Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996

P-ISSN: 2958-5988

Improving Medical Dataset Accuracy with PSO Feature Selection and Machine Learning Classification Techniques

Ahsan Masroor*[1], Dr. Muhammad Affan Alim[2], Dr. Waleej Haider[3], Syed Affan Hussain[4]

[1] College of Computing and Information Sciences, Karachi Institute of Economics and Technology, Karachi, Pakistan. *Corresponding author: ahsanmath@yahoo.com

[2] Department of Computer Science, IQRA University, Karachi, Pakistan.

[3,4] Department of Computer Science, Sir Syed University of Engineering & Technology, Karachi, Pakistan.

DOI: https://doi.org/10.63163/jpehss.v2i4.1389

Abstract

The redundancy in the dataset usually effects the performance of the model in the sence of accuracy, computational time, cost of storage and rilaiability of the data analysis. A cleaned and noise free data can help achieving satisfactory performance of the model. The one-vs-rest approach serves as the fitness function for Particle Swarm Optimisation (PSO) to resolve the classification problem. The global optimization in machine learning reduces or removes irrelevant redundant data to provide accuracy. A good method of feature selection includes investigation, sample classification to avoid incomprehensibility. In this paper swarm optimization is used to implement feature selection. The support vector machines and one verses rest method prove to be the fitness function of PSO classification problem. Moreover, In applications such as data mining, machine learning, medical data processing, and pattern classification, feature selection is vital. Furthermore, our test results show that PSO-based feature selection improves machine learning model performance on medical datasets accuracy of 98.13%, thereby enabling successful and accurate diagnosis and forecast

Keywords: *Machine learning (ML)Particle swarm optimization (PSO), Random Forest(RF), Support vector machines (SVM)*

INTRODUCTION

Cancer kills over 10 million individuals each year. Cancer's arising significance as the main cause of death mirrors a sharp decline in the incidences of strokes and coronary cardiac disease in many nations. Cancer is often categorized into two types: malignant and benign. Benign cells cannot spread quickly in the human body; hence they are classified as noncancerous. Malignant stage is treated as cancer-causing because cancer cells multiply, damaging neighboring organs and spread cancer through the body.Early identification and treatment of cancer are critical to achieving a cure and increasing life. Nowadays, improvements in machine learning and deep learning models are becoming more important and helpful in identifying cancer. This review investigates several current research articles over the last 5 years on the uses of machine learning ,image processing and deep learning (DL) addresses in automated identification and classification of many malignancies using different types of images. A critical stage in reducing the dimensionality of dataset, feature selection improve the competence, interpretability and simplification. Dataset is the most informative properties in machine learning and data analysis.

A subset of feature is selected that effectively represents the inherent structure that and associate in the data while eliminating unwanted or simulate data. Parameter nomination can optimize the machine

learning model efficiency, interpretability, and clarity by minimizing dataset dimensionality. Nominating parameter is critical step in machine learning since it can markedly optimize the reliability of grouping algorithm by decreasing dimensionality. Additionally, many practical function the data set are elevate dimensiontional leading parameter nomination complicated.The expletive of dimensionality can top to overfitting, unfortunate performance, and enlarged computing difficulty. Consequently, categorizing an optimal selection of acceptable characteristics is critical for improving classification accuracy and lowering computing costs. Methods for feature selection optimize the accuracy, interpretability, and efficiency of machine learning models by selecting the data's most relevant subset of features. Feature selection techniques may be classified in various different ways. The most prevalent classification is among filters, wrappers, embedding, and hybrid approaches.

Filter Methods

Filter approaches choose features based on performance criteria, regardless of data modelling procedure. Only after identifying the best traits, Modelling algorithms can make use of them. Filter approaches can score single characteristics or assess whole feature subsets. The first step in filter feature selection is to create an evaluation tool that will be utilized to evaluate the significance of each feature. The four most frequently used assessment metrics in filter approaches are variance, mutual information, chi-squared test and correlation coefficient.

Next, using the given evaluation metric, compute the relevance score of each feature. The characteristic is more relevant to the target variable if the relevance score is higher. After calculating the relevance score for each feature, rank the features in descending order of relevance score. This ranking serves as a foundation for picking the most relevant elements for study. Finally, depending on the rank order, choose the best attributes and use them for additional analysis. The number of characteristics to choose from is determined by the complexity of the analysis and the model's performance. Each of these strategies takes benefits and drawbacks, and the method used is determined by the type of data and the situation at hand. Although filter techniques are effective and simple to use, they may not always capture the interactions between characteristics and the target variable. As a result, they may overlook essential qualities that are only meaningful when combined with others.

Wrapper Methods

Wrappers act as a black box, evaluating subsets of features based on how effectively a modelling technique performs. So, for classification jobs, a The wrapper evaluates subsets depending on classifier performance (such as Naïve Bayes or SVM). When clustering, a wrapper evaluates subsets based on how well an individual method for clustering works (K-means, for example).

Embedded Method

Embedded approaches differ from Filter and Wrapper methods because It is difficult to separate the procedures for feature selection and learning.

Embedded methods Reduce time by doing feature selection throughout the learning process, removing the requirement for two-step induction, as shown in the wrapper technique [3]. Embedded approaches beat wrappers in terms of making better use of accessible data without the need for distinct sets both validation and training. Furthermore, they can find a solution faster because every examined piece of data doesn't need a predictor to be updated from scratch.

To solve optimization problems we use Particle Swarm Optimization (PSO) algorithm that also inspired by the feature selection and social behavior. In PSO, a swarm of particles moves through a exploration space to find the optimum solution. Every particle updates its position based on its own experience as well as the observations of its neighbors, representing a possible solution. PSO is a powerful tool for feature selection, as it can efficiently search for the optimal subset of features in a high-dimensional space

Literature Review

A novel approach to heart disease prediction has been developed by combining Naïve Bayes and PSO. Prior to applying the Naïve Bayes method, important characteristics were selected using PSO. Researchers found that the classification accuracy was higher than when Naïve Bayes was used alone [1]. The verified results show that PSO reduces feature space without compromising prediction accuracy. The Flower Pollination approach is used to optimize isomap features, which are then classified using the Real Adaboost classifier [2]. PA-KNN represents optimizations for K-NN using PSO and ACO techniques. A comparison between multiple techniques for data mining on the heart disease dataset. Determining an effective based on performance method for heart disease prediction[3]. This method provides a novel approach called CMTMSOM, that makes use of conic quadratic programming, an effective instrument for convex and continuous optimization. Focusing on the mean-shift outlier regression model, this approach has been invented by integrating the resilience of M-estimation with the stability of Tikhonov regularization [4]. Artificial neural networks trained with stochastic gradient descent worked more effectively than the other methods, reaching a 95% accuracy in classification and a 14.69 average deviation reduction.

[5]. The portion of a feed forward neural network (FNN) that will improve its capacity for generalization and rate of convergence (learning speed); to identify new fields of inquiry that will help scientists create new [6]. In the ML method, normalization techniques like sigmoid normalization improve the accuracy of trained neural networks[7].The illness was previously predicted using the Ridge-Adaline Stochastic Gradient Descent Classifier (RASGD). The categorization model is regularised utilizing weight decay methods, namely least absolute shrinkage, selection operator, and ridge regression. To minimise the cost function of the classifier[8]. To diagnose cardiac illness, four machine learning models are used, including random forest (RF), decision tree (DT), AdaBoost (AB), and K-nearest neighbour (KNN)[10]. A generalised method was developed to assess the strength of the key parameters that influence heart disease prediction.notably RF and KNN, demonstrate good accuracy, in internet-based cloud hosting platform.[9][11]. The algorithm known as PSO evaluates feature subsets based on SVM classification accuracy in this code's wrapper-based feature selection method by maximising the fitness function, the PSO particles look for the optimal subset of features.

Comparing to filter or embedding techniques, this study's wrapper-based PSO approach provides accurate and model-specific selection by evaluating feature subsets based on SVM performance. Despite being computationally demanding, it guarantees improved feature relevance and classification accuracy, which is consistent with the goal of the study.[16].

I. PROBLEM STATEMENT AND ITS PROPOSED SOLUTION

A. Problem Description

This method intentions to determine the influence of Particle Swarm Optimization (PSO) in feature assortment for cataloguing models (SVM and RF). The assessment will expose if PSO improves classification correctness or not. The emphasis is on assessing the efficiency of PSO as a feature assortment technique by associating performance metrics (like accuracy) among models that use it and those that don’t.

B. Solution Framework

This context suggests a step-by-step explanation for associating two methods to attribute selection and cataloguing by Particle Swarm Optimization (PSO). The attention is on investigating the effect of PSO on cataloguing accuracy using Support Vector Machine (SVM) and Random Forest (RF) classifiers.

II. Methodologies and Techniques

To illustrate the various features of the proposed feature selection technique, two separate scenarios

have been explored. First scenario Feature selection is performed out with PSO, and after which the feature chosen are subjected to the Support Vector Machine (SVM) & Random Forest (RF) algorithms. Subsequently the models' accuracy will be assessed Second scenario, PSO is not applied throughout Feature selector similar applications are utilized to estimate the accuracy of the RF and SVM models. The proposed model compares FS performance with and without PSO to show how PSO impacts model accuracy.

Table 1:Sample dataset used for experimentation

	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	...	gene_16373	gene_16374	gene_16375	gene_16376
0	0.0	2.017209	3.265527	5.478487	10.431999	0	7.175175	0.591871	0.0	0.0	...	8.750533	7.421257	4.692126	1.334282
1	0.0	0.592732	1.588421	7.586157	9.623011	0	6.816049	0.000000	0.0	0.0	...	6.638879	7.991732	5.709045	0.811142
2	0.0	3.511759	4.327199	6.881787	9.870730	0	6.972130	0.452595	0.0	0.0	...	8.205754	10.375778	1.839758	0.000000
3	0.0	3.663618	4.507649	6.659068	10.196184	0	7.843375	0.434882	0.0	0.0	...	8.093185	8.424771	5.502251	0.434882
4	0.0	2.655741	2.821547	6.539454	9.738265	0	6.566967	0.360982	0.0	0.0	...	7.522228	12.176650	10.305423	0.360982

5 rows × 16383 columns

The dataset comprises DNA or gene-expression sequences of patients, encompassing about 16,383 gene characteristics (from gene_0 to gene_16382). Each row corresponds to an individual patient, with numerical values denoting gene expression levels (e.g. gene_1 = 2.017209, gene_4 = 10.431999,etc.) A sample dataset is presented in the Table 1.

This dataset is high-dimensional due to its large number of features, which poses issues common to gene-expression or DNA-sequencing datasets. The primary objective of utilizing this dataset is to employ machine learning classifiers and feature-selection techniques to accurately predict or classify cancer patients versus non-cancer individuals, or to distinguish among different cancer types.

Microarray data are often highly redundant, noisy, and unequal in their dimensionality[12]. Many genes are considered irrelevant to the investigated classes. In this study, a novel selection of features technique based on PSO has been proposed for classification of highly dimensional cancer microarray information. Feature selection significantly enhances the effectiveness of machine learning models, particularly in medical datasets where the number of features is frequently high. In this paper, we present a unique method for selecting features in medical datasets using Particle Swarm Optimization (PSO)[13]. Overfitting, inferior model performance, and higher computing complexity may result form the curse of dimensionality.

To optimize classification efficiency pso inspires social behavior to find the best attribute subset. PSO is applied to medical dataset by using various classification method and evaluate to check how model efficiency is impacted. A fitness function is designed to evaluate the capacity of a particle’s selected attributes.

The UCI machine learning repository is examined for cancer gene dataset upgraded the cancer gene dataset that include expression stage and labels for each gene is evaluated[14]. Response such as pandas or numpy ,data can be successfully imported and processed into a parameter matrix (x) and label vector (y) facilitates machine learning tasks. The PSO optimisation has been set for (the number of iterations) iterations using (number of particles) particles. For each dimension, the particle's position range are defined as (particular range, e.g., (min, max)). (fitness function) is a fitness function that is used for optimisation. So as to ensure a successful and reproducible optimisation procedure, these parameters were chosen using information. [15].

Parameter nomination substantially optimize the precision of machine learning models, significantly in medical dataset with ample feature.in this study, we detect a normal approach for introducing feature in medical dataset using pso, we can choose parameter using random forest and support vector machine. svm weight feature during the training stage focused on their consequence on the decision marks, pondering attributes with larger weight as more essential. examining these weight after training

can half in detecting the essential component. Random Forest to tackle non linearity, component interaction and overfitting creates a strenthful parameter nomination criteria.

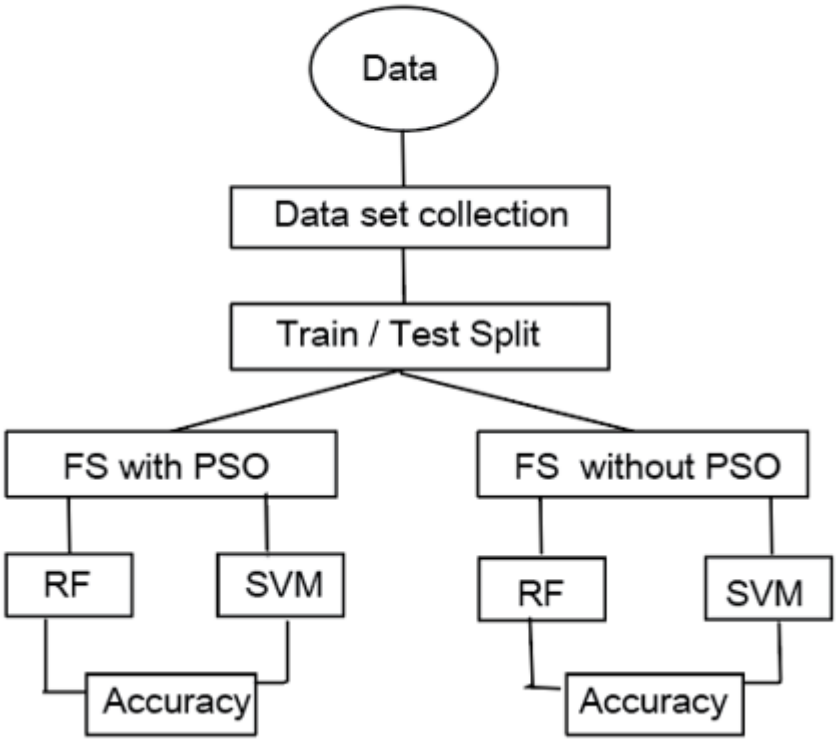

Figure 1: Proposed Model with Different Optimization Techniques

IV.Result Discussion:

Performance increased when we applied PSO with SVM. The original parameter of the SVM when the accuracy was 38.50% we kept it.Using PSO to modify the hyper parameters particularly fine-tuning gamma to 0.001, we saying a significant increase in accuracy, success 98.13%. This discovery highlights PSO's worth in fine-tuning SVMs and maximizing their forecast powers. Remarkably, the gamma parameter, which pedals the consequence of individual training data, had a serious role in finding this significant accuracy rise. The accuracy increase from applying PSO for feature selection and fine-tuning the SVM hyper parameters is significant. Our findings demonstrate the potential for PSO to improve the efficiency of machine learning mathematical models, especially when paired with focused hyper parameter adjustments.

A comparative study has been prented in Table 2. Fangfang Chen et al. has worked on human blood cell analysed and used FTIR Spectroscopy Data for expetimentation. They have used SVM algorithm and achieved 87.07% accuracy wth PSO. Similarly, a breast cancer analysis has been performed in [18] using PSO-SVM (with Cuckoo Search). They have achieved 89.02% accuracy. It is observed that the proposed model have achieved a significant improvement in the accuracy.

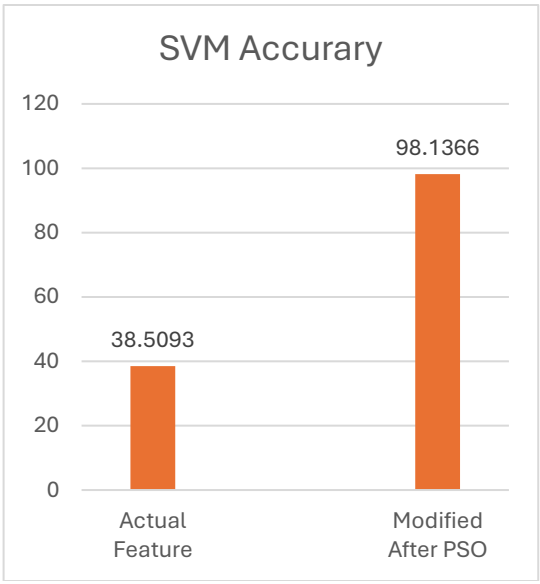

Figure 2: Accuracy of SVM on Proposed Features

In below diagram Figure 2 the results show the ability of PSO to increase the accuracy of random forest. The significant rise in accuracy from 89.2187 to 96.7187. In RF the accuracy has increased by around 7.444%.PSO is an optimization strategy that identifies relevant attributes and improves model performance by reducing noise and unnecessary information. The 7.444% gain in accuracy might be attributed to a well-balanced choice of hyper parameter along with suitable feature selection using PSO. These choices have resulted in a strong and efficient model capable of detecting essential patterns in data while avoiding overfitting

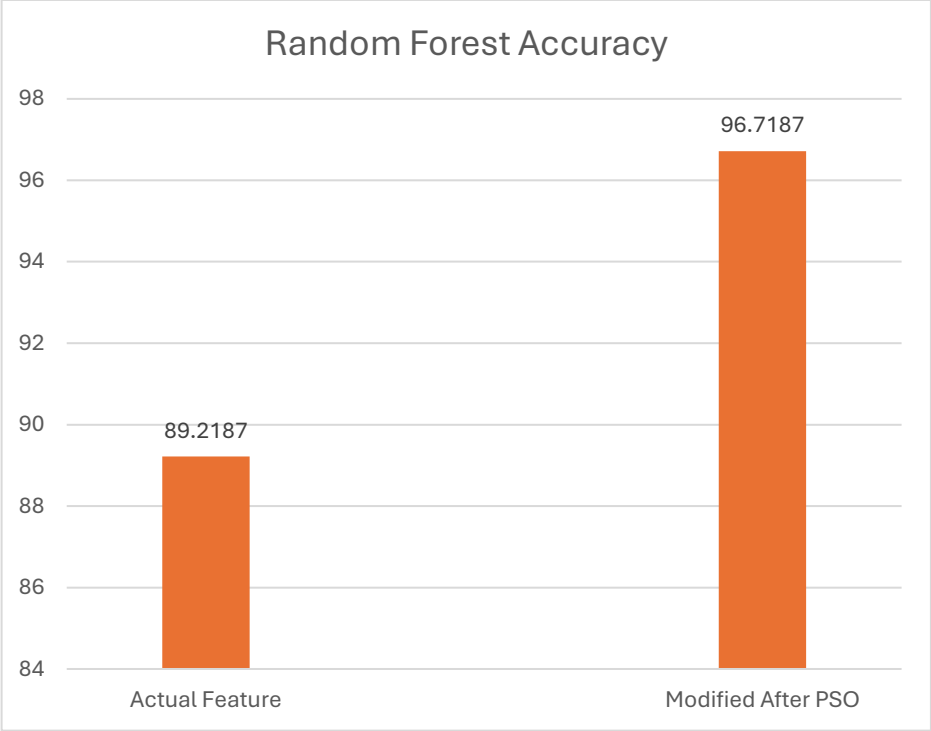

Figure 3: Accuracy of Rendome Forest Algorithm with Proposed Features

Table 2: A Comparative Analysis of the performance of Existing and Proposed Model

Algorithm	**Feature Selection Method**	**Accuracy with PSO**	**Accuracy without withou PSO**	**Refrences**
SVM	Wrapper based PSO+SVM	98.13	38.50	Proposed
RF	Random forest classifier	96.7187	89.2187	Proposed
SVM	PCA + PSO-SVM wrapper hybrid	87.07	83.62	[17]
RF,SVM	PSO-SVM (with Cuckoo Search)	89.02	85.46	[18]

V. Conclusion and Recommendations

The combination of PSO feature selection and SVM/Random Forest classification yields a robust and efficient approach that It could significantly improve the efficiency of machine learning models. particularly in scenarios with large feature dimensions and feature relevance critical for accurate prediction. The UCI dataset's results lead to the following conclusions. The initial actual feature count is 16384. It was then modified, resulting in a feature count of 1093, with an accuracy of 98.1366%.

REFERENCES

[1] Dulhare, U. N. (2018). Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomedical Research, 29(12), 2646-2649.

[2] Prabhakar, S. K., & Rajaguru, H. (n.d.). A framework for schizophrenia EEG signal classification optimization algorithms. IEEE

[3] Khourdifi, Y., & Bahaj, M. (n.d.). K-nearest neighbour model optimized by particle swarm optimization and ant colony optimization for heart disease classification. In Proceedings of the International Conference on Big Data and Smart Digital Environment

[4] Taylan, P., Yerlikaya-Özkurt, F., Bilgiç Uçak, B., & Wilhelm, G. (2021). A new outlier detection method based on convex optimization: Application to diagnosis of Parkinson's disease. Journal of Applied Statistics. Taylor & Francis.

[5] Govindarajan, P., Soundarapandian, R. K., & Gandomi, A. H. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications. Springer

[6] Khan, W. A., Chung, S. H., Awan, M. U., & Wen, X. (2020). Machine learning facilitated business intelligence (Part I): Neural networks learning algorithms and applications. Industrial Management & Data Systems. Emerald

[7] Dimitrovska, I., Malinovski, T., & Krstevski, D. (n.d.). Machine learning solution based on gradient descent algorithm for improved business process outcomes

[8] Deepa, N., Prabadevi, B., Maddikunta, P. K., & Gadekallu, T. R. (2021). An AI-based intelligent system for healthcare analysis using Ridge Adaline Stochastic Gradient Descent classifier. The Journal of Supercomputing

[9] Absar, N., Das, E. K., Shoma, S. N., Khandaker, M. U., & Miraz, M. H. (n.d.). The efficacy of machine-learning-supported smart system for heart disease prediction. Health Care Sciences & Services

[10] Louridi, Nabaouia, Samira Douzi, and Bouabid El Ouahidi. "Machine learning-based identification of patients with a cardiovascular defect." Journal of Big Data 8.1 (2021): 133

[11] Folorunso, Sakinat Oluwabukonla, et al. "Heart disease classification using machine learning models." International conference on informatics and intelligent applications. Cham: Springer International Publishing, 2021
[12] Bolón-Canedo, Verónica, et al. "A review of microarray datasets and applied feature selection methods." Information sciences 282 (2014): 111-135
[13] Alrefai, Nashat, and Othman Ibrahim. "Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets." Neural Computing and Applications 34.16 (2022): 13513-13528
[14] Alhenawi, Esra'A., et al. "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." Computers in biology and medicine 140 (2022): 105051
[15] Shami, Tareq M., et al. "Particle swarm optimization: A comprehensive survey." Ieee Access 10 (2022): 10031-10061.
[16] Wong, M. T., He, X., Yeh, W. C., Ibrahim, Z., & Chung, Y. Y. (2014, November). Feature selection and mass classification using particle swarm optimization and support vector machine. In International Conference on Neural Information Processing (pp. 439-446). Cham: Springer International Publishing.
[17] Fadlelmoula, A., Catarino, S. O., Minas, G., & Carvalho, V. (2023). A review of machine learning methods recently applied to FTIR spectroscopy data for the analysis of human blood cells. Micromachines, 14(6), 1145.
[18] Liu, X., & Fu, H. (2014). PSO-Based Support Vector Machine with Cuckoo Search Technique for Clinical Disease Diagnoses. The Scientific World Journal, 2014(1), 548483.