Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-1 Issue-5, December 2021

Multi-Objective Optimization Based Feature Selection Algorithms for Big Data Analytics: A Review

Aakriti Shukla, Damodar Prasad Tiwari

Abstract*: Dimension reduction or feature selection is thought to be the backbone of big data applications in order to improve performance. Many scholars have shifted their attention in recent years to data science and analysis for real-time applications using big data integration. It takes a long time for humans to interact with big data. As a result, while handling high workload in a distributed system, it is necessary to make feature selection elastic and scalable. In this study, a survey of alternative optimizing techniques for feature selection are presented, as well as an analytical result analysis of their limits. This study contributes to the development of a method for improving the efficiency of feature selection in big complicated data sets.

***Keywords*: Big Data, Feature Selection, Optimization, Data Mining.**

I. INTRODUCTION

Categorization is an important job in data mining and machine learning that divides every example in a data collection into various groups based on the information provided by its attributes. Without prior knowledge, determining which traits are useful is challenging. As a result, the data collection typically includes a huge number of functions, both relevant and unrelated [1]. Many important applications have complex pattern categorization or modeling tasks, which necessitate the use of feature selection approaches to reduce complexity and eliminate repetitive, irrelevant features or noisy dominated.

Feature selection is a difficult and computationally intensive operation due to two key problems. The first point of concern is the intricacy of feature interaction. The 2nd point of issue is the high dimensional space; the overall amount of solutions accessible for a database with a huge dataset makes selection of features difficult [2]. The findings of published efforts in the area of enhancement suggested solving a problem using swarm intelligence and evolutionary approaches, that began as static optimization and reveal a pseudo environments-based technique that can resolve and enhance an integrated multi issue to decrease or improve the outcome features. Multi-objective optimization issues [3] are a type of optimization problem that arises in real-world applications and has many competing goals. The implementation of feature selection as a multi-objective optimization process can bring certain benefits whether the classification approach is supervised or unsupervised. Supervised classification approaches are used to improve classifiers performance while lowering the amount of data, as large feature sets might lead to overfitting. As a consequence, a multi-objective optimization strategy that sufficiently combines classifiers performance and attribute quantity provides a reasonable formulation of this problem. Problems with many goals are referred to as multi-objective optimization (MOO). Engineering, social studies, economics, agriculture, aviation, automobile and Math are all examples of fields where this type of challenge can be encountered. The multi-objective evolutionary algorithm is a stochastic optimization technique (MOEA). MOEAs, like other optimization algorithms, are used to identify optimal Pareto solutions for specific issues, but they differ from community solutions. The MOEA's optimization method is remarkably similar to that of evolutionary algorithms, with the exception of the usage of a dominance relationship. Scalarization and the Pareto technique are two ways of tackling the MOO issue. The Pareto technique is used to develop a compromise solution (tradeoff) that can be shown in the form of a Pareto optimal front (POF) end if the intended solutions and performance measurements are separate. Furthermore, the Scalarization approach is included in the evolutionary algorithm as part of the performance measurements that constitute a scalar function. When the proper decisions must be made when a trade-off must be made between two goals that are normally in conflict, multi-objective challenges arise. Multi-objective optimization is compensated by objective functions that attenuate or optimize various conflicts.

II. LITERATURE REVIEW

Bing Xue's study [1] provided a comprehensive overview of EC solutions to address decision making issues, encompassing all commonly used EC algorithms and focused on key features including representation, performance metrics, applications and search procedures.

Manuscript received on 29 November 2021. | Revised Manuscript received on 10 December 2021 | Manuscript Accepted on 15 December 2021| Manuscript published on 30 December 2021.

***** Correspondence Author

Aakriti Shukla*, Department of Computer Science and Engineering, Bansal Institute of Science & Technology, Bhopal (M.P.), India: aakritishukla2512@gmail.com

Dr Damodar Prasad Tiwari, Department of Computer Science and Engineering, Bansal Institute of Science & Technology, Bhopal (M.P.), India

© The Authors. Published by Lattice Science Publication (LSP). This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Important issues and challenges were also discussed.

A variety of EC algorithms have recently garnered a lot of attention for handling feature selection problems, according to this survey. In GAs, GP, and PSO, improving the representation to extract features and optimize classifiers, such as SVMs, is a frequent strategy. According to the findings of this research, significant progress has been made in improving the efficiency of multi-objective selecting features in aspects of reliability and set of parameters, paving the direction for new advancements. Qasem Al-Tashi [2] discussed main multi-objective feature selection challenges and strategies, which included all of the regularly used multi-objective algorithms and concentrated on the most important components. According to the (NFL) hypothesis, there has never been and will never be an optimization approach that can solve all problems. For multi-label classification data, Azam Asilian Bidgoli [3] suggested a binary many-objective feature selection approach that picks the ideal subset of features with the highest classification accuracy, lowest computing cost and fewest features. The binary suggested operators is based on oppositional learning and a qualified majority. compared to the NSGA-II feature selection method on multilabel data is additionally, the suggested method was tested on eight real-world multi-label datasets. Corina Cimpanu [4] presented the feature selection method for EEG classification as an optimization methodology developed using single and multi-objective GA in this study. Relevant features for n-back memory task assessment are selected from wave components received from a restricted range of electrodes. The analysis of a set of individuals generated at random or by SOO-ER demonstrates that characteristic subsets of various computational difficulty can produce similar incorrect negative rates. Because of a specific purpose, the usefulness of imposing direct control over the amount of features stands out in this scenario.

El-Hasnony, Ibrahim M. [5] This research provides a new paradigm for addressing and lowering feature selection barriers using a hybridization of grey wolf optimization (GWO) and particle swarm optimization (PSO). The proposed framework combines the K-nearest neighbor classifier with Euclidean separation matrices to get the optimum answers. In the analyses, twenty datasets are used, and statistical analyses are run to verify that perhaps the proposed framework is reliable and efficient throughout all comparative stages. The main purpose of the research was to identify the optimal set of attributes and fine-tune SVM parameters. Ehsan Eslami [6] suggested a new hybrid PSO–SVM approach to solve these primary objectives simultaneously. Two types of PSO versions, continuous-valued and binary versions, were combined to maximize the best SVM model parameters and the optimal function subset. The proposed technique outperforms other FS algorithms in terms of picking subsets of limited characteristics and improved classification accuracy on six well-known datasets. Qasem Al-Tashi [7] presented a binary version of the multi-objective grey wolf optimizer to solve the issue of feature selection. The continuous form of MOGWO was converted to its binary form using a sigmoid binary transfer function, allowing the technique to be used in the selecting features. The subset of functions was also analyzed using a wrapper Artificial Neural Network. The results also indicate that approaching component choice as a challenge with many goals is much more productive than considering the risks as a single solution, because BMOGWO-S can more efficiently explore the region for a variety of different answers. Dragi Kimovski [8] considers the approach to be a significant step ahead in the design of multi-objective evolutionary issues with a great amount of choice variables, such as the attribute choice challenge under consideration here. Feature selection has also been viewed as a multi-objective unsupervised clustering problem. By exploiting the independent evolution of population groups that interact after a given number of generations, this work leads to the parallel implementation of MOEA. S. Yadav et al. [9] suggested the method incorporates a variety of features and was created without the use of any domain-specific resources or software. The GENETAG, Aimed, GENIA, Bio Creative II (BC-II) gene mention recognition datasets are used as standards to assess the methods. The classifier performs better with a small feature set than with a bigger feature set, in contrast to the system constructed with a larger feature set. Ye Tian [10] proposed framework termed as PlatEMO and the attribute selection challenge, then explained how to implement the value selection challenge to PlatEMO, and last compared the output of eight MOEAs for selecting features. To explain how to use PlatEMO to solve novel MOPs, this paper employed a case study on the feature selection problem. MOEAs can be made more efficient for selecting features, and novel operators and selection techniques for MOEAs in handling the attribute selection challenge should indeed be created on PlatEMO. All of the studied MOEAs outperform a traditional feature selection strategy on most datasets. NSPSOFS and CMDPSOFS, two PSO-based multi-objective attribute selection techniques, were investigated by Mengjie Zhang [11]. The two feature selection algorithms were compared to two traditional methods (LFS and GSBS), a single objective algorithm (ErFS), a two-stage approach (2SFS), and three well-known multi-objective algorithms on 12 benchmark data sets of different complexity (NSGAII, SPEA2, and PAES). As multi-objective algorithms, NSPSOFS and CMDPSOFS are considered to be better productive than single-objective algorithms at exploring the optimum solution for a number of no dominated solutions. Users can choose their favorite solutions to address specific demands by looking at the Pareto front created by multi-objective algorithms. The contributions of various researchers are summarized in this section. Table I demonstrates that the majority of researchers concentrated their efforts on developing single- and multi-objective feature selection algorithms. Changing these characteristics, however, is not cost-effective. As a result, some solutions must be integrated with current designs in order to enhance effectiveness.

III. RESEARCH GAPS

Curse of Dimensionality: Among the most prominent problems in the data engineering industry is the curse of dimensionality, which means that a greater characteristic dimension increases computing difficulty and reduces the efficiency. The efficiency of big data classification and analysis will be harmed if the features are dimensioned incorrectly.

© Copyright: All rights reserved.

Due to the high dimensionality of the data, issues such as redundancy, missing samples, and no relationship between features arise.

If high-dimensional data with little resemblance is used, there will be more errors and machine learning techniques will have a harder time producing reliable findings. To mitigate for the dimensionality curse, low-dimensional feature sets with related or similar feature sets must be fed into data mining or machine learning algorithms. To address the obstacles associated with higher-dimensional data, the dimensionality of the information to be processed and visualized must be reduced [4].

Overfitting Problem: One of the primary challenges in feature selection and dimension reduction strategies is overfitting. Traditional dimension reduction algorithms can identify relationships among accessible feature sets for high-dimension data, but they cannot avoid the impact of the overfitting difficulty on the actual assessment of findings, which remains a difficult barrier.

Missing Value: In nearly every aspect of life, constant changes in size, volume, format and data patterns have resulted in a missing information issue. Missing data machine learning training is also a difficult task.

Table I. Comparative Result Analysis

Ref	Technique Used	Results	Limitations
[12]	Fuzzy C-Means	Accuracy = 96.3%	Increased computational time.
[13]	Genetic Algorithm	Maximal relevance is approx. 5.	Diverse solution on large feature set.
[14]	Distributed Fuzzy Rough Set	Accuracy = ~94%	Scalability issue.
[15]	Multilayer Co-Evolutionary	Accuracy = ~94%	With increase in noise level accuracy drops steeply.
[16]	PSO-GWO	Accuracy = ~90%	Overfitting problem.
[17]	Genetic Algorithm	Accuracy = ~94%	Increased computational time.
[18]	Particle swarm optimization	-	Large computational cost.
[19]	Canonical PSO	Performs better than PSO	Increased computational time.
[20]	MO-PSO	Error rate = 0.5	Computational speed is effected by increasing iteration
	MOGA	Error rate = 0.47	
	MOCSO	Error rate=0.01	
[21]	Shared Nearest Neighbor clustering	-	Replication of same blocks
[22]	Inclusive Similarity based Clustering	Time =50-100 sec Accuracy =~90%	Time consumption was high.

IV. CONCLUSION

The overfitting problem of classifiers for large datasets can be reduced by using dimension reduction or feature selection techniques. This research focuses on providing an analytical evaluation of current research issues in feature selection on big data. In comparison to traditional attribute choosing approaches, it is clear from a survey of the literature that bio-inspired (swarm intelligence, genetic algorithms, etc.) is the most common way for locating relevant characteristics.

REFERENCES

1. B. Xue, M. Zhang, W. N. Browne and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," in IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606-626, Aug. 2016, doi: 10.1109/TEVC.2015.2504420. [CrossRef]
2. Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili and H. Alhussian, "Approaches to Multi-Objective Feature Selection: A Systematic Literature Review," in IEEE Access, vol. 8, pp. 125076-125096, 2020, doi: 10.1109/ACCESS.2020.3007291. [CrossRef]
3. A. A. Bidgoli, H. Ebrahimpour-Komleh and S. Rahnamayan, "A Many-objective Feature Selection Algorithm for Multi-label Classification Based on Computational Complexity of Features," 2019 14th International Conference on Computer Science & Education (ICCSE), 2019, pp. 85-91, doi: 10.1109/ICCSE.2019.8845067. [CrossRef]
4. C. Cîmpanu, L. Ferariu, T. Dumitriu and F. Ungureanu, "Multi-Objective Optimization of Feature Selection procedure for EEG signals classification," 2017 E-Health and Bioengineering Conference (EHB), 2017, pp. 434-437, doi: 10.1109/EHB.2017.7995454. [CrossRef]
5. I. M. El-Hasnony, S. I. Barakat, M. Elhoseny and R. R. Mostafa, "Improved Feature Selection Model for Big Data Analytics," in IEEE Access, vol. 8, pp. 66989-67004, 2020, doi: 10.1109/ACCESS.2020.2986232. [CrossRef]
6. E. Eslami and M. Eftekhari, "An effective hybrid model based on PSO-SVM algorithm with a new local search for feature selection," 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE), 2014, pp. 404-409, doi: 10.1109/ICCKE.2014.6993448. [CrossRef]
7. Q. Al-Tashi et al., "Binary Multi-Objective Grey Wolf Optimizer for Feature Selection in Classification," in IEEE Access, vol. 8, pp. 106247-106263, 2020, doi: 10.1109/ACCESS.2020.3000040. [CrossRef]
8. Dragi Kimovski, Julio Ortega, Andrés Ortiz, Raúl Baños, "Parallel alternatives for evolutionary multi-objective optimization in unsupervised feature selection", Expert Systems with Applications, Volume 42, Issue 9, 2015, pp. 4239-4252. [CrossRef]
9. Yadav, S., Ekbal, A. & Saha, S. "Feature selection for entity extraction from multiple biomedical corpora: A PSO-based approach", Soft Comput 22, 6881–6904 (2018). https://doi.org/10.1007/s00500-017-2714-4[CrossRef]
10. Y. Tian, S. Yang, X. Zhang and Y. Jin, "Using PlatEMO to Solve Multi-Objective Optimization Problems in Applications: A Case Study on Feature Selection," 2019 IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 1710-1717, doi: 10.1109/CEC.2019.8789953. [CrossRef]
11. B. Xue, M. Zhang and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach," in IEEE Transactions on Cybernetics, vol. 43, no. 6, pp. 1656-1671, Dec. 2013, doi: 10.1109/TSMCB.2012.2227469. [CrossRef]
12. Muhammad Attique Khan, Habiba Arshad, Wasif Nisar, Muhammad Younus Javed, and Muhammad Sharif.: An Integrated Design of Fuzzy C-Means and NCA-Based Multi-properties Feature Reduction for Brain Tumor Recognition. Signal and Image Processing Techniques for the Development of Intelligent Healthcare Systems. 1-28 (2020). [CrossRef]
13. U. F. Siddiqi, S. M. Sait and O. Kaynak.: Genetic Algorithm for the Mutual Information-Based Feature Selection in Univariate Time Series Data. IEEE Access. 8, 9597-9609 (2020). [CrossRef]
14. L. Kong et al.: Distributed Feature Selection for Big Data Using Fuzzy Rough Sets. IEEE Transactions on Fuzzy Systems. 28, 846-857 (2020). [CrossRef]
15. W. Ding, C. Lin and W. Pedrycz: Multiple Relevant Feature Ensemble Selection Based on Multilayer Co-Evolutionary Consensus MapReduce. IEEE Transactions on Cybernetics. 50, 425-439 (2020). [CrossRef]
16. M. El-Hasnony, S. I. Barakat, M. Elhoseny and R. R. Mostafa: Improved Feature Selection Model for Big Data Analytics. IEEE Access. 8, 66989-67004 (2020). [CrossRef]
17. X. Liu, Y. Liang, S. Wang, Z. Yang and H. Ye.: A Hybrid Genetic Algorithm With Wrapper-Embedded Approaches for Feature Selection. IEEE Access. 6, 22863-22874 (2018). [CrossRef]

© Copyright: All rights reserved.

18. S. Fong, R. Wong, and A. Vasilakos.: Accelerated PSO swarm search feature selection for data stream mining big data. Services IEEE Transactions on Computing. 9, 33–45 (2016). [CrossRef]
19. Gu S, Cheng R, Jin Y.: Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Comput. 22, 811–822 (2018). [CrossRef]
20. D. Yan, H. Cao, Y. Yu, Y. Wang and X. Yu.: Single-Objective/Multiobjective Cat Swarm Optimization Clustering Analysis for Data Partition. IEEE Transactions on Automation Science and Engineering, 17(3). pp. 1633-1646 (2020).
21. S. Wang and C. F. Eick.: MR-SNN: Design of parallel Shared Nearest Neighbor clustering algorithm using MapReduce. IEEE International Conference on Big Data Analysis (ICBDA). pp. 312-315 (2017). [CrossRef]
22. J. Sangeetha and V. S. J. Prakash.: An Efficient Inclusive Similarity Based Clustering (ISC) Algorithm for Big Data. World Congress on Computing and Communication Technologies (WCCCT). pp. 84-88. (2017). [CrossRef]

AUTHORS PROFILE

Aakriti Shukla, I have completed my schooling from Vidya Bhumi Public School, CBSE board with Maths and Science as stream. I completed my Bachelor of Engineering in Computer Science and Engineering in 2018 from Truba Institute of science and technology, Bhopal (RGPV). I was working as an Android Application Developer in Wildnet Technologies Pvt Ltd wherein learned several aspects of Android and worked on several applications for Indian as well as International Clients. I am also well versed in Salesforce CRM platform and have worked with Marketing Cloud projects. Currently I am pursuing my MTech from Bansal Institute of science and technology, bhopal.

Damodar Tiwari, I Pursued phD from Rajiv Gandhi Proudyogiki Vishwavidylaya , Bhopal and have several research paper published like "A proper fit virtual machine migration approach for the load balancing in cloud" , "Impact of IPv4, IPv6 and dual stack interface over wireless networks" , "A virtual machine migration approach in cloud to optimise resource utilisation" , "Application of Viral System Algorithm in Load Balancing of Cloud Environment" , "LMP-DSR: Load balanced multi-path dynamic source routing protocol for mobile ad-hoc network". I am associated with Bansal group since 2006 and was initially designated as an assistant professor in Computer Science Department at Bansal Institute of science and technology. Currently, I am the director in Bansal Institute Of Science and Technology, Bhopal.

© Copyright: All rights reserved.