridges and buildings), communication (radar design, networking), data mining (classification, prediction, clustering, system modeling), and other engineering and scientific problems are successfully solved with them.

RESEARCH METHODOLOGY

Analyzing and synthesizing previous research on the use of different meta-heuristic algorithms for feature selection in high-dimensional datasets is the goal of a meta-analytical review of meta-heuristic algorithms for feature selection in high-dimensional datasets. The main objective is to evaluate the performance of various meta-heuristic algorithms when feature selection is applied to high-dimensional datasets. Because these datasets usually contain more characteristics than samples, they provide special difficulties for machine learning algorithms. Because meta-heuristic approaches can handle complex search spaces and perform global optimization, they are widely used. Meta-heuristic algorithms' relative efficacy in feature selection tasks. Determine which algorithms perform better than others on a variety of datasets. To find pertinent papers, use scholarly resources like IEEE Xplore, SpringerLink, ScienceDirect, and Google Scholar. the particular parameters applied to each algorithm, like the inertia weight (for PSO) or the population size, iterations, or mutation rate (for GAs). Analyze subgroups according to many criteria, such as the kind of meta-heuristic method, the properties of the dataset, and the performance metrics employed. Use statistical tests like Egger's or Begg's test or graphical techniques like funnel plots to assess possible publication bias. Use established checklists or rating systems to evaluate the included studies' quality. The objective of this meta-analytical review of meta-heuristic algorithms for feature selection in high-dimensional datasets is to

clearly synthesize the body of existing research, spot trends in algorithm performance, and provide guidance for practitioners and researchers in the fields of feature selection and machine learning.

RESULTS AND DISCUSSIONS

We have taken publicly available datasets from LIBSVM/KDD/UCI repositories. We performed experiments on 7 datasets. The information of each dataset is summarized in Table 1. We used python for implementing the algorithm. Data is divided into "balanced" subsets and to control overfitting, we used "Extremely randomized trees (extra-trees) classifier" on the data sub-samples. The datasets used in this work are illustrated.

Table 1: Datasets used for experiments

Dataset	Repository	No. of instances	No. of features
Arrhythmia	UCI	453	277
Coil_2000	KDD	9,823	84
Ozone_level	UCI	2,535	73
Libras_move	UCI	359	91
Spectrometer	UCI	532	92
Scene	LIBSVM	2,408	293
Optical_digits	UCI	5,621	63

The above datasets are cautiously chosen and are selected to have different number of features, classes, data types and instances. The dataset comprises of 453 instances and 277 features. The data used in COIL_2000 dataset comprises of information of clients of Insurance Company. It contains 85 features that comprise of information of product usage and socio-demographic data that is obtained from zip area codes.

The dataset comprises of 73 features and 2,535 instances. For analysis of these movements "mapping operation" is performed by mapping curve F in representation with 91 features that represent coordinates of movement.

The spectrometer dataset comprises of 532 instances from IRAS-LRS database. The valuable flux information is contained in 48 red-band channels and 44 blue-band channels (92 spectral intensities). The IRAS-LRS program has observed high-intensity over 2 continuous spectral bands. It has 293 features that are numerical having value in the range of 0 to 1. The dataset comprises of 63 features and 5,621 instances.

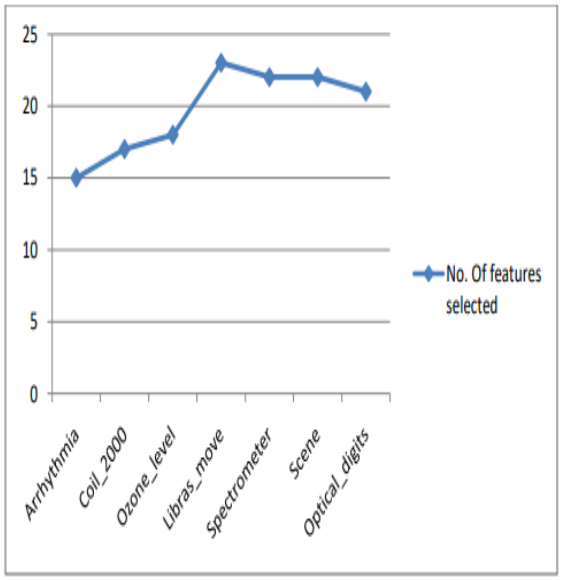
The proposed approach is tested on various high-dimensional datasets that are fetched from UCI and LIBSVM repositories. Table 2 represents the accuracy rate, the number of features selected for each dataset, AUROC, Recall, Precision and F1 Score.

Table 2: Number of features selected and results of classification performance (%)

Dataset	No. of selected features	Accuracy	AUROC	Recall	Precision	F1 score
Arrhythmia	14	94.08	98.11	95.35	93.37	94.82
Coil_2000	16	92.98	98.02	90.43	95.37	92.41
Ozone_level	17	93.56	96.15	93.30	93.30	88.97
Libras_mo	22	77.78	85.35	66.	84.27	75.3

ve				37		1
Spectrometer	21	97.49	98.75	97.40	96.46	95.73
Scene	21	85.44	72.39	75.95	73.49	73.27
Optical_digits	20	94.43	97.90	96.33	97.25	97.51

Since the HUS Boot and RUS Boot approaches have already been used for comparable datasets without carrying out the feature selection process, we contrasted our strategy with theirs. Based on the results, our method's accuracy was the greatest at 97.49 for the "Spectrometer" dataset, however it was much lower than the HUSBoot method for the Libras_move dataset.



Graph 1: Number of features selected for various datasets

Graph 1: The quantity of characteristics chosen for different datasets. The shows the accuracy percentage attained for different datasets. After doing feature selection on various datasets, we significantly decreased the amount of features, which further helped us arrive at the best solution and raise the accuracy value.

CONCLUSIONS

A promising method for feature selection that strikes a balance between exploration and exploitation is the use of meta-heuristic algorithms. These algorithms can effectively find a subset of features that increase the dataset's correctness and decrease its dimensionality. The experimental findings show that by efficiently scanning the feature space, this method produces encouraging results. The performance of machine learning models can be enhanced by using meta-heuristic algorithms as a pre-processing step to choose pertinent features. Create meta-heuristic algorithms that maximize several goals, including computational speed, feature count, and accuracy. The parameters utilized to compare the performance of the suggested technique include decreasing the number of selected features and optimizing the predicted accuracy. We have contrasted our findings with those of the "RUSBoost" and "HUSBoost" approaches, which did not carry out the feature selection task. It turns out that the suggested methodology produces higher accuracy than methods that make use of the entire set of attributes. We also contrasted the accuracy attained using alternative techniques. Feature selection is essential for improving model performance, particularly when dealing with high-dimensional datasets. Furthermore, guiding the practical implementation of these algorithms in real-world contexts would need understanding the trade-offs between accuracy and computing complexity as well as creating benchmarks for comparison.

REFERENCES

1. *Amparo Alonso-Betanzos (2019), "Ensembles for feature selection: A review and future trends", Information Fusion, issn:1566-2535, vol.52, pages.1-12. https://doi.org/10.1 016/j.inffus.2018.11.008*
2. *Bogdanović, M. (2011), "On Some Basic Concepts of Genetic Algorithms as a Meta-Heuristic Method for Solving of Optimization Problems", Journal of Software Engineering and Applications, issn:1945-3124, vol.4, pages.482-486.*
3. *N. Sánchez-Maroño (2012), "An ensemble of filters and classifiers for microarray data classification", Pattern Recognition, issn:0031-3203, vol.45, issue.1, pages.531-539. https://doi.org/10.1016/j.patcog.2011.06.0 06*
4. *R Cheng (2016), "Feature Selection for High Dimensional Classification using A Competitive Swarm Optimizer", Soft Computing, issn:1433-7479, https://doi.org/10. 1007/s00500-016-2385-6*
5. *Raca Todosijević (2017), "Solving the maximum min-sum dispersion by alternating formulations of two different problems", European Journal of Operational Research, issn:1872-6860, vol.260, issue.2, pages.444-459. https://doi.org/10.1016/j.ejor.2016. 12.039*
6. *Rashidi-Nejad, M. (2013), "Using UPFC and IPFC Devices Located by a Hybrid Meta-Heuristic Approach to Congestion Relief", Energy and Power Engineering, issn:1947-3818, vol.5, pages.474-480.*
7. *Vyas, O. (2014), "A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty", Journal of Data Analysis and Information Processing, issn:2327-7203, vol.2, pages.95-105.*
8. *Wang, Y. (2019), "Different Feature Selection of Soil Attributes Influenced Clustering Performance on Soil Datasets", International Journal of Geosciences, issn:2156-8367, vol.10, pages.919-929.*
9. *Marco F. Duarte (2018), "Graph autoencoder-based unsupervised feature selection with broad and local data structure preservation", Neurocomputing, issn:0925-2312, vol.312, pages.310-323.https://doi.org/10.1016/j.neucom.2018 .05.117*
10. *Gang Chen (2017), "Isolation, characterization and multipotent differentiation of mesenchymal stem cells derived from meniscal debris", Asia-Pacific Journal of Sports Medicine, Arthroscopy, Rehabilitation and Technology, issn:2214-6873, vol.9, pages.94-95. https://doi.org/10.1016/j.asmart.2017.05.2 12*