World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Feature selection for image classification using ESSO

Kumar Siddamallappa U *

Department of studies in Computer Science, Davangere University, Davangere, India.

World Journal of Advanced Research and Reviews, 2023, 19(03), 1170–1176

Publication history: Received on 16 August 2023; revised on 24 September 2023; accepted on 27 September 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.19.3.2005

Abstract

Feature selection is one of the major components of the data processing flow that correctly selects real-time entities for categorization. It is utilized in numerous research fields, including the machine learning approach. Feature selection and classification are beneficial to the processing of biomedical data in high-importance, high-dimensional datasets. Due to numerous obstacles in research, the current implementations are inadequate for predicting classification accuracy. To handle the classification challenge, we require dedicated neural network and classification model approaches. Initial features are selected using a method called "fuzzy c-means clustering with rough set theory" and are subsequently classified using "support vector machines." The learning machines' accuracy is seriously harmed by irrelevant and redundant features. The curse of dimensionality makes it difficult to locate data clusters in high-dimensional space. Data in the irrelevant dimensions may cause a lot of noise when the dimensionality grows. In addition, the current approach has a significant flaw: it is extremely time-consuming. The proposed solution used efficient feature subset selection in high-dimensional data to fix these problems. To begin with, our recommended method used Enhanced Social Spider Optimization (ESSO) computation, in which the standard Social Spider Optimization is improved with the aid of optimal radial based calculation to select the best highlights. When categorizing data, the Optimal Radial Basis Function Neural Network (ORBFNN) is used. Methods for calculating Artificial Bee Colony (ABC) are used to streamline RBFNN's sufficiency in characterizing smaller scale show information.

Keywords: ESSO; RBFNN; ORBFNN; ABC

1. Introduction

In medical domains, data mining is particularly beneficial for knowledge discovery and classification of diagnostic outcomes. We pre-process data or information using natural language processing (NLP) to create datasets or information that is ready to classify for optimization results. Due to the time series' real-world entities, biomedical data processing is challenging due to the raw data's high dimensionality. Due to the large dimensionality of the data, the feature selection and classification algorithms are insufficient to give classification results. Thus, an advanced study is required to select the appropriate features to demonstrate great performance.

The feature selection retains information or establishes rule priority for data type definition selection. Classification brings the type definition back to life in order to give the optimal result. Feature selection is a critical component of the data process because it correctly selects real-time entities in the data, which results in categorization. It is employed in a variety of study fields, one of which is machine learning. The communities of data mining statistics, image processing, and pattern analysis are all actively engaged in research. The critical component of feature selection is to minimize the number of categorization steps required. By utilizing low-level features, it is possible to classify data in order to enhance performance with the least amount of time complexity.

* Corresponding author: Dr. Kumar Siddamallappa U (Orcid Id: 0000-0002-1975-3868)

The primary goal of redundant feature selection is to select a subset of input features by dispersing out high points that have no discernible material for the selective feature scenario. Emphasizing the feature selection process can significantly improve the understanding ability of future classifier models and frequently result in a model that sums up better to inconspicuous focuses. Additionally, it is frequently the case that identifying the optimal subset of prophetic traits is a major challenge in and of itself. For instance, the specification feature may determine whether or not a risky medical operation is necessary for therapy based on the selected elements. The feature selection problem is the most common type of classification problem, as it involves a large number of dimensions. Due to inconsistency in feature selection, the problem is that a subset of unimportant features is chosen to boost the algorithm's execution speed, but the classification accuracy is low.

The research primarily focuses on the classification problem and how to handle it by utilizing effective feature selection for CNN for easy feature prediction and classification in high-dimensional datasets. They are as follows:

- To enhance feature selection performance, a rough set-based Multiple Kernel Fuzzy C-means selecting the model, and an enhanced social spider, fruit fly algorithm is utilized.
- To enhance feature selection (ESSO) in order to create a more efficient strategy to reprocess data points based on feature selection, which will include category as the label for classes.
- To use a raw collection of data to train a neural classifier (SVM) using an appropriate feature selection model.
- To enhance classification accuracy by the use of multi-attribute case prediction for categorization via RBFNN.
- To develop an optimum neural fuzzy classifier for classification and efficient feature selection using attribute subset feature selection (ABC).

2. Literature review

The feature selection technique has evolved into one of the most essential strategies in the field of content-programmed order (sundari et al. 2013). Another technique for picking content features in light of Information Gain and Genetic Algorithm is offered. This technique derives the characteristic based on data acquired from the occurrence of events. Meanwhile, for evidence sifting frameworks, this strategy has boosted wellbeing capacity by needing a comprehensive assessment of weight, substance, and vector closeness measurement, among other aspects.

Kaveh and Rostami (2021) proposed using POLSAR images to classify land cover. This research will employ HBBOSVM to classify RADARSAT 2 POLSAR images from San Francisco, California, using a biogeography-based support vector machine (HBBOSVM). We reduced the number of features we use while improving categorization. Preprocessing, feature selection, and categorization are all included. Speckle removal and feature extraction were pre-processed. If you wish to find the optimal features, utilize a biogeography-based optimization method and an onlooker bee (ABC). Using SVM, pixels were identified by land cover. In order to get ground truth samples for Google Earth, Pauli RGB, high-resolution photos, and the national land cover database (NLCD 2006). Its performance was compared against BBOSVM, ABCSVM, PSOSVM, and others. The HBBO is tested on 20 benchmarks. Overall and average HBBOSVM accuracy is 96%, which is better than other results. The HBBOSVM outperforms others in kappa, average accuracy, and convergence trend. Effective meta-heuristic for benchmark problems is HBBO. Both optimization and machine learning have benefits.

Xie et al. (2021) presented two PSOs for feature selection. The goal is to overcome early convergence and poor utilization of near-optimal solutions. The first PSO variation offered uses rectified personal and global best signals, spiral search-based local exploitation, swarm leader augmentation using Gaussian distributions, and mirroring and mutation to enhance worst solutions. The second proposed PSO model improves on the first by adding four new strategies: adaptive exemplar breeding with multiple optimal signals, nonlinear function-oriented search coefficients, exponential and scattering swarm leader selection procedures, and worst solution enhancement. They outperform 15 classic and advanced search strategies for discriminative feature selection across 13 data sets.

Abualigah et al. (2020) say that picking the most important characteristics in data mining pre-processing is key to developing new useable features. Using the informative subset, a classification model was more accurate than one built from all features. This technique can enhance data mining speed and computing time. SCA and GA were combined in this study to choose the optimal characteristics quickly and efficiently. This hybrid method outperformed SCA and GA by two times. It will be compared to the original Sine Cosine Algorithm (SCA) and other comparable algorithms, such as Ant Lion Optimization (ALO) and Particle Swarm Optimization (PSO), using 16 datasets from UCI Machine Learning library.

Hasnony et al. (2020) built models for Huge Data feature selection, yet processing massive data remains difficult. Massive volumes of data hamper data mining. Feature selection is a pre-processing step to discover the most informative features and improve classification accuracy. Choosing features takes time. PSO and a new binary version of grey wolf have wrapper features. Euclidean separation matrices optimize the KNN classifier. A chaotic tent map helps the algorithm avoid a local optimal solution. Sigmoid functions transform a continuous vector into a binary vector for feature selection. Cross-validation K-fold prevents machine learning model overfitting. Grey wolf and particle swarm optimization have yielded comparable outcomes. Using feature ratio, classification accuracy, and calculation time, 20 datasets are studied to evaluate the proposed model's performance and efficacy. PSO and GWO recovered 336 and 393 features, respectively, from 20 datasets. Overall accuracy is 90%, compared to 86.8% and 81.6% for the other techniques. PSO and GWO take 245.6 and 272 seconds to process all datasets.

Khiarak et al (2019) authors describe a strategy that uses an adjusted fluffy min-max classifier (ICAMFMCN) to determine the ideal subset of features. They then consider credit probability to improve classification accuracy and system scalability. The classification's performance is endorsed and perceived when an actual credit set is selected from a UCI dataset. Assets prove classification accuracy. Exploratory results show that existing data mining technologies can be included into ICA-MFMCN.

Allam et al. (2017) summarized current feature dimension reduction strategies based on improvement algorithms and offered another strategy to optimize arrangement space. Houari et al. (2016) developed a dimensionality reduction strategy based on Copulas and Forward Substitution to regulate information given in a high-dimensional space and address the impact of excess measurements on the final results. This method works well for dimensionality reduction on the following datasets: Diabetes, Waveform, Smartphone Human Activity Recognition, and Thyroid Datasets.

3. Methodology

The most important component in output features is clustering, which is accomplished by adding a kernel function to c-means in order to obtain linear and nonlinear relationships when mapping the data. The kernel combination of fuzzy relations identifies the rule set theory from the marginal weights of the micro array dataset. By selecting non-redundant characteristics in the majority of cases. The issue is significant for some, actual clustering applications in which several data points are used to choose features in biological applications. To use kernel-based clustering to such applications, it is frequently necessary to combine the entire features from multiple sources into a single high-dimensional data set. By utilising this weighting scheme, attributes are changed with numerous weights to fit the observed feature value for further classification.

Feature selection has demonstrated its usefulness in a variety of applications by allowing for the construction of simpler and more comprehensive models for the purpose of decreasing categorization issues, optimising learning performance, and providing clean, intelligible data. The fruit fly algorithm is centred on the objective in order to perform feature selection and hence improve learning performance. This approach divides the population of selected fruit flies into two groups, with one group searching for the optimal solution in a large space and the other searching for the optimal solution in a small space. The specification optimizes the characteristic for algorithm convergence.

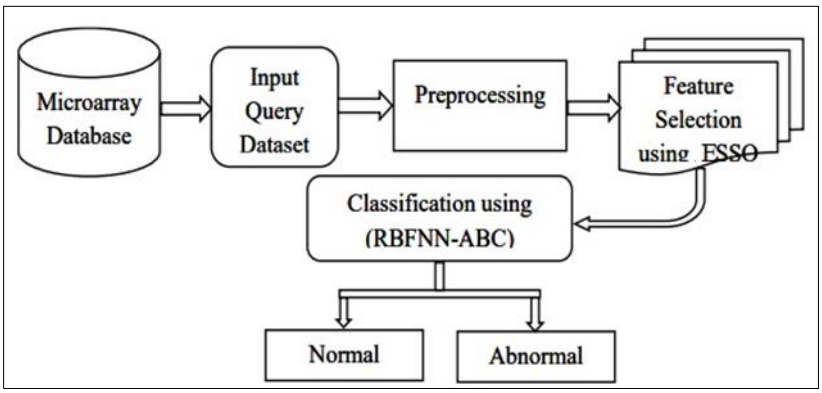

Figure 1 Optimal feature selection

The fruit fly method is based on the search element of sensing redundant features and weighting them closer to the optimal value over multiple iterations. The iteration is based on the feature value's maximum and minimum value

subset identification. Following that, the classifier is responsible for the selected features. The order is carried out here by applying the Optimized RBFNN technology to collect microarray data as common or exceptional. Techniques such as the Artificial Bee Colony (ABC) algorithm helps to improve the appropriateness of RBFNN.

There are three parts to the methodology. First, there is pre-processing to remove any noise; second, optimal feature selection; and third, classification to determine the appropriate category for each piece of data.

3.1. Pre-processing

The occurrence of alarming repercussions may occur when the calculation of dataset information is available, utilising partially separated non-value data to manage the concerns. As a result, it is critical to have a more accurate portrayal of information alongside its brilliance prior to its appraisal. Due to the proximity of worthless information, learning distinguishing proof in the training procedure becomes exponentially more unpleasant. The term "handling" is used more frequently in relation to information planning, and additionally distinct steps are necessary. The data set containing non-essential information is urged to undergo preprocessing. Insignificant data may manifest as clamour, irregularity, and misplaced characteristics. The proximity of extraneous data worsens the yield. With the final goal of resolving such challenges in mind, preprocessing is preferred, as it results in good outputs. When it comes to the information mining task, data pre-handling is critical. It is used for both the training and transformation of unique datasets. Our proposed technique is designed with the end goal of isolating all-out data and making use of non-straight-out data pre-handling. No longer-needed data is discarded, and steps continue to use the entire data.

3.2. Selection of Features

The process of selecting a subset of uninteresting characteristics (factors, pointers) for use in a scheduled display update is known as feature selection. It is also known as variable selection, trademark selection, or variable subset selection. Here, ideal segments are selected through the use of Enhanced Social Spider Optimization (ESSO) modeling. The conventional Social Spider Optimization technique is improved here with the assistance of fly improvement figuring. MSSO has been framed using the Fruit Fly Optimization (FFO) as a result of SSO.

3.3. Enhanced social spider optimization (ESSO) algorithm

The ESSO anticipates that the entire request space will be an open web in which all social spiders will communicate with one another. Each inclination plan within the request space is used to represent an area in the aggregate web in the suggested approach. All spiders acquire characteristics based on the health assessment of the system to which the social spider communicates. With the eventual goal of improving feature selection, the renewed spiders are gradually refreshed using fruit fly optimization calculations. The fruit fly demonstrates an iterative method for predicting features at a great distance by sensing the closest data and applying it to the succeeding features. The fruit fly search algorithm incorporates ideal conditions, such as redundant attributes, in order to choose the desired feature.

3.4. Population Update

The algorithm generates two different inquiry operators (spiders) that belong to a discrete class. Depending on sexual orientation, each discrete value max is focused by a technique including a variety of developing operators that simulate characteristics conducive to the behaviours that are typically expected inside the state of the dataset.

3.5. A neural network is used to determine the classification's best radial function.

ORBFNN employs the parameters as a component of the RBFNN (number of neurons, their individual focuses, radii and streamlined with the assistance of phoney honey bee settlement computation). Using the help of Broom head and Lowe, RBF networks were incorporated into the neural network building process. A network of nearby units is more likely to help the development of various near-reaction components of human identity, in contrast to more traditional models. Many sections of an eager plan can generate neurons with a near-tuned response characteristic. For example, a feature selection plan, a neural network, or a cell that sees to minor social events or new visual features within the search field can all generate neurons with this feature.

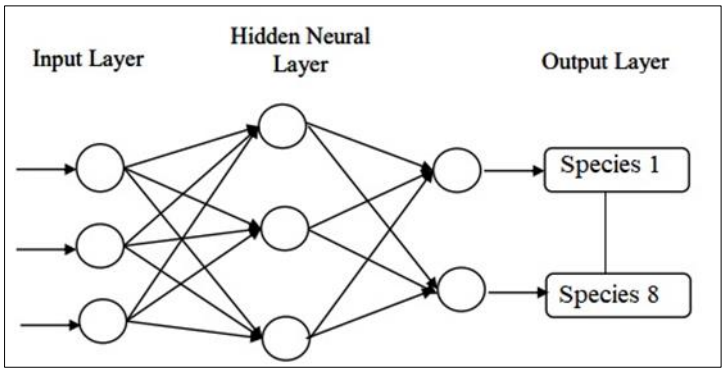

Figure 2 Structure of proposed RBFNN technique

In the RBFNN, the input layer, the hidden layer, and the output layer are all three layers of the network. Data classification is done using the RBFNN. The RBFNN's requirement to work with a defined purpose is met by applying the data's optimally chosen aspects.

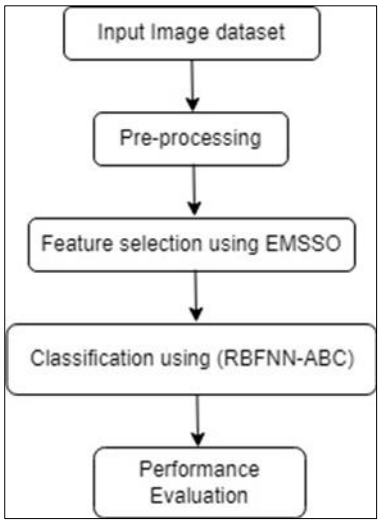

Figure 3 Flow chart of proposed work

The data in the input layer is balanced. With the aid of a non-facilitate constraint, the concealed layer converts the data from the information space to the unnoticeable locale. The Gaussian measure is typically employed in RBF hidden nodes. The output layer is immediate and directs the network's response. It has been observed that if a sufficient number of units are transmitted, an RBFNN can become unverifiable. The abundance of RBFNN is increased with the use of a forgery honey bee settlement state computation technique such as ABC.

3.6. Steps involved in ABC Algorithm

- Stage 1: is to initialize the data source that will sustain the micro arrays.
- Stage 2: Fitness Assessment
- Stage 3: The third stage, known as the employed bee stage
- Stage 4: Evaluation of the new source's viability comes at stage 4 of the process.
- Stage 5: Greedy selection takes place during the fifth and final stage.
- Stage 6: Observer ABC assessment
- Stage 7: The seventh stage involves the organization of the artificial bee network.
- Stage 8: The eighth stage is the Stop Criteria. Repeat stage2, until an unrivalled health or most extreme number of accentuations are met, at this point the course of action which is holding the best health respect is picked, and it is resolved as the best parameters for speculation.

4. Conclusion

Predicted features can now be selected and classified in a new way. A high-dimensional microarray dataset is used to estimate an improved social spider optimization technique in order to identify the best features. Using an ORBFNN classifier and an effective feature selection technique, the arrangement is carried out from that moment onward (ESSO). Based on a variety of metrics, including precision, accuracy, recall and F1 score on the Feature selection using MSSO, it can be concluded that the proposed grouping system surpasses the earlier methods on five benchmark datasets.

Flowering plant dataset is reported to be 94% accurate. Results show that the proposed ORBFNN classifier-based system outperforms existing techniques in terms of precision, affectability, and specificity. In other words, our approach offers the best possible framework for arranging data. A more dynamic estimation of optimal execution performance will be possible in the future, allowing the analyst to use a variety of feature selection methodologies.

References

[1] Rostami, O., & Kaveh, M. (2021). Optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC) and support vector machine (SVM): a combined approach of optimization and machine learning. Computational Geosciences, 25(3), 911-930.

[2] El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., & Mostafa, R. R. (2020). Improved feature selection model for big data analytics. IEEE Access, 8, 66989-67004.

[3] Xie, H., Zhang, L., Lim, C. P., Yu, Y., & Liu, H. (2021). Feature selection using enhanced particle swarm optimisation for classification models. Sensors, 21(5), 1816.

[4] Abualigah, Laith & Aldulaimi, Akram & Al Shinwan, Mohammad & Shehab, Mohammad (2019), A Proposed Hybrid Feature Selection Method for Data Mining Tasks, International Journal of Science and Applied Information Technology

[5] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M. R., Razeghi, F., Mazaheri, S., Zamani-Harghalani, Y., & Moosavi-Tayebi, R. (2020). New hybrid method for feature selection and classification using meta-heuristic algorithm in credit risk assessment. Iran Journal of Computer Science, 3(1), 1-11.

[6] Azhagu Sundari, B and Antony Selvadoss Thanamani, 2013, 'Feature Selection based on information gain', International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, no. 2.

[7] Houari, R, Bounceur, A, Kechadi, MT, Tari, AK and Euler, R, 2016, 'Dimensionality reduction in data mining: a copula approach', Expert Systems with Applications, vol. 64, pp. 247-260.

[8] Cruz, DPF, Maia, RD, da Silva, LA and De Castro, LN, 2016, 'Beerbf: A bee-inspired data clustering approach to design rbf neural network classifiers', Neurocomputing, vol. 172, pp. 427-437.

[9] Allam, M., & Nandhini, M. (2017). A study on optimization techniques in feature selection for medical image analysis. International Journal on Computer Science and Engineering (IJCSE), 9(3), 75-82.

[10] Abualigah, Laith & Aldulaimi, Akram & Al Shinwan, Mohammad & Shehab, Mohammad (2019), A Proposed Hybrid Feature Selection Method for Data Mining Tasks, International Journal of Science and Applied Information Technology.

[11] Pan JS, Hu P, Chu SC (2021) Binary fish migration optimization for solving unit commitment. Energy 226:120329

[12] H.Mehmood, "A Review of Feature Selection Techniques in Bioinformatics", Pakistan Journal of Art and Culture, vol.1, no.1, (2018).

[13] Too J, Abdullah AR (2021) A new and fast rival genetic algorithm for feature selection. J Supercomput 77(3):2844–2874

[14] Agrawal P, Ganesh T, Mohamed AW (2021) Solving knapsack problems using a binary gaining sharing knowledge-based optimization algorithm. Compl Intel Syst pp. 1–21

[15] Yan C, Liang J, Zhao M, Zhang X, Zhang T, Li H (2019) A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. Anal Chim Acta 1080:35–42

[16] Arslan S, Ozturk C (2019) Multi hive artificial bee colony programming for high dimensional symbolic regression with feature selection. Appl Soft Comput 78:515–527

[17] Liu Y, Wang Y, Ren X, Zhou H, Diao X (2019) A classification method based on feature selection for imbalanced data. IEEE Access 7:81794–81807

[18] BenSaid F, Alimi AM Online feature selection system for big data classifcation based on multiobjective automated negotiation. Pattern Recognit Page no.107629, 2021.

[19] Rostami M, Berahmand K, Nasiri E, Forouzande S, Review of swarm intelligence-based feature selection methods. Eng Appl Artif Intel, 2021.

[20] Alomari OA, Makhadmeh SN, Al-Betar MA, Alyasseri ZAA, Doush IA, Abasi AK, Zitar RA (2021) Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators. Knowl-Based Syst 223:107034.

[21] Mansour NA, Saleh AI, Badawy M, Ali HA (2021). Accurate detection of Covid-19 patients based on feature correlated naïve bayes (FCNB) classification strategy. J Ambient Intel Hum Comput 1–33

[22] M.Qaraad, S. Amjad, I.I.Manhrawy, H.Fathi, B.A.Hassan, and P.El.Kafrawy. "A Hybrid Feature Selection Optimization Model for High Dimension Data Classification." IEEE Access, vol. 9, (2021), pp.42884-42895.

[23] Anil K Jain. Fundamentals of digital image processing. Englewood Cliffs, NJ: Prentice Hall, 1989.

[24] Bl Basavaprasad and M Ravi. A study on the importance of image processing and its applications. IJRET: International Journal of Research in Engineering and Technology, 3, 2014.

[25] Robert J Schalkoff. Digital image processing and computer vision, volume 286. Wiley New York, 1989.

[26] Aggelos K Katsaggelos. Digital image restoration. Springer Publishing Company, Incorporated, 2012.