International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT12390542

Piecewise Linear Approximation-Driven Primal SVM Approach for Improved Iris Classification Efficiency

Shital Solanki*[1], Dr. Ramesh Prajapati[2]

*[1]Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India

[2] Professor, CE/IT Department, Shree Swaminarayan Institute of Technology, Gandhinagar, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 01 Oct 2023

Published: 15 Oct 2023

Publication Issue

Volume 9, Issue 5

September-October-2023

Page Number

286-292

ABSTRACT

Classification, a crucial aspect of machine learning, revolves around the meticulous analysis of data. However, the complexity of diverse life forms on Earth poses a challenge in distinguishing species that share similar attributes. The iris flower, with its subspecies exemplifies this challenge. The aim of the paper is to develop a methodology that not only enhances classification accuracy but also effectively addresses computational efficiency, facilitating faster and more practical categorization of iris patterns. This novel approach named Piecewise Linear Approximation based SVM (PLA-SVM) is applied to flower species classification and is benchmarked against alternative machine learning techniques. Implementation is carried out utilizing MATLAB – GUROBI interface of and GUROBI Solver. The performance metrics such as accuracy, precision, F1 score and ROC – AUC Curve are used to compare proposed algorithm performance. This comprehensive analysis enables a comparative study of diverse algorithms, ultimately validating the proposed PLA-SVM technique using the Iris dataset. The numerical implementation results shows that the PLASVM outperforms the existing standard classifiers in terms of different performance matrices.

Keywords : SVM, Piecewise Linear Approximation, ROC, AUC

I. INTRODUCTION

In recent years, advancements in machine learning have led to the development of diverse algorithms that excel in solving complex classification problems, ranging from traditional methods like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees, to more sophisticated techniques such as neural networks and ensemble methods. Each algorithm brings its strengths and limitations to the classification task, necessitating careful consideration of algorithm selection, parameter tuning, and performance evaluation. Accurate and efficient classification of iris patterns holds paramount importance across diverse fields such as botany and biometrics. As the relevance of pattern recognition techniques continues to escalate, the pursuit of refining classification approaches remains ongoing. Support Vector Machines (SVMs) have firmly established themselves as a pivotal tool for tackling classification tasks. It has been applied very successfully in a variety of applications like classification, forecasting, and estimation in small-sample cases. However, SVM algorithms do come with a range of intricacies related to kernel selection, hyperparameter tuning, computational efficiency, handling large and imbalanced datasets, and balancing accuracy with model complexity that demand careful consideration. SVM training time can increase significantly as the dataset size grows, making it less suitable for very large datasets. The SVM model can consume a substantial amount of memory, especially for high-dimensional data, and SVM performance can even degrade when dealing with imbalanced datasets. Understanding and managing these complexities are essential for effectively utilizing SVMs in classification tasks. This research addresses fundamental challenges in SVM-based classification tasks by presenting Piecewise Linear Approximation based SVM, a unique approach that leverages the strength of linear programming which is proposed by Shital and Ramesh [1].

The findings of the present work may offer guidance in selecting appropriate algorithms for a variety of classification tasks beyond the realm of botany, reaffirming the significance of the Iris flower classification problem as a cornerstone in the field of machine learning.

The rest of the paper is arranged as follows: Section II gives detailed literature review of the iris classification problem. Section III explains detailed methodology of the proposed SVM. The experimental implementation results are explained in Section IV. The paper ends in conclusion by proposing concluding remarks in Section V.

II. RELATED WORKS

Several machine learning algorithms are commonly utilized for iris flower classification using benchmark iris flower dataset [2]. Some of these are decision trees, random forests, support vector machines, and K-nearest neighbours. T. Gupta and associates have done a comparative study using Logistic Regression Support Vector Machine and K-Nearest Neighbours for Iris flower species classification [3]. During their study, exploratory data analysis has been done to pre-process the dataset, which includes tasks like eliminating NULL, duplicate, and irrelevant values, ensuring data readiness for machine learning models. The models achieved impressive maximum accuracy scores of 96.43%, 98.21%, and 94.64% for logistic regression, SVM, and KNN, respectively. A ring Four-channels RC system approach was employed by Qian Yu and other fellow researchers to solve the Iris flower classification based on unidirectional coupled VCSELs [4]. This research explores the impact of several parameters on this RC system's recognition performance, such as bias currents, external injection strength, frequency detuning, feedback strength, coupling strength, and the number of virtual nodes. This RC system achieves a classification accuracy of 100%.

Hemalatha et al. conducted study on iris flower species classification [5]. The research compares various algorithms, such as Gaussian Naive Bayes, KNN, SVM, and Decision Tree. On the Iris dataset, the Gaussian Naive Bayes model outperforms other models with 100% accuracy. The study applies cross-validation and investigates numerous tactics, demonstrating the efficiency of the Gaussian Naive Bayes algorithm in the end. Poojitha et al conducted study using unsupervised clustering techniques like K-means and a neural network clustering tool in MATLAB, for effectively categorization of the Iris dataset into its species groups without requiring explicit supervision [6]. The ensemble classification technique also performed very well on iris classification by L. Pawar and a research group where the base model is used to classify the iris plant based on its flower pattern [7]. Later, an ensemble model is proposed to enhance the classification performance. This might include using techniques like voting, bagging, or boosting to combine the base models. Pinto et al implemented KNN, Logistic regression and SVM iris dataset classification using Scikit tool reported that SVM classifier gives best accuracy compared to KNN and logistic regression models [7].

In the present work, the implementation of the PLA-SVM on iris dataset is carried out using non-separable configuration of dataset that is using virsicolor and viriginica type of species of iris flower.

III. PROPOSED METHODOLOGY

Recently, Shital and Ramesh introduced and validated a novel approach for designing linear Support Vector Machines (SVMs) based on separable linear programming [1] [9]. They proposed Piecewise Linear Approximation Support Vector Machine (PLA-SVM) for linear separable dataset and successfully validated on the fault classification of a laboratory gas turbine engine. They derived the λ – formulation of the original SVM's objective function and constraints which converts the original non-linear quadratic optimization SVM design problem in to approximating linear programming problem which is simpler and faster. The linear PLA-SVM design ensures a global solution, due to the incorporation of mixed integer linear programming and branch and bound algorithms, core components of the GUROBI Optimizer solver [8].

In the present work, the linear PLA-SVM proposed by Shital and Ramesh [1] has been implemented on iris dataset using versicolor and virginica classes and petal length and petal width feature. Due to the non-separability of the present problem, the PLA-SVM proposed in [1] is modified to incorporate soft-margin primal form of SVM. The modification incorporates ξ as the slack variable representing the degree of misclassification and C as the regularization parameter, controlling the trade-off between maximizing the margin and minimizing the misclassifications. A larger C allows for fewer misclassifications but may result in a narrower margin. The PLA-SVM is designed using Primal quadratic optimization problem of SVM design and solved using GUROBI – MATLAB [10] interface of GUROBI solver. The solution of the proposed method will be the optimal values of SVM parameters w, b, and ξ's that satisfy all the constraints PLASVM. The implementation of the proposed algorithm includes the following steps:

1. Data collection /Data preparation
2. Exploratory Data Analysis (EDA)
3. Data pre-processing.
4. Obtain the λ – formulation of the original SVM's with added C and ξ's
5. Apply the PLA-SVM algorithm on training dataset
6. Testing the model on unseen data or testing data
7. Performance evaluation of the model.

The proposed PLA-SVM offers many benefits as:

1) It simplifies optimization by directly solving the primal optimization problem of SVM.
2) Solution depends on initial search domain of the SVM parameters and not on initial guess.
3) It can deal with large datasets efficiently due the linear programming approach and generalizes well.

IV. EXPERIMENTAL IMPLEMENTATION

In this section, we implement the PLA-SVM algorithm on Iris flower classification problem. The dataset has been explained in details. To show the effectiveness of the proposed approach, the PLA-SVM has been compared with the existing classifiers.

A. The Dataset

The Iris dataset is a widely used dataset from the UCI Machine Learning Repository. [2]. It was introduced by British statistician and biologist Ronald Fisher in 1936. This flower dataset contains three species, namely iris setosa, iris versicolor and iris virginica as shown in Figure 1. It is commonly used as a benchmark dataset for practicing and demonstrating various machine learning algorithms and techniques.

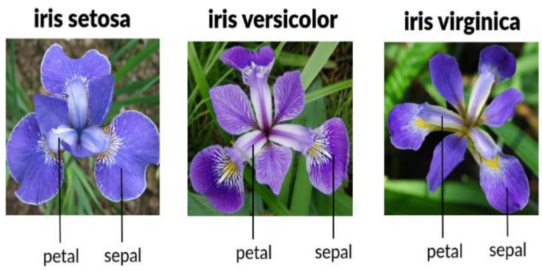
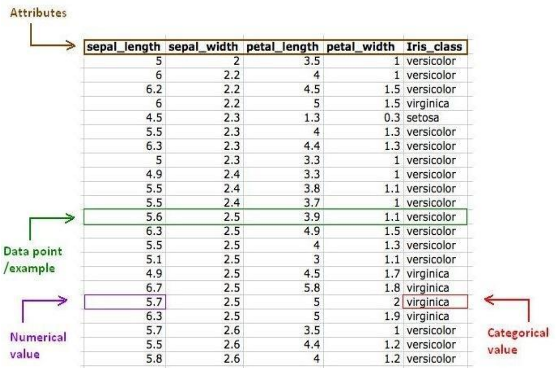


Figure 1: Iris flower subspecies

The dataset has 150 data samples; each class contains 50 samples. The dataset includes features like sepal length, petal length, sepal width, and petal width.as shown in Figure 2. In the present work, we have used



sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor

petal length and petal width features and classes used are versicolor and virginica.

Figure 2: Sample Iris dataset and its attributes

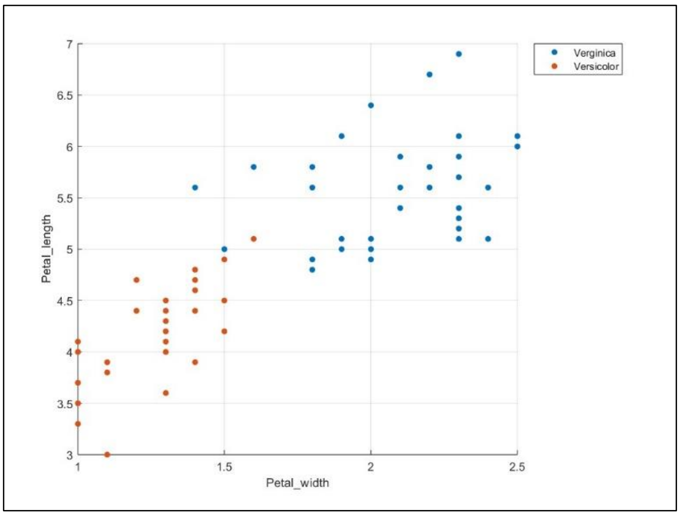


Figure 3: Scatter plot of an iris dataset showing Petal length and Petal Width

B. Implementation of PLA-SVM

Before creating the model and training it, the pre-processing of the dataset is required. To better understand the dataset's spread and relationships, a scatterplot has been generated. The scatter plot of an iris showing petal length and petal width is shown Figure 3 which indicates that the dataset is non-separable.

There are no null values in the dataset and no zeros. So, there is no need to handle missing or null values. Through the correlation analysis, we found a positive correlation exists in the iris dataset. This information is useful not only for exploring the data but also for making informed choices when constructing predictive models or deriving conclusions from the dataset. From observation, it is found that petal length's wide range (1.0-6.9) and highest standard deviation, among others, indicate significant variance among samples. The high standard deviation emphasizes its discriminative potential, making petal length an important feature in differentiating iris species.

To execute the implementation of PLA-SVM, the iris dataset is partitioned into two segments for training and testing using a 70:30 ratio giving 70 observations for training dataset and 30 observations as test dataset. We have used 5-fold cross validation and the value of regularization hyperparameter C is set to the value of 1 using grid search method.

To apply the PLA-SVM, we have set the initial search box for weight vector w = [w_1 w_2] and bias b as w_1 = [-5 50], w_2 = [-5 50], and b = [-5 50] respectively with breakpoints of 100 for all parameters. Set the initial search box for ξ_i's = [0 50] and breakpoints as 50, where i=1 to n, n=number of datasets that is 70 in this case. We have set the tolerance as 1.00e-04 which is default to GUROBI optimizer.

Now, PLA-SVM problem is set up now so load the iris dataset into MATLAB and run the PLASVM algorithm to train the model with optimal SVM parameters. Once the model is obtained, we have use test dataset to validate the model using performance matrices.

All implementation work was done on an Intel Core i7 processor running at 4.8 GHz, 8 GB of RAM, and 1TB of external storage, in MATLAB- GUROBI interface.

C. Discussion

The research encompasses an exploratory data analysis on and a Rigorous 5-fold cross-validation process, meticulously evaluating outcomes through performance metrics such as accuracy, precision, F1 score and Area Under the ROC curve.

The PLA-SVM implemented in MATLAB- GUROBI interface finds the optimal values of SVM hyper plane parameters in 0.7035 time. The optimal values obtained are

w_1_opt = 0.8661, w_2_opt = -1.643, b_{opt} = -0.3621. The Figure 4 shows the confusion matrix obtained with the PLA-SVM for test dataset.

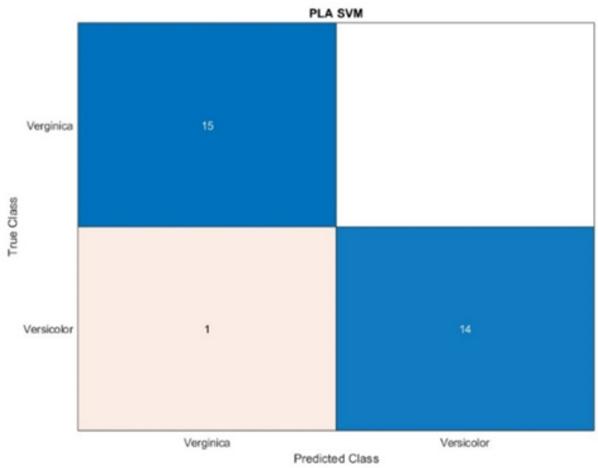


Figure 4. Confusion matrix of the obtained PLA-SVM Model

The optimal model obtained with the PLA-SVM algorithm is tested and validated on testing data. The obtained results are compared with the existing bench mark classifiers obtained using the same dataset and same settings. Table 1 shows the performance matrices such as accuracy, precision, F1 score, AUC score and training time for PLA-SVM along with other classifiers.

The PLASVM classifier distinctly stands out in performance, showcasing commendable metrics across the board. It boasts an impressive accuracy of 96.7%, an F1 score of 0.9677, and an almost impeccable area under the ROC curve at 0.9956. Furthermore, its remarkably swift training duration of 0.7035 seconds underpins its capacity for prompt and efficient outcomes. This swiftness suggests its potential to be adept even in scenarios demanding real-time processing or when navigating vast datasets.

Diving into the other models, we notice a striking uniformity in performance metrics among Logistic Regression (Kernel), Ensemble Bagged Trees, and Tri-layered Neural Network Decision classifiers. All three models register an accuracy of 93.3% and F1 scores of 0.9375. A salient feature across these models is their impeccable precision of 100%, emphasizing their adeptness in significantly reducing false positive errors, which is paramount in many applications.

Table 1: Performance of different Classifiers

Sr No.	*Classifier Name*	*Accuracy (%)*	*Precision (%)*	*F1 Score*	*Area Under the ROC Curve*	*Training Time (Sec)*
1	*Logistic Regression (Kernel)*	*93.3*	*100*	*0.9375*	*0.9867*	*4.4156*
2	*Ensemble Bagged Trees*	*93.3*	*100*	*0.9375*	*0.9800*	*4.1661*
3	*PLASVM*	*96.7*	*100*	*0.9677*	*0.9956*	*0.7035*
4	*Tri layered Neural Network Decision*	*93.3*	*100*	*0.9375*	*0.9600*	*3.1590*
5	*K Nearest Neighbor*	*90.0*	*93.33*	*0.9032*	*0.9000*	*1.1072*

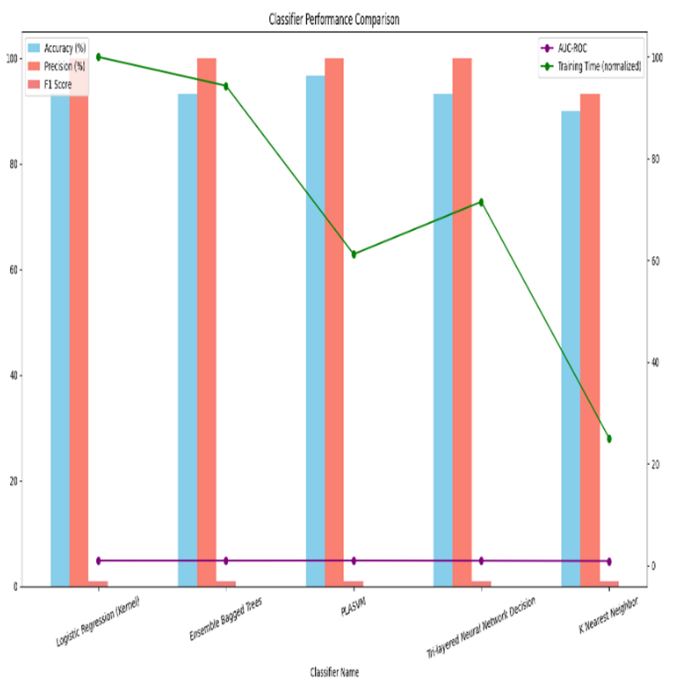


Figure 5. Performance comparison of different classifiers.

The K Nearest Neighbor classifier, while trailing slightly in terms of some metrics like accuracy, precision, F1 score, and AUC, still brings its strengths to the table. Its nimble training time coupled with its robust design could position it as a suitable candidate for scenarios where a straightforward approach and easy interpretation are pivotal.

When it boils down to choosing an appropriate model, it's essential to recognize that decisions should not be solely hinged on performance metrics. Various external factors, such as the model's explainability, dataset characteristics, the nature of the problem at hand, and the repercussions of prediction errors (both false positives and negatives), play a pivotal role. It's crucial to weigh each model's advantages and drawbacks against these overarching factors to ensure an informed selection.

Figure 5 showcases a comparative performance analysis of PLA-SVM with Logistic Regression (Kernel), Ensemble Bagged Trees, KNN and Tri-layered Neural Network Decision classifiers.

V. CONLUSION

This research presents a novel and computationally efficient method to design PLASVMs. It directly solves the SVM design's fundamental optimization problem. To develop the PLASVM model, the suggested method employs separable linear programming concepts and piecewise linear approximation techniques. This approach employs piecewise linear programming and the GUROBI optimizer solver to optimize the primal SVM hyperparameters. The PLA-SVM is successfully validated on a well-known iris data set in a multi-feature binary class and non-separable setup. The generated results are compared to those of previously developed algorithms. According to the findings, it is shown that the PLASVM method outperforms than Logistic Regression (Kernel), Ensemble Bagged Trees, and Tri-layered Neural Network Decision classifiers and KNN classifiers with reference to the chosen performance measures. PLA-SVM also found it quicker than the previous method for the specific iris flower categorization task. Furthermore, the findings may have a broader influence on plant species recognition, pattern recognition, and classification tasks across multiple domains. Future research can explore possibilities for researching new datasets and expanding the recommended approach to encompass a broader spectrum of plant species.

VI. REFERENCES

[1]. Shital Solanki, Dr. Ramesh T. Prajapati "Efficient Primal Support Vector Machine Design Using Piecewise Linear Approximation Based Linear Programming Optimization Techniques" Seybold report, volume 18 (5) 2023, pp 1062-77, doi:10.17605/OSF.IO/2JYMK.

[2]. Dataset Fisher, R.A. 1936. UCI Machine Learning Repository:Iris Data Set. Available at:http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013.

[3]. T. Gupta, P. Arora, R. Rani, G. Jaiswal, P. Bansal and A. Dev, "Classification of Flower Dataset using Machine Learning Models," 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-6, doi: 10.1109/AIST55798.2022.10065178.

[4]. Q. Yu et al., "Iris Flower Recognition Using Four-channels Ring Photonic Reservoir Computing Based on Unidirectional Coupled VCSELs," in IEEE Journal of Quantum Electronics, doi: 10.1109/JQE.2023.3292064.

[5]. Hemalath . D., Y. T. Reddy, L. K and S. C. Davusaheb Saajid, "Iris Flower Species Detection using Machine Learning Technique," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-4, doi: 10.1109/CONIT59222.2023.10205566.

[6]. Poojitha V, Shilpi Jain, "A Collection of IRIS Flower Using Neural Network CLustering tool in MATLAB '', International Journal on Computer Science And Engineering(IJCSE).

[7]. L. Pawar, J. Singh, R. Bajaj, G. Singh and S. Rana, "Optimized Ensembled Machine Learning Model for IRIS Plant Classification," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1442-1446, doi: 10.1109/ICOEI53556.2022.9776724

[8]. Gurobi Optimization Inc., 2022, Gurobi Optimizer Reference Manual, version 10.0. Houston, Texas.

[9]. S. S. Rao, Engineering Optimization: Theory and Practice. John Wiley & Sons, 2019

[10]. The MathWorks Incorp., MATLAB Version 9.13.0.2049777, vol. R2022b. Natick, Massachusetts, 2022.

Cite this article as :

Shital Solanki, Dr. Ramesh Prajapati, "Piecewise Linear Approximation-Driven Primal SVM Approach for Improved Iris Classification Efficiency", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 5, pp.286-292, September-October-2023. Available at doi : https://doi.org/10.32628/CSEIT12390542

Journal URL : https://ijsrcseit.com/CSEIT12390542