Technium Vol. 6, pp.56-65 (2023)
ISSN: 2668-778X
www.techniumscience.com

A Comparison Between the Performance of Features Selection Techniques: Survey Study

Nadia Mohammed Majeed[1]
Nadia.20csp72@student.uomosul.edu.iq

Fawziya Mahmood Ramo[1]
fawziyaramo@uomosul.edu.iq

[1]Computer Science Department, University of Mosul. Mosul / Iraq

Abstract Feature selection is one of the most popular and crucial methods of data processing used in different machine learning and data mining approaches to avoid high dimensionality and increase classification accuracy. Additionally, attribute selection aids in accelerating machine learning algorithms, improving prediction accuracy, data comprehension, decreasing data storage space, and minimizing the computational complexity of learning algorithms. For this reason, several feature selection approaches are used. To determine the essential feature or feature subsets needed to achieve classification objectives, several feature selection techniques have been suggested in the literature. In this research, different widely employed feature selection strategies have been evaluated by using different datasets to see how efficiently these techniques may be applied to achieve high performance of learning algorithms, which improves the classifier's prediction accuracy

Keywords: feature selection, dimensionality reduction, classification

1. Introduction

Due to the increase in the amount of data we deal with in the current era, the extraction of relevant basic data has become increasingly necessary. When using traditional model-based machine learning methods, higher-dimensional data increases search area, computational time, and takes longer to process. And noise can occur, affecting the model building, and resulting in a loss of efficiency [1]. The increase in the dimensions of data with multiple variables and access to a good representation of it is one of the problems faced by machine learning applications and the data mining process, as well as areas related to artificial intelligence. Therefore, attribute selection techniques are used to deal with large datasets. The main purpose of dimensionality reduction is to improve prediction performance by removing irrelevant and redundant features and reducing complexity to reach a simpler representation of data, which leads to easier pattern recognition and many other intelligent operations[2].

2. Feature Selection Process

The four basic steps of the feature selection process are the creation of the feature subset, the evaluation of the subset, the stopping criteria, and the validation of the results. A candidate subset is selected for the assessment using a heuristic search method as the technique for creating feature subsets used to improve the efficiency of the feature selection method, like sequential and random searches. These search techniques focus on the incremental addition or elimination of features. An evaluation criterion is used to evaluate the goodness of the created subset. If the newly created subset is better than the

previous subset, it is also used to replace the previous subset. These two processes are repeated until the stopping condition is achieved. The final optimal feature subset is then verified using other tests or previous knowledge[3]. Figure (3-1) demonstrates how features are selected.

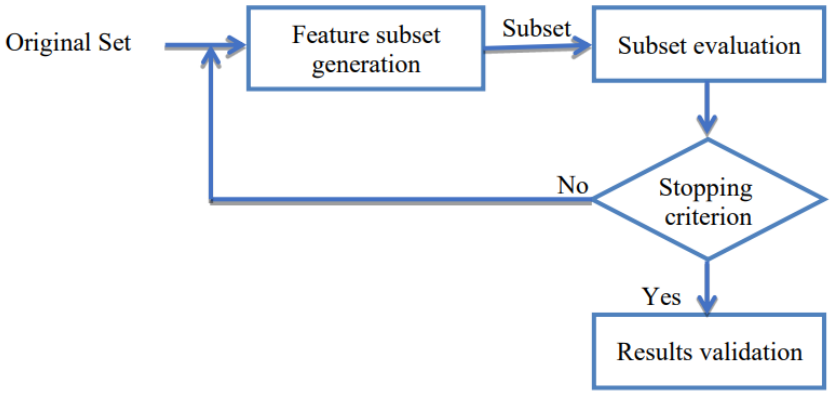

Figure 1: feature selection process

The search starting point can be an empty set, a full set, or a randomly generated subset used to generate subsets. From the starting point, it can scan subsets of features in different directions, including forward, backward, and random. Forward search refers to a search that begins with an empty set of features, this strategy iteratively adds a single feature to the existing subset and features that are gradually combined into larger and larger subsets until the assessment of the rate of error does not improve further or until other criteria are met. During the iteration, the error is evaluated for each of the remaining features that are added (one by one) to the existing feature subset. Then the feature that makes the most improvement is permanently added to the relevant subset file of the model. Whereas in backward search, starting from all features, iteratively the least significant feature is removed one by one from the current subset based on the evaluation criterion or until the stopping criterion is achieved. To avoid the local maxima problem, random search adds or removes features randomly [4][3] [5].

3. Related Works

Chuang, Li-Yeh, et al. [6] employed an improved binary particle swarm optimization (IBPSO) to perform a feature selection, and K-nearest neighbor (K-NN) acts as an IBPSO evaluator for problems involving the classification of gene expression data. Experimental findings demonstrate that our approach effectively decreased the number of genes (features) required and facilitated the selection of genes. In comparison to the best outcomes previously covered the average accuracy of the classification for the suggested approach improved by 2.85%. The suggested approach can be used as an efficient pre-processing method to assist in feature selection process optimization since it improves classification accuracy.

Darzi, Mohammad, et al. [7] proposed a genetic algorithm (GA) and case-based reasoning (CBR) to address feature selection in diagnosing breast cancer. (GA) employed to determine every possible subset of attributes, and (CBR) was utilized to estimate the assessment outcome of each subset. In comparison to other models with the WBCD dataset, the findings demonstrate how well the suggested model performed after feature selection, where they have an accuracy of 97.37 %.

Uzer, Mustafa Serter , et al. [8] suggest a hybrid technique that utilize the artificial bee colony (ABC) for selecting features to select the important attributes and decrease the dimension of the feature vector. Support vector machines (SVM) is then utilized for classification to calculate the accuracy rate. This paper's main objective is to explore how removing duplicate and irrelevant characteristics from datasets help improve classification accuracy. To increase the reliability of the classifier, k-fold cross-validation was employed. The suggested approach achieved classification accuracy for the diagnosis of hepatitis, liver diseases, and diabetes datasets from the UCI database of 94.92%, 74.81%, and 79.29%, respectively. Findings obtained indicate that the performance of the suggested technique is quite successful when compared to other results obtained and appears to be very promising for pattern recognition applications.

Hosseinzadeh, Faezeh, *et al*.[9] proposed an approach based on sequence-derived structural and physicochemical attributes of proteins that participate in both types of cancers. To predict and detect the kind of lung cancer, three different machine-learning models (SVM, ANN, and NB) were tested . When algorithms were applied to datasets generated using attribute weighting models rather than the original dataset, their performance in predicting the tumor type of lung cancer increased. Moreover, SVM and SVM Linear methods produced the best cancer-type diagnosis (82%). Whereas the Neural Net model application on the SVM dataset resulted in the highest ANN accuracy (88%). The outcomes also demonstrated that feature weighting could be useful to both processing speed and producing more accurate findings to forecast the kind of lung cancer tumors.

Mafarja, Majdi, *et al*.[10] Used the binary dragonfly method as a wrapper feature selection technique and as an evaluator a K-NN classifier was used. The performance of the proposed technique was evaluated using eighteen UCI datasets. The experimental results showed superior performance of the Binary Dragonfly (BD) approach compared to the results of Particle Crowd Optimization (PSO), and (GAs) in terms of the number of attributes selected and the ability to search the feature space for the most relevant features for classification tasks, and also obtained better results in most of the UCI datasets. For example when using the cancer dataset (BreastEW), the average classification accuracy obtained using the selected features was (0.961), while the classification accuracy (0.945) was obtained when using all the features.

Belina,S. *et al*.[11] suggested a novel Hybrid Wrapper and Filter based FS (HWFFS) technique for selecting the best subset of features from datasets to predict Chronic Kidney Disease (CKD) using a (SVM) classifier. There are 24 characteristics for predicting CKD or non-CKD in the UCI statistics collection. From the original dataset, at least 16 characteristics were chosen. The effectiveness of the reduced feature set has been evaluated using the SVM classifier. The findings show that, compared to existing approaches like NB and ANN, the proposed SVM with the HWFFS algorithm delivers results with a lower error rate of 10.00%. This leads to the conclusion that the proposed work outperforms other classifiers.

Khellat-kihel, Badra, and Mohamed Benyettou [12] proposed methodology for feature selection that involves Ant Colony (ACO), Artificial Bee (ABC), and Firefly Algorithm (FA) algorithms to identify the most relevant features in the UCI dataset. Genetic algorithm can create a new set of chromosomes as the initial set generated by the three algorithms used (ACO, ABC, FA) after feature selection. Neural network algorithm was used to improve classifier performance. The results of the study showed that the rate of error in classification accuracy is lower after selecting features compared to the results when using the original features in the data sets used. for example when using the Spect Heart data set, the rate of error before selecting feature is (0.1491) for three algorithms (ABC, ACO, FA), while the rate of error after selecting the attributes is lower as it is (0.0625, 0.0804, 0.0568), respectively.

Yasen, Mais, *et al*.[13] suggested the implementation of Dragonfly Algorithm (DA) as an optimization strategy for the weights of each link between the Artificial Neural Network (ANN) neurons in a method known as (ANN-DA) which is utilized to predict disease and uses real data as a case study. In order to assess ANN-DA, we will also propose a model called ANN-ABC that uses ABC. Different measures have been used to compare both optimization techniques experimentally to other well-known classifiers. The findings indicate that ANN-DA outperformed ANN-ABC and basic ANN in terms of effectiveness When used in selecting features. ANN-DA outcomes were also more accurate throughout all datasets, for example when utilizing Hepatitis datasets the accuracy of ANN-DA is (94.340).

Anuradha, and Vasantha Kalyani David. [14] Whale Swarm Algorithm (WSA) proposed for carrying out the feature selection process. The subsets are acquired for various numbers of iterations using three various classifiers for the fitness function: Logistic Regression, Random Forest, and K-N N. For classification, the dataset depending on the chosen subsets is then utilized. Four various classifiers, namely Gaussian Naive Bayes, Support Vector Classification, Random Forest, and Logistic Regression, have their classification accuracy compared. WSA using Logistic Regression as the fitness function (WSA-LR), among these provides an average subset of eight features, and Random Forest Classifier's accuracy is determined to be 85.7%, which is higher than that of the other classifiers.

Latha, C. Beulah Christalin, and S. Carolin Jeeva [15] proposed the ensemble algorithms: bagging, boosting, stacking, and majority voting were used in the experiments to predict heart disease. An increase in accuracy of up to 6.92% occurs when bagging is used, when using boosting, the maximum accuracy improvement is 5.94%, Stacking increased accuracy by up to 6.93%, and when individual classifiers are grouped with majority voting, accuracy increased by up to 7.26 %. A comparison of the findings revealed that vote by majority leads to the most tremendous accuracy increase for the Cleveland heart dataset. The Performance is further improved by using feature selection techniques. Feature selection demonstrates enhancement in majority voting with all of the attribute sets, the best accuracy was 3.29%, by the feature set FS2. Feature selection methods have also enhanced the accuracy of ensemble algorithms.

vinmalar,F. Leena, and A. Kumar Kombaiya [16] Used the Improved Dragonfly optimization Algorithm (IDA) to decrease the dataset's dimensionality for lung cancer gene expression. To select an important subset of features, IDA uses the wrapper feature selection technique. To recognition of lung tumor subtypes, random subspace (RS), artificial neural network (ANN), and minimal sequential optimization (SMO) classifiers were used, the features selected applied in these algorithms to predict lung cancer effectively. Finally, the results proved that the proposed IDA-SMO has better accuracy for diagnosing lung cancer than other methods, where the accuracy of IDA-SMO was (0.93), which is greater by (3.79%) than the IDA-RS and IDA-ANN methods for diagnosing lung cancer.

Too, Jingwei, and Seyedali Mirjalili [17] suggested Hyper Learning Binary Dragonfly Algorithm (HLBDA) as a wrapper-based approach to determining the ideal subset of attributes for a specific classification issue. Twenty-one datasets were collected from the UCI repository and used to validate the performance of the proposed method. Besides, the proposed model has been applied to COVID-19 datasets. Eight methods from the literature are compared to the suggested HLBDA. The obtained results show that the proposed HLBDA is superior to the other algorithms in most of the datasets. In addition, HLBDA was able to increase the classification accuracy by decreasing the number of selected attributes.

Sydney Mambwe Kasongo [18] proposed machine learning methods to create effective intrusion detection system utilizing the NSL-KDD dataset. To select the best features, feature selection technique based on wrapper method was used with the Genetic Algorithm (GA). The extra-trees (ET) algorithm was applied as a fitness function in GA. Seven feature subsets were chosen by the GA for the multiclass

classification process. The following ML techniques were used in the modeling process: decision tree (DT), (SVM), random forest (RF), extra-trees (ET), extreme gradient boosting (XGB), and naïve Bayes (NB). For the multiclass and binary classification techniques, trials were performed. The outcomes showed that applying the GA algorithm enhances the efficiency of the chosen classifiers, where GA-DT was found to have an AUC of 89% and a detection accuracy of 89.26% for the two-way classification. With its multiclass classification approach, the GA-XGB was able to detect intrusions with an accuracy of 87.26%.

Majeed, Nadia Mohammed, and Fawziya Mahmood Ramo [19] Propose a wrapper feature selection method depending on the Dragonfly algorithm. The best subset of attributes to accurately diagnose breast cancer as benign or malignant is found using the dragonfly optimization approach. Hard vote technique is used as a fitness function to evaluate each feature subset. ensemble hard voting combines the SVM, K-NN, Naïve Bayes (NB), Decision Tree (DT), and Random Forest machine learning methods. The voting ensemble classifier has the highest accuracy value as compared to individual classifiers, according to the experiment's findings. The suggested technique demonstrated that when training the subset of features, the voting classifier's accuracy is high at 98.24%, while training all features produced an accuracy of 96.49%, utilizing the (WDBC) Dataset from the UCI repository.

Table 1: A comparative analysis of feature section approaches for high-dimensional classification using various algorithms

Paper title	Authors name	Methods used	year	Results
Improved binary PSO for feature selection using gene expression data	Chuang, Li-Yeh et al.	Improved binary particle swarm optimization (IBPSO) to perform a feature selection, and a K-nearest neighbor (K-NN) acts as an IBPSO evaluator	2008	Average of accuracy is 2.85%.
Feature selection for breast cancer diagnosis: A case-based wrapper approach	Darzi, Mohammad , et al.	genetic algorithm (GA) and case-based reasoning (CBR) to address feature selection	2011	Accuracy is 97.37 %.

Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification	Uzer, Mustafa Serter , et al.	(ABC) ,Support vector machines (SVM)	2013	Accuracy for the diagnosis of hepatitis, liver diseases and diabetes datasets 94.92%, 74.81%, and 79.29%, respectively.
Prediction of lung tumor types based on protein attributes by machine learning algorithms	Hosseinzadeh, Faezeh, *et al*.	SVM, ANN, and NB	2013	SVM and SVM Linear methods accuracy is (82%). Whereas the Neural Net model ANN accuracy (88%).
Binary Dragonfly Algorithm for Feature Selection	Mafarja, Majdi, *et al*.	binary dragonfly method as a wrapper feature selection technique and K-Nearest Neighbor classifier	2017	Accuracy obtained using the selected features was (0.961), while the classification accuracy (0.945) was obtained when using all the features.
Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction	S. Belina V.J. Sara, K. Kalaiselvi	Hybrid Wrapper and Filter based FS techniques, (SVM) classifier	2018	HWFFS with SVM results a lower error rate of 10.00%.
Hybrid Bio-Inspired Approach for Feature Subset Selection	Khellat-kihel, Badra, and Mohamed Benyettou	(ACO), (ABC), and (FA) algorithms for feature selection, Genetic algorithm, Neural network algorithm	2018	Error rate before selecting feature is (0.1491) for (ABC, ACO, FA), while the rate of error after selecting the attributes is lower as it is (0.0625, 0.0804, 0.0568), respectively.

Optimizing Neural Networks using Dragonfly Algorithm for Medical Prediction	Yasen, Mais, *et al.*	Dragonfly Algorithm (DA), Artificial Neural Network (ANN), ABC	2018	Accuracy of ANN-DA is (94.340).
Feature Selection Using Whale Swarm Algorithm and a Comparison of Classifiers for Prediction of CARDIOVASCULAR DISEASES	Anuradha, and Vasantha Kalyani David	(WSA) for feature selection. Logistic Regression, Random Forest, and K-Nearest Neighbours for classification	2019	Random Forest Classifier's accuracy is determined to be 85.7%, which is higher than that of the other classifiers.
Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques	Latha, C. Beulah Christalin, and S. Carolin Jeeva	bagging, boosting, stacking, and majority voting	2019	An increase in accuracy of up to (6.92%, 5.94%, 6.93%, 7.26 %) occurs when used bagging, boosting, Stacking, and majority voting, respectively. whereas the best accuracy was 3.29%, for voting With the use of feature selection
An Improved Dragonfly Optimization Algorithm based Feature Selection in High Dimensional Gene Expression Analysis for Lung Cancer Recognition	vinmalar,F. Leena, and A. Kumar Kombaiya	(IDA) for feature selection, (RS), (ANN), and (SMO) for classification	2020	the accuracy of IDA-SMO was (0.93), which is greater by (3.79%) than the IDA-RS and IDA-ANN methods

A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study	Too, Jingwei, and Seyedali Mirjalili	(HLBDA) as a wrapper-based approach to determining the ideal subset of attributes	2021	The obtained results show that the proposed HLBDA is superior to the other algorithms in most of the datasets, and able to increase the classification accuracy by decreasing the number of selected attributes
Genetic Algorithm Based Feature Selection Technique for Optimal Intrusion Detection	Sydney Mambwe Kasongo	(GA) for feature selection, (ET) algorithm was applied as a fitness function in GA. (DT), (SVM), (RF), (ET), (XGB), and (NB) for classification.	2021	GA-DT was found to have an AUC of 89% and a detection accuracy of 89.26% for the two-way classification. With its multiclass classification approach, the GA-XGB was able to detect intrusions with an accuracy of 87.26%.
Implementation of Features Selection Based on Dragonfly Optimization Algorithm	Majeed, Nadia Mohammed and Fawziya Mahmood Ramo	wrapper feature selection method depending on the Dragonfly algorithm, hard vote technique, SVM,(K-NN), (NB), (DT), and Random Forest	2022	when training the subset of features, the voting classifier's accuracy is high at 98.24%, while training all features produced an accuracy of 96.49%

4. Conclusion

This survey shows a study on different feature selection techniques applied to high dimensional data in the Classification process. The study indicates that feature selection is a critical step in machine learning applications and it provides an effective way to analyze high dimensional data by reducing not relevance and redundant data. However, selecting the correct feature selection algorithm constitutes an enormous challenge for machine learning problems. Many feature selection approaches have been suggested to address this problem in recent years.

References

[1] G. Kicska and A. Kiss, “Comparing swarm intelligence algorithms for dimension reduction in machine learning,” *Big Data Cogn. Comput.*, vol. 5, no. 3, 2021, doi: 10.3390/bdcc5030036.

[2] C. F. Selection and P. Correlation, “NEW ORGANIZATION PROCESS OF FEATURE SELECTION BY FILTER WITH CORRELATION-BASED FEATURES SELECTION METHOD,” vol. 3, no. 21, pp. 39–50, 2022.

[3] M. K. H. AL-Malali, “Behavioral Sense Classification using Machine Learning Algorithms,” pp. 1–144, 2021.

[4] L. Brezočnik, I. Fister, and V. Podgorelec, “Swarm intelligence algorithms for feature selection: A review,” *Appl. Sci.*, vol. 8, no. 9, 2018, doi: 10.3390/app8091521.

[5] Q. Al-Tashi *et al.*, “Binary Multi-Objective Grey Wolf Optimizer for Feature Selection in Classification,” *IEEE Access*, vol. 8, pp. 106247–106263, 2020, doi: 10.1109/ACCESS.2020.3000040.

[6] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, “Improved binary PSO for feature selection using gene expression data,” *Comput. Biol. Chem.*, vol. 32, no. 1, pp. 29–38, 2008, doi: 10.1016/j.compbiolchem.2007.09.005.

[7] M. Darzi, A. AsgharLiaei, M. Hosseini, and H. Asghari, “Feature selection for breast cancer diagnosis: A case-based wrapper approach,” *World Acad. Sci. Eng. Technol.*, vol. 53, no. 5, pp. 1142–1145, 2011.

[8] M. S. Uzer, N. Yilmaz, and O. Inan, “Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification,” vol. 2013, 2013.

[9] F. Hosseinzadeh, A. H. Kayvanjoo, and M. Ebrahimi, “Prediction of lung tumor types based on protein attributes by machine learning algorithms,” *Springerplus*, vol. 2, no. 1, pp. 1–14, 2013, doi: 10.1186/2193-1801-2-238.

[10] M. M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, and S. Mirjalili, “Binary Dragonfly Algorithm for Feature Selection,” *Proc. - 2017 Int. Conf. New Trends Comput. Sci. ICTCS 2017*, vol. 2018-Janua, pp. 12–17, 2017, doi: 10.1109/ICTCS.2017.43.

[11] S. B. V. J. Sara and K. Kalaiselvi, “Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction,” *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 190–195, 2018, doi: 10.14419/ijet.v7i2.31.13438.

[12] R. Kaushik and B. Keswani, “Hybrid Bio-Inspired Approach for Feature Subset Selection,” vol. 14, no. 03, pp. 10–14, 2018.

[13] M. Yasen, N. Al-Madi, and N. Obeid, “Optimizing Neural Networks using Dragonfly Algorithm for Medical Prediction,” *2018 8th Int. Conf. Comput. Sci. Inf. Technol. CSIT 2018*, pp. 71–76, 2018, doi: 10.1109/CSIT.2018.8486178.

[14] K. David, “Feature Selection Using Whale Swarm Algorithm and a Comparison of Classifiers for Prediction of CARDIOVASCULAR DISEASES,” *Int. J. Res. Anal. Rev.*, vol. 6, no. 2, pp. 123–130, 2019.

[15] C. B. C. Latha and S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” *Informatics Med. Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100203.

[16] F. L. vinmalar* and D. A. K. Kombaiya, “An Improved Dragonfly Optimization Algorithm based Feature Selection in High Dimensional Gene Expression Analysis for Lung Cancer Recognition,” *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 8, pp. 896–908, 2020, doi: 10.35940/ijitee.h6302.069820.

[17] J. Too and S. Mirjalili, “A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study,” *Knowledge-Based Syst.*, vol. 212, p. 106553, 2021, doi: 10.1016/j.knosys.2020.106553.

[18] S. M. Kasongo, “Genetic Algorithm Based Feature Selection Technique for Optimal Intrusion Detection,” no. June, pp. 1–22, 2021, doi: 10.20944/preprints202106.0710.v1.

[19] N. M. Majeed, “Implementation of Features Selection Based on Dragonfly Optimization Algorithm,” vol. 4, no. 10, pp. 44–52, 2022.