RESEARCH ARTICLE

SAMRIDDHI Volume 14, Issue 4, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Feature Reduction in Classification Tasks using Bio-inspired Optimization Algorithms

Rachna Kulhare*, S Veenadhari

Department of Computer Science and Engineering, RNTU, Bhopal, Madhya Pradesh, India

Abstract

In big data, there is a major difficulty that requires data mining to be conducted with elevated data in big technology, which would be gaining a lot of traction nowadays. When it comes to Big Data, feature selection approaches are seen to be a game changer since they can assist minimize the complexity of data, making it simpler to study and translate it into meaningful information. To enhance classification performance, feature selection removes unnecessary and redundant characteristics from the dataset. In this paper, Grey Wolf Approaches based on Quantum leaping neighbor memeplexes termed as QLGWONM is proposed. The result shows that when compared to the some bio-inspired algorithms such as PSO, GWO, ABA, CSA models, the suggested model performed well in terms of accuracy and have accuracy of 100% for brain tumor, CNS, Lung dataset and 97.1% for Ionosphere dataset and 99% for NSL-KDD.

Keywords: Big Data, Feature Extraction, Machine Learning, GWO, Classification.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i04.12

Introduction

Feature Selection (FS) has been extensively studied in data mining,[1-3] patterns recognition,[4,5] and machine learning.[6,7] Choosing which aspects of a dataset to keep and which to eliminate to create a more useful profile is referred to as "feature selection." The purpose of FS is to preserve the strong properties of the estimation method, so making it more exclusive and, as a result, more effective.[8]

Traditional methods of knowledge extraction will not function in this setting since they were not intended for use in this manner when they were developed. Big data is a mix of approaches that permit processing massive amounts of data. Big data is a concept that has been broken down into its component parts, which are as follows: velocity of data, diversity of data, validity of data, value of data and volume of data.[9] The first three are described as the method of data generation as well as the methods that are used to obtain and maintain the information. The components of genuineness and importance will be those that interact with one another, including quality and practicality. The most significant challenges associated with big data are assessing computer skills and subject knowledge, maintaining the confidentiality of data, and mining data. Because of the difficulties presented by these factors, the processes of data processing and mining are absolutely necessary for the progression of innovation.[10,11]

The FS method's objective in the context of large-scale mining of information and data is to eliminate redundant, irrelevant, and noisy characteristics while simultaneously grouping subsets of the data based on the key aspects of the original set. Because ineffective features have the potential to confound the learning system, researchers may be able to build a better model by deleting them from the data source. This would result in compute costs and lower memory.

Corresponding Author: Rachna Kulhare, Department of Computer Science and Engineering, RNTU, Bhopal, Madhya Pradesh, India, e-mail: rachna12kulhare@gmail.com

How to cite this article: Kulhare, R., Veenadhari, S. (2022). Feature Reduction in Classification Tasks using Bio-inspired Optimization Algorithms. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 14(4), 72-78.

Source of support: Nil

Conflict of interest: None

When processing high-dimensional data, one method for dimensional reduction is called feature selection. This method chooses a subset of the data that has characteristics that may be used in model creation. It has the benefit of being correctly preserved if it keeps a subset of its distinguishing characteristics as well as their practical interpretations in its original feature sets. This might lead to an enhancement in both accessibility[12] and comprehensibility. This feature selection will gain the required aspects while simultaneously discarding unnecessary and unwanted features, lowering costs without jeopardizing the product's functionality. Techniques for selecting features that are connected to such search tactics include the "filter",[14] "embedding approaches" [15] and "wrapper".[13] By using the model's acquired knowledge, the Wrapper method determines whether or not a character is relevant. Until it reaches a high level of performance, it will continue to choose subsets of characteristics and estimate training presentations based on the features that it selects. The fact that it examines the entirety of the search region[16,17] makes it a slow and hardly employed option. Instead of relying on learning algorithms, which are quite effective, the filter approaches make use of the qualities of the data to determine the significance of the features.[18]

The combination of embedded selection with the proposed model has produced an approach that integrates the rewards of either the filter or wrapping techniques. This has made it possible for the combination of embedded selection and model development to take advantage of the benefits of both techniques. First, there are no additional interactions within the learning process; second, because feature sets are not examined, wrapper techniques are more effective.[19] The categorization technique will aid data mining since it will categorize the data into organized groups or classes. These groups and classes can then be mined for information. Both the extraction of information and the creation of a strategy for the future will benefit from this. Thanks to this categorization, which will be broken down into two phases, users can make decisions that are in their best interests: In the beginning, there will be a learning strategy that will investigate a massive data collection.[20] The subsequent phase will consist of either an investigation or a verification of the correctness of the classification patterns. The categorization will be based on the application of models and object class labels to establish the category of an item whose category is not known. Neural networks, classification rules, decision trees, and mathematical formulas, in addition to KNN and Nave Bayes classifiers,[21] are only some of the classification methods used. A fresh approach has been proposed to enhance the analysis of massive amounts of data for this paper.

Related Work

C. Fahy et al.[22] suggest using dynamic feature masking for clustering high-dimensional data sets containing a large amount of information. After redundant features have been masked, clustering may begin across the beneficial qualities that have not been masked. As the relevance of a trait is reevaluated, the mask is adjusted to reflect this change. Previously unimportant characteristics are revealed, while those that have gained significance are concealed. The proposed method is independent of any particular algorithm and may be utilized with any of the two existing intensity-based clustering algorithms. These strategies frequently lack a technique to cope with drifting features and struggle when applied to huge datasets. The proposed dynamic feature mask improves the effectiveness of clustering while at the same time lowering the amount of time required for the underlying algorithm to perform its tasks. The F-score that was obtained is 0.62. The CMM value is 0.87, while the purity value is 0.93.

Joseph et al.[23] The study that is being recommended builds a feature selection algorithm method for the text-based classifying method by employing the strategies of ant-colony optimizing (ACO) and artificial neural networks (ANN). The utilization of Reuter's data set allowed for the demonstration of the efficiency of this hybrid approach. The results of carrying out the requested task indicate and provide proof of the competitiveness of the enterprise. An appropriate subset of features is found by the ANN algorithm within the data set that is provided. As a consequence of this, the problem of feature selection is effectively resolved by the ACO–ANN hybrid technique. It has also been used in a scenario involving a significant amount of data, and the outcomes of that application have been analyzed. The score for F-1 is 89.87 out of 100. There is an 81.35 percent degree of accuracy. Both precision and recall score 77.34 out of a possible 80.14 points.

Adaptive Boosting for Feature Selection is a revolutionary new dynamic FS strategy for data streams, and it was presented by Jean et al.[24] (ABFS). In addition to the method we have provided, it expands the use of feature selection-specific statistics from batch learning to streaming situations. The next step is to evaluate ABFS based on these criteria, considering both simulated and actual conditions. Because of this, ABFS can boost the classification rates of many different types of students, which in turn leads to an improvement in the efficient use of computer resources. An accuracy of 89.01 percent has been assigned to the ABFS-HAT prediction. We also investigate the recommended model's recall and selection accuracy as well as its level of complexity. The recommended method operates independently of the classifier, and the data show that ABFS can improve the performance of a wide variety of different classifiers in a variety of contexts. ABFS greatly increases the amount of computing time and memory needs utilization of Bayes and regression trees learners, despite the fact that it makes notable reductions in the quantity of the streaming data. Because the data show that processing speeds, as well as storage and memory consumption rates, have improved, the KNN classifier is an intriguing exception to consider.

X. Liu et al.[25] presented the hybrid feature selection algorithm employing an embedded wrapper approach and referred to it as HGAWE. This algorithm combines embedding normalization processes (local search) with evolutionary algorithms (global search). The findings demonstrate that it is superior to existing combination methods regarding the selection of features and classification accuracy. The proposed research has a sensitivity ranging from 0.724 to 0.935. The proposed action has a specificity ranging from 0.851 to 0.998, and it has an accuracy that ranges from 76.94 to 97.08%. According to the results of the studies, the HGAWE method is superior to a variety of other techniques that are currently used for feature selection regularization estimation.

Table 1: Comparative table of the recent researches

Ref	*Method*	*Result*
[22]	Clustering algorithms	F-score = 0.62. Purity0.93 and CMM = 0.87.
[23]	ACO and ANN	F-1 score = 89.87. Accuracy = 81.35%. Precision and recall = 77.34 and 80.14.
[24]	Adaptive Boosting for Feature Selection (ABFS)	Accuracy = 89.01%.
[25]	FS based on wrapping method and embedded feature	The sensitivity = 0.724-0.935. The specificity= 0.851-0.998 and The accuracy is 76.94%- 97.08%

[26]	Co-evolutionary algorithm	For NB and NB+CCEAFS
		precision = 0.70 and 90.20 %.
		Recall= 79.8 and 87.80%.
		F1 score = 83.70 and 88.80%,
		Accuracy= 79.78 and 87.79 % and
		No of features = 1024 and 201, respectively

[27]	link based particle swarm optimization (LBPSO)	For REUTER C selected features are 909 which is 48%. Purity measure= 0.8174 . accuracy = 96.1783%. Minimum Rand Index and Normalized Mutual Information (NMI) =0.66 and 0.56. Maximum NMI= 0.8451 and 0.8806 respectively
[28]	co-evolution CC	Dataset, accuracy, specificity and sensitivity of the proposed NB + CCEAFS are 87.79, 91.20 and 49.80%, respectively.
[29]	fireflies gravitational ant colony optimization (FGACO)	Sensitivity-98.43 %, specificity-98.21%, accuracy-98.9%, the number of selected features- 183, respectively, average efficiency = 98.4625%.

Table 2: Comparison of Accuracy Evaluation

Dataset	*PSO*	*GWO*	*ABA*	*CSA*	*QLGWONM*
Brain Tumor	0.8333	0.944	0.944	0.833	1
CNS	1	1	0.916	0.833	1
Lung	0.975	0.975	0.95	0.975	1
ionosphere	0.957	0.957	0.9285	0.928	0.971
NSL_KDD	0.97	0.98	0.98	0.97	0.99

PSO= Particle Swarm Optimization, GWO= Gray Wolf Optimization, ABA = Artificial Butterfly Algorithm, CSA= Crow Search Optimization

It can successfully identify significant characteristics of bio-mark genes in high-dimensional biological datasets, forecast the class of patients, and appropriately create the learning model. The HGAWE approach has shown to be a more helpful instrument, particularly in feature selection and learning prediction.

A. N. M et al.[26] In this work, the influence of the cooperative co-evolutionary technique for feature selection on six commonly used ML classification algorithms was investigated. The performance of the classifiers was demonstrated with and without selecting features. Because the dataset's properties were retained, SVM beat LR in most situations, but LR surpassed SVM in others. When the CCEAFS is used, however, NB consistently outperformed the other classifiers. Precision is 90.70 and 90.20 percent for NB and NB+CCEAFS, respectively. Recall rates are 79.8% and 87.80%, respectively. For the Qsar oral toxicity dataset, the F1 score is 83.70 and 88.80 percent, the accuracy is 79.78 and 87.79 percent, and the number of features is 1024 and 201, respectively.

Neetu et al.[27] The link based particle swarm optimization approach is proposed in this study as a novel feature selection method for unsupervised text clustering (LBPSO). In order to pick important features, this technique provides a novel

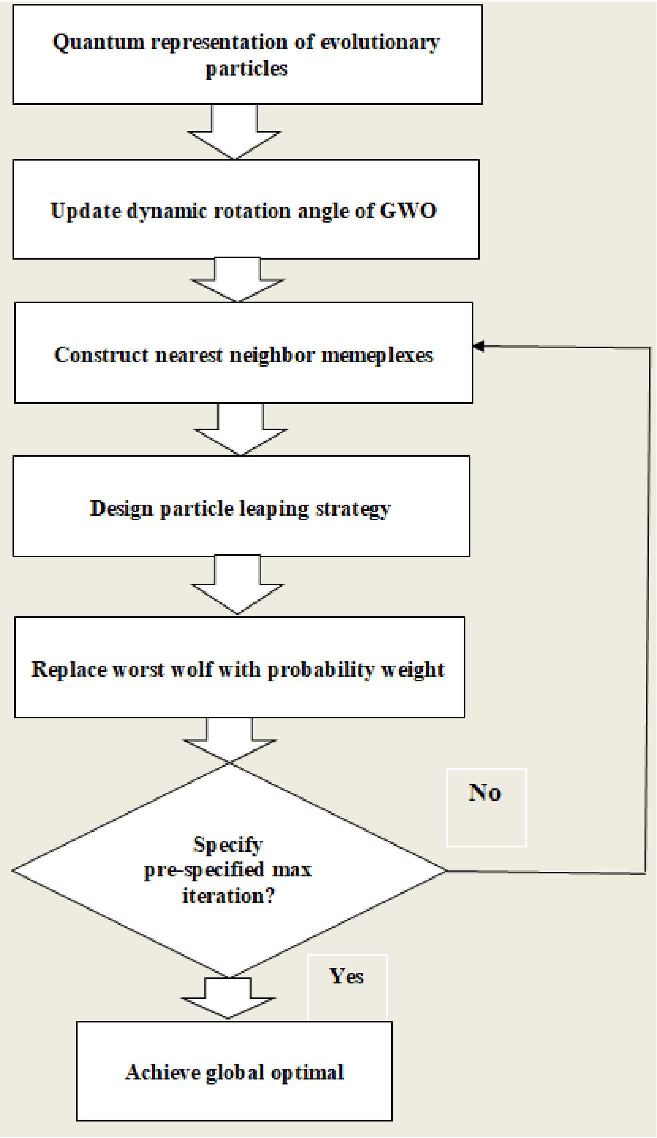
Quantum representation of evolutionary particles
Update dynamic rotation angle of GWO
Construct nearest neighbor memeplexes
Design particle leaping strategy
Replace worst wolf with probability weight
Specify pre-specified max iteration?
No
Yes
Achieve global optimal

Figure 1: QLGWONM steps.

neighbor selection mechanism in BPSO. The performance of LBPSO is better to other PSO-based algorithms, according to our assessment metrics. The chosen characteristics for REUTER C are 909, which is 48%. The purity value is 0.8174. TDT2A has a maximum accuracy of 96.1783 percent. The Minimum Rand Index (MRI) and Normalized Mutual Information (NMI) are 0.66 and 0.56, respectively. TDT2A has a maximum of 0.8451 and 0.8806 correspondingly. In this situation, the suggested feature selection process is a text clustering algorithm, which results in more related groupings. The suggested technique may be integrated with other meta heuristic algorithms in the future to provide more useful features while also boosting the search capabilities of the algorithm.

Rashid et al.[28] The use of cooperative co-evolution (CC) with a dynamical decompositions for FS was examined in this work. It presented a random feature grouping approach for feature selection using CC and tested six machine learning classifications on 7 datasets. The trials revealed that the FS procedure did not substantially impact the classifiers' performance. It also looked at the impact of feature selection on a variety of datasets, including those with a large number of samples but few features and those with a small number of samples but numerous features. The suggested work's efficacy is validated by a comparison of the classifying performance outcomes in terms of accuracies, sensitivities, and specificity. The suggested NB+CCEAFS has an accuracy, specificity, and sensitivity of 87.79, 91.20, and 49.80 percent, respectively, when tested on the QSAR Oral Toxicity Dataset.

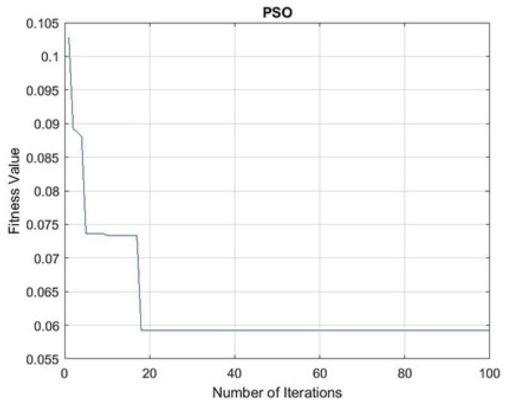

Figure 2: FV vs. Number of iteration for PSO Algorithm.

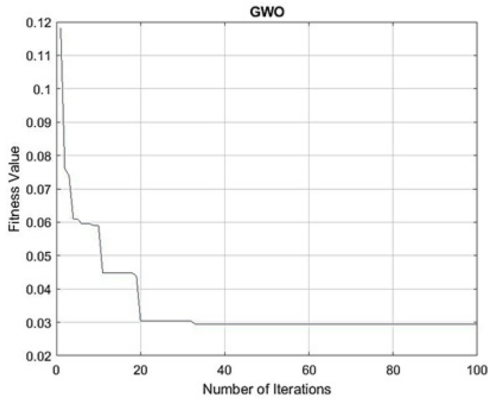

Figure 3: FV vs. No. of iteration for GWO algorithm.

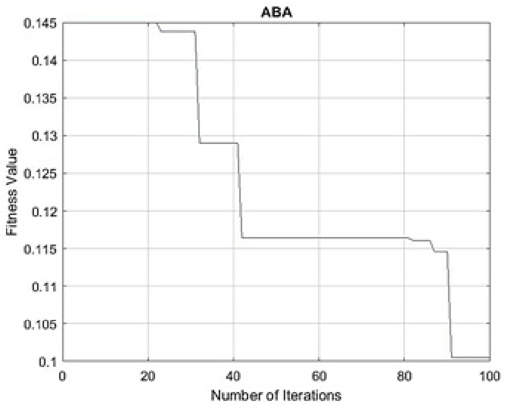

Figure 4: FV vs No. of iteration for ABA model.

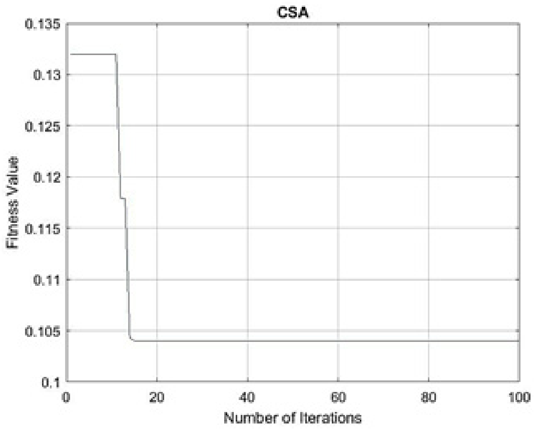

Figure 5: FV vs no. of iteration for CSA Algorithm.

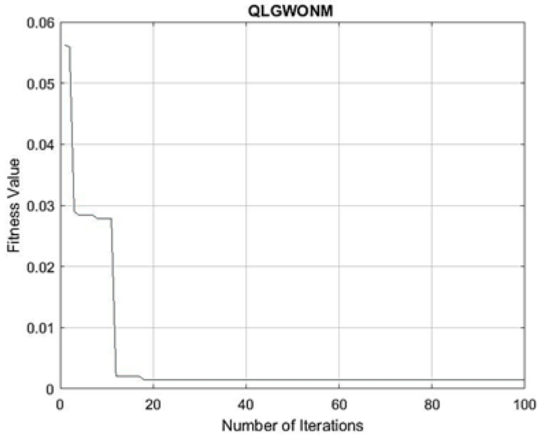

Figure 6: Fitness value vs Number of iteration for QLGWONM Algorithm).

Al Farraj et al.[29] defined that this work presents a strategy for optimizing feature selection and soft computing strategies for lowering the dataset's dimensionality. Initially, the data was gathered from a variety of sources, some of which included inconsistencies, limiting the system's effectiveness. Then, the inconsistencies and noisy data were eliminated using a normalized technique. The firefly gravitational ant colony optimization (FGACO) technique was then used to choose the optimum characteristics. During the selection process, this optimized FS properly analyses the qualities as well as relevance of the feature. All facts regarding certain predictive analytics are included in the specified feature. The experimental findings suggest that FGACO outperforms. The average effectiveness of the feature selection approach is 98.4625%. Sensitivity, specificity, accuracy, and the number of chosen characteristics for bank marketing datasets are 98.43, 98.21, 98.9, and 183, respectively Table 1.

Proposed Methodology

The FS is the method of creating a fully new database devoid of duplicated or unneeded features. Then it also implies that the initial data structure is preserved, and any critical stuff is not wasted or compromised. When there are a lot of samples and a lot of characteristics, feature selection techniques become quite important. These strategies are extensively utilized by consumers since they can effectively minimize dimensionality. It is a way to find characteristics that can quickly and accurately characterise an initial dataset. This research aims to provide a suggested paradigm for massive datasets processing challenges relying on quantum jumping GWO with nearest-neighbor memplexes of dimension reduction.

Because of overexploitation, many enhanced GWO methods are prone to being caught in a locally optimal as when the search space becomes more multidimensional, causing their efficiency to decrease. Exploring is seen to be an excellent way of learning much more about global optimum. On the other hand, extensive exploration degrades the quality of the chosen wolf's search. We present a quantum jumping GWO with nearest-neighbor memeplexes to allow better optimal matching among exploratory and exploitative of GWO for the fuzzy attribute reduction of complicated large data (QLGWONM). QLGWONM is shown in Figure 1 as a whole. It uses both a coordinates rotation gate and a dynamic rotation angle technique to explored the search space, find the global best area throughout fuzzy attribute reduction, and speed up premature convergence.

We propose the QLGWONMs to develop the deconstructed fuzzy attribute subsets in the quantum representations, rotational mechanisms, and wolf leaping approach, as indicated above. The fundamental procedures are outlined in Figure 1. This employment process in QLGWONM employs the nearest-neighbor memeplexes jumping search methodology to locate a global optimal wolf location with the appropriate combination of exploratory and exploitative strategies. This allows for the most accurate results possible. The fact that QLGWONM is able to obtain sufficient data about all quantum particles or wolves in its NNM is helpful since it allows for the enhancement of its dominating efficiency and the reduction of fuzzy attribute subsets.

Results and Discussion

Analysis of the change in iteration numbers with fitness values is shown to evaluate sentiment analysis and accuracy. The result shows how the overall number of iterations and fitness values have changed over time. In Figure 2, Fitness value (FV) vs number of iterations were shown for PSO algorithm. The fitness decreases from 0.103 to 0.6 for 0–20 number of iterations and almost constant after 20 iterations. The accuracy of the PSO is evaluated on 5 datasets brain tumor CNS, Lung, ionosphere and NSL_KDD. PSO have an accuracy of 0.8333 for brain tumor 0.975 for Lung, 0.975 for ionosphere, 0.97 for NSL_KDD.

In Figure 3, FV vs no. of iteration were shown for GWO model. The fitness decreases from 0.12 to 0.03 for 0-20 number of iteration and almost constant after 20 iterations. The accuracy of the PSO is evaluated on 5 datasets brain tumor, CNS, Lung, ionosphere and NSL_KDD. PSO have an accuracy of 0.944 for brain tumor, 1 FOR CNS, 0.975 for Lung, 0.957 for ionosphere, 0.98 for NSL_KDD.

In Figure 4, FV vs no. of iteration was shown for ABA model. The fitness decreases from 0.13 to 0.03 for 0–28 number of iteration and almost constant after 28 iterations. The accuracy of the PSO is evaluated on 5 datasets brain tumor, CNS, Lung, ionosphere and NSL_KDD. PSO have accuracy of 0.944 for brain tumor, 0.916 for CNS, 0.955 for Lung, 0.9285 for ionosphere, 0.98 for NSL_KDD.

In Figure 5, FV vs no. of iteration was shown for CSA model. The fitness decreases from 0.132 to 0.105 for 0–18 number of iteration and almost constant after 18 iterations. The accuracy of the PSO is evaluated on 5 datasets brain tumor, CNS, Lung, ionosphere and NSL_KDD. PSO have accuracy of 0.8333 for brain tumor, 0.833 for CNS, 0.975 for Lung, 0.928 for ionosphere, 0.97 for NSL_KDD.

In Figure 6, FV vs no. of iteration were shown for QLGOWNM model. The fitness decreases from 0-20 number of iteration and almost constant after 20 iterations. The accuracy of the PSO is evaluated on 5 datasets brain tumor, CNS, Lung, ionosphere and NSL_KDD. PSO have accuracy of 1 for brain tumor, 1 for CNS, 1 for Lung, 0.971 for ionosphere, 0.99 for NSL_KDD.

Table 2 shows the performance analysis of optimization algorithms over classification tasks. In this, we have taken datasets on 5 classification tasks such as brain tumor, CNS, Lung, ionosphere and NSL_KDD. On every datasets, QLGWONM results in better feature extraction as compared to other optimization algorithms.

Conclusion

Big data has increasingly gained traction in a variety of industries, including deep learning, pattern classification, healthcare, commercial, and infrastructure. Data analysis is essential for transforming the data into much more precise information that can be fed into decision-making processes. Information retrieval gets increasingly challenging as databases grow more varied and complicated. One way is to employ attribute selection, preprocessing, which minimizes the scale of the situation and makes computing and interpretation easier. Any data-mining technique will benefit from preprocessing since it creates a dependable and acceptable source. The selection of appropriate features may help us comprehend complicated data's properties and underlying structure, as well as increase the model's performance. Based on the proposed QLGWONM technique, this paper offers a unique hybrid feature selection model for 5datasets. Compared to the PSO, ABA, GWO, CSA model, the suggested model performed well in terms of accuracy and have an accuracy of 100% for brain tumor, CNS, Lung dataet and 97.1% for Ionosphere dataset and 99% for NSL-KDD. Compared to the weighted closest neighbor, the experimental results revealed excellent insights in both time utilization and feature weights.

References

[1] M. Zaffar, M. A. Hashmani, and K. S. Savita, "Performance analysis of feature selection algorithm for educational data mining," 2017 IEEE Conf. Big Data Anal. ICBDA 2017, vol. 2018-January, pp. 7–12, Feb. 2018, doi: 10.1109/ICBDAA.2017.8284099.

[2] A. K. Verma, S. Pal, and S. Kumar, "Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method-a Comparative Study," Appl. Biochem. Biotechnol., vol. 190, no. 2, pp. 341–359, Feb. 2020, doi: 10.1007/S12010-019-03093-Z.

[3] L. Abualigah and A. J. Dulaimi, "A novel feature selection method for data mining tasks using hybrid Sine Cosine Algorithm and Genetic Algorithm," Clust. Comput. 2021 243, vol. 24, no. 3, pp. 2161–2176, Feb. 2021, doi: 10.1007/S10586-021-03254-Y.

[4] H. M. Harb and A. S. Desuky, "Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization," Int. J. Comput. Appl., vol. 104, no. 5, pp. 975–8887, 2014.

[5] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, M. Meram, and A. Prathik, "Innovative feature selection and classification model for heart disease prediction," J. Reliab. Intell. Environ. 2021, pp. 1–11, Aug. 2021, doi: 10.1007/S40860-021-00152-3.

[6] P. Ghosh et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[7] H. M. Aydin, M. A. Ali, and E. G. Soyak, "The analysis of feature selection with machine learning for indoor positioning," SIU 2021 - 29th IEEE Conf. Signal Process. Commun. Appl. Proc., Jun. 2021, doi: 10.1109/SIU53274.2021.9478012.

[8] A. Kaur, K. Guleria, and N. Kumar Trivedi, "Feature Selection in Machine Learning: Methods and Comparison," 2021 Int. Conf. Adv. Comput. Innov. Technol. Eng. ICACITE 2021, pp. 789–795, Mar. 2021, doi: 10.1109/ICACITE51222.2021.9404623.

[9] F. BenSaid and A. M. Alimi, "Online feature selection system for big data classification based on multi-objective automated negotiation," Pattern Recognit., vol. 110, p. 107629, Feb. 2021, doi: 10.1016/J.PATCOG.2020.107629.

[10] M. Rong, D. Gong, and X. Gao, "Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends," IEEE Access, vol. 7, pp. 19709–19725, 2019, doi: 10.1109/ACCESS.2019.2894366.

[11] S. Meera and C. Sundar, "A hybrid metaheuristic approach for efficient feature selection methods in big data," J. Ambient Intell. Humaniz. Comput. 2020 123, vol. 12, no. 3, pp. 3743–3751, Jan. 2020, doi: 10.1007/S12652-019-01656-W.

[12] A. N. M. Bazlur Rashid and T. Choudhury, "Knowledge management overview of feature selection problem in high-dimensional financial data: Cooperative co-evolution and Map Reduce perspectives," Probl. Perspect. Manag., vol. 17, no. 4, pp. 340–359, 2019, doi: 10.21511/PPM.17(4).2019.28.

[13] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of Feature Selection Using Wrapper for Numeric Dataset with Random Forest Algorithm," Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021, pp. 179–183, Sep. 2021, doi: 10.1109/ISEMANTIC52711.2021.9573249.

[14] Y. Jiang, X. Liu, G. Yan, and J. Xiao, "Modified Binary Cuckoo Search for Feature Selection: A Hybrid Filter-Wrapper Approach," Proc. - 13th Int. Conf. Comput. Intell. Secur. CIS 2017, vol. 2018-January, pp. 488–491, Feb. 2018, doi: 10.1109/CIS.2017.00113.

[15] N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the wrapper feature selection evaluators on twitter sentiment classification," ICCIDS 2019 - 2nd Int. Conf. Comput. Intell. Data Sci. Proc., Feb. 2019, doi: 10.1109/ICCIDS.2019.8862033.

[16] S. Fong, R. Wong, and A. V. Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data," IEEE Trans. Serv. Comput., vol. 9, no. 1, pp. 33–45, Jan. 2016, doi: 10.1109/TSC.2015.2439695.

[17] F. Moslehi and A. Haeri, "A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection," J. Ambient Intell. Humaniz. Comput. 2019 113, vol. 11, no. 3, pp. 1105–1127, Jun. 2019, doi: 10.1007/S12652-019-01364-5.

[18] G. T. Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," IEEE Access, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[19] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, ''K-nearest neighbor classification,'' in Data Mining in Agriculture. New York, NY, USA: Springer, 2009, pp. 83–106.

[20] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," Proc. 2014 Sci. Inf. Conf. SAI 2014, pp. 372–378, Oct. 2014, doi: 10.1109/SAI.2014.6918213.

[21] B. Chakraborty and A. Kawamura, "A new penalty-based wrapper fitness function for feature subset selection with evolutionary algorithms," https://doi.org/10.1080/24751839.2018.1423792, vol. 2, no. 2, pp. 163–180, Apr. 2018, doi: 10.1080/24751839.2018.1423792.

[22] C. Fahy and S. Yang, "Dynamic Feature Selection for Clustering High Dimensional Data Streams," IEEE Access, vol. 7, pp. 127128–127140, 2019, doi: 10.1109/ACCESS.2019.2932308.

[23] R. Joseph Manoj, M. D. Anto Praveena, and K. Vijayakumar, "An ACO–ANN based feature selection algorithm for big data," Clust. Comput. 2018 222, vol. 22, no. 2, pp. 3953–3960, Mar. 2018, doi: 10.1007/S10586-018-2550-Z.

[24] J. P. Barddal, F. Enembreck, H. M. Gomes, A. Bifet, and B. Pfahringer, "Boosting decision stumps for dynamic feature selection on data streams," Inf. Syst., vol. 83, pp. 13–29, Jul. 2019, doi: 10.1016/J.IS.2019.02.003.

[25] N. Kushwaha and M. Pant, "Link based BPSO for feature selection in big data text clustering," Futur. Gener. Comput. Syst., vol. 82, pp. 190–199, May 2018, doi: 10.1016/J.FUTURE.2017.12.005.

[26] A. N. M. B. Rashid, M. Ahmed, L. F. Sikos, and P. Haskell-Dowland, "Cooperative co-evolution for feature selection in Big Data with random feature grouping," J. Big Data, vol. 7, no. 1, pp. 1–42, Dec. 2020, doi: 10.1186/S40537-020-00381-Y/FIGURES/2.

[27] O. AlFarraj, A. AlZubi, and A. Tolba, "Optimized feature selection algorithm based on fireflies with gravitational ant colony algorithm for big data predictive analytics," Neural Comput. Appl. 2018 315, vol. 31, no. 5, pp. 1391–1403, Jul. 2018, doi: 10.1007/S00521-018-3612-0