SML Project Report

Reducing Dimensionality and Multi-Class Classification

Ayush Srivastava
Computer Science and Artificial Intelligence
Indraprashta Institue of Information Technology Delhi
Delhi, India
ayush21457@iiitd.ac.in

Anurag Gupta
Computer Science and Artificial Intelligence
Indraprashta Institue of Information Technology Delhi
Delhi, India
anurag21451@iiitd.ac.in

***Abstract*—This paper aims to use various machine learning algorithms to perform optimized classification on a data set with low samples and high attributes. We create additional features by clustering the entire dataset to refine the dataset. We perform dimensionality reduction using Principal Discriminant Analysis and Linear Discriminant Analysis and then utilize five different traditional classification techniques for fruit classification, after removing anomalies. After cross-validating each model separately, we analyze the accuracy of each model and use a voting classifier to get final predictions.**

I. Introduction

In many real-life cases, a dataset available for classification has a low number of samples and a high number of attributes, which presents challenges for traditional classification algorithms.

According to the Hughes phenomenon, as the dimensionality increases, the number of data points required for good performance of any machine learning algorithm increases exponentially. Hence, we first decrease the dimensions in the dataset if we only possess a limited number of samples.

Reducing the number of dimensions on a large scale leads to many obscurities in the dataset. We, hence look into clustering to gain additional features before dimensionality reduction.

In order to reduce dimensions, we have two main approaches - feature selection and feature transformation. We use feature transformation since we cannot select which feature is more valuable than the other from the dataset.

We arrive at two significant models for dimensionality reduction - Principal Component Analysis and Linear Discriminant Analysis. We use both algorithms consecutively, which gives us the best results.

For classification on this refined dataset, we adapt five different classification techniques - Gradient Boost Classifier, Random Forest Classifier (Ensemble of Decision Trees), Multi-Class Logistic Regression, Gaussian Naive Bayes Classifier, and K nearest neighbor. We finally ensemble all five of these into a voting classifier and make predictions based on this.

II. Literature Review

A. Logistic Regression

Logistic regression is a versatile statistical method that has a wide range of applications across different fields. Its ability to predict the probability of specific outcomes makes it a powerful tool for decision-making in marketing research, sports, and medical research[1].

B. kNN

K-nearest neighbors (KNN) is a non-parametric machine learning algorithm that is commonly used in various fields, including computer vision, natural language processing and recommender systems. KNN is one of the most frequently used machine learning algorithm in disease prediction[2]

C. Random Forest Classifier

Random forests are an effective tool in prediction. Because of the Law of Large Numbers they do not over fit. They are widely used in computer vision, bio informatics and finance.RF has been successfully applied in many scientific realms such as, the bio informatics, proteomics, and genetics [3][4][5][6][7]

D. Gaussian Naive Bayes Classifier

Gaussian Naive Bayes (GNB) classifier is a simple probabilistic machine learning algorithm. Its assumption of independent and normally distributed features makes it a useful tool for text classification[8], fraud detection, medical diagnosis, and credit scoring.

E. Gradient Boosting Classifier

Gradient Boosting Classifier (GBC) is a powerful machine learning algorithm that has a wide range of applications across different fields. Its ability to combine the results of multiple weak learners makes it a useful tool for improving the accuracy of predictions in various domains, including finance[9], marketing, computer vision, health care, natural language processing, and recommender systems.

F. Principal Component Analysis

Principal Component Analysis (PCA) is a popular dimensionality reduction technique that has a wide range of applications across different fields. Its ability to identify the most important features of a dataset makes it a useful tool for data visualization, signal processing,social sciences and on clinical and biochemical parameters in children with CAH.[10]

G. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a popular dimensionality reduction technique that has a wide range of applications across different fields. Its ability to find the most discriminative features that can separate the different classes makes it a useful tool for pattern recognition, computer vision, bioinformatics, finance, marketing, and NLP.

H. Local Outlier Factor

Local Outlier Factor (LOF) is a popular anomaly detection technique that has a wide range of applications across different fields. Its ability to identify anomalous data points that are not detectable by traditional methods makes it a useful tool for network intrusion detection, credit card fraud detection, medical diagnosis, manufacturing, environmental monitoring, and social network analysis.

I. DBSCAN

DBSCAN is a popular clustering algorithm that has a wide range of applications across different fields. Its ability to identify clusters of data points based on their density makes it a useful tool for image analysis, anomaly detection, recommendation systems, environmental monitoring, social network analysis, and market segmentation.

III. Material

We use a dataset from Kaggle's website that classifies fruits into 20 unique categories. The dataset has 1000 samples and 4096 features.

We use the implementation of kNN, Logistic regression, Gradient Boost classfier, and all the other models of dimensionality reduction and anomaly detection from sklearn library of Python.

IV. Methodology

A. Pre-Processing

We first normalize the dataset since many classification models like Logistic Regression work better on normalized datasets.

B. Splitting the training dataset

We split the training dataset into a new dataset and a validation set to test our model.

C. Clustering for additional features

We consider two different clustering algorithms -DBSCAN and k-means clustering. We get better results on cross-validation with DBSCAN and use it in the final model. The value of eps and minPts is repeatedly determined by cross-validating the final dataset.

D. Dimensionality Reduction

We use both PCA and LDA to reduce our dataset. We set the components of LDA to 19 (since we have 20 classes). For determining the optimum value of n-components in PCA, we loop over a range of values from 50 to 500 and select the best based on the validation set accuracy (described later in detail).

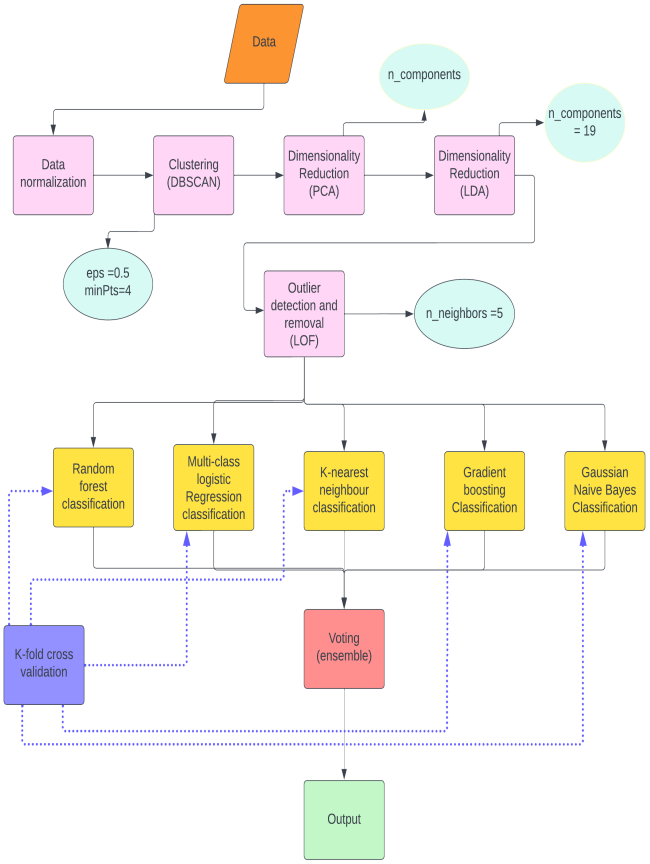

Fig. 1. Flowchart summarizing the method

E. Outlier Detection and Removal

In order to remove highly irrelevant values from the training dataset, we use the Local Outlier Factor algorithm to detect and remove outliers from the training dataset. After trying and testing different values, we set the parameter of the kth neighbor to 5 in the algorithm.

F. Classification Algorithms

We utilize five classification techniques - Gradient Boost Classifier, Random Forest Classifier (Ensemble of Decision Trees), Multi-Class Logistic Regression, Gaussian Naive Bayes Classifier, and K nearest neighbor. For multi-class Logistic Regression, we use the 'one-versus-all' approach.

G. Hyperparamter Tuning

As mentioned in D., we take a range of values for n-components of PCA and loop over them. In each iteration, we perform a grid search cross-validation for each model separately and get the best values for their parameters. For each iteration, we compute the accuracy for the validation set created based on the parameter values obtained and store the n-components value with the highest accuracy.

In this way, we obtain the best value of n-components of PCA for each model and parameter values of the model using grid-search k-fold cross-validation.

We store these tuned hyperparameters values, train our model based on these, and predict upon the test dataset.

H. Ensemble of the Five Models

We already ensemble decision trees in the random forest model and gradient boost classifier used in F. Now, we ensemble our five classifiers based on the Voting Classifier that takes the vote of each model and makes the prediction based on the mode of the prediction of each model.

V. Results and Discussions

A. Hyperparameter Values and Validation Accuracies

We obtain the final values of the parameters for each model as follows:

TABLE I
Model Parameters and Accuracies

Model	n-components of PCA	Validation Accuracy
GNB	250	0.8125
kNN	300	0.8125
Logistic Regression	275	0.8223
Random Forest	250	0.7927
Gradient Boost	200	0.7525

B. Final Accuracy

When we predict the test dataset on our final trained model, we get an accuracy of 0.84615.

VI. Conclusion and Future Scope

We can see that we have successfully our dataset to from 4096 to 19 dimensions and obtained an above 80 accuracy. However, we can include some other improvements also to obtain better accuracies.

We can see that we know the number of optimum clusters since the number of classes are known. Hence, we can try and optimize the k-means clustering method.

We can also implement weighted voting based on the validation accuracies obtained from each model.

Acknowledgment

We would like to take this opportunity to express our gratitude to Professor Koteswar Rao Jerripothula for his invaluable guidance and support throughout our project. His expertise and insights have been instrumental in shaping the direction of our work, and we are truly grateful for his time and effort.

We would also like to thank Indraprastha Institute of Information Technology Delhi for providing us with the resources and facilities necessary to complete this project. The university's commitment to fostering academic excellence has been instrumental in enabling us to pursue our research goals.

References

[1] Schober, Patrick MD, PhD, MMedStat*; Vetter, Thomas R. MD, MPH†. Logistic Regression in Medical Research. Anesthesia Analgesia 132(2),2021,pp 365-366.

[2] Uddin, S., Haque, I., Lu, H. et al. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12, 6256 ,2022.

[3] Leo Breiman. Random Forests. Statistics Department University of California Berkeley, CA 94720,2017

[4] Chen, X., Wang, M., and Zhang, H. . The use of classification trees for bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1, 2011, pp 55–63.

[5] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., et al. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10:213, 2011.

[6] Calle, M. L., Urrea, V., Boulesteix, A. L., and Malats, N.. AUC-RF: a new strategy for genomic profiling with random forest. Hum. Hered. 72, 2011,pp 121–132.

[7] Sarica A, Cerasa A and Quattrone A Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front. Aging Neurosci. 9:329, 2017.

[8] Raschka, Sebastian. Naive Bayes and Text Classification I - Introduction and Theory. 10.13140/2.1.2018.3049,2014.

[9] Yiheng Sun, Tian Lu, Cong Wang, Yuan Li, Huaiyu Fu, Jingran Dong, Yunjie Xu. TransBoost: A Boosting-Tree Kernel Transfer Learning Algorithm for Improving Financial Inclusion,2014.

[10] Ljubicic ML, Madsen A, Juul A, Almstrup K and Johannsen TH. The Application of Principal Component Analysis on Clinical and Biochemical Parameters Exemplified in Children With Congenital Adrenal Hyperplasia. Front. Endocrinol. 12:652888,2014.