Ensemble Methods for Classification

Abhishek Sushil
Infosys Centre for AI
Indraprastha Institute of Information Technology
Delhi, India
abhishek21441@iiitd.ac.in

Alhad Sethi
Department of Mathematics
Indraprastha Institute of Information Technology
Delhi, India
alhad21445@iiitd.ac.in

***Abstract*—High accuracy classification on data sets wherein the number of features far exceeds the number of samples (p<<n) leads to a challenging optimization landscape. Here, we present a detailed explanation of the motivation behind, the construction and execution of an ML Model for classifying different fruits using the scikit-learn library[1] for the given dataset. We use an ensemble method for classification, combining a logistic regression classifier with a random forest approach. Prior to training we pre-process the data, including dimensionality reduction and scaling, to achieve higher downstream accuracy. We justify our approach via rigorous cross-valdiation and experimentation.**

I. Introduction

As part of our coursework for the Statistical Machine Learning (CSE342) course at IIIT Delhi, we have learnt various aspects of machine learning and have applied them to the problem presented. This document is an examination of our methodology and results in creating a model for classifying fruits.

II. Methodology

A. Local Outlier Factor Analysis

We used Local Outlier Factor[2] to determine outliers in the data set. After adjusting *MinPts* in the K-Neighbourhood and rigorously testing (via k-fold cross validation on a rudimentary classification algorithm), we arrived at the conclusion that the data contained a low number of outliers.

B. Principal Component Analysis

Observations of the given data set revealed that the number of raw features provided was far higher than the number of data points, furthermore a majority of these features lacked a significant amount of variance and therefore would not prove to be helpful in classification.
In order to work on better features we used PCA, which identifies directions of maximum variance and projects data on those directions specifically, thus, reducing the dimensional complexity (along with feature extraction and noise reduction).

C. Scaling

We used the inbuilt MinMax Scaling to scale our data. This operation scales features to a range of [0,1] or a user-specified range by subtracting the minimum value of the feature and dividing by the difference between the maximum and minimum value of the feature.

This is useful when we have features with different scales and ranges, and you want to ensure that they are on a comparable scale to avoid dominance by one feature over others. It is a vital pre-processing step for our further operations.

D. Linear Discriminant Analysis

Linear discriminant analysis (LDA)[3] is a supervised learning algorithm used for dimensionality reduction and classification. It aims to find a linear combination of features that maximizes inter-class variance while minimizing the intra-class variance. After performing LDA, the reduced-dimensional data can be used as input for other classifiers we used further along our model. LDA also provides better results when data is scaled as done in prior steps.

Testing also revealed that the PCA-reduced data is vaguely Gaussian since the LDA classification alone provided results that were within reasonable proximity to our final test accuracies.

E. K-Means Features

K-Means is an unsupervised machine learning algorithm that is commonly used for clustering data. The algorithm groups together similar data points based on their distance from each other.

We arrived at the conclusion that high values of k, that is close to actual number of unique data labels would over-fit the model and thus wouldn't help distinguish the different labels (i.e. within the data, the labels aren't easily distinguishable by a direct distance metric). On the other hand, extremely low values of k (<5) would simply provide no significant information to the model and adding this feature would be equivalent to adding another low variance feature to the data.

F. Multinomial Logistic Regression

Logistic regression is a classification algorithm that is commonly used for binary classification problems. However, there are situations where you may have more than two classes and need to use logistic regression for multiple classes.

We found we were obtaining slightly better test accuracies for one vs all method of multi-class regression. For each

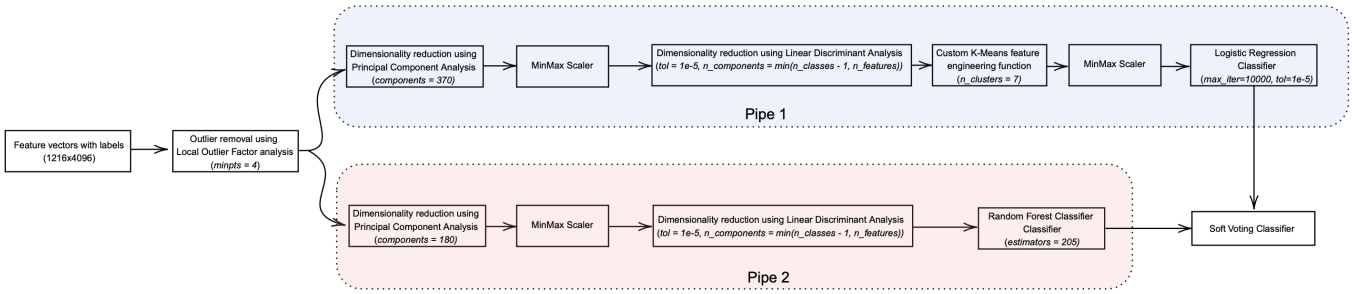

Fig. 1. Pipeline diagram

model, one class is considered the positive class and the other classes are combined into a single negative class. During prediction, each model is used to predict the probability of the positive class, and the class with the highest probability is chosen as the predicted class.

G. Random Forest Classifier[4]

After creating several decision trees, based on different subsets (with replacement) of the data set, the Random Forest Algorithm relies on the mode of the final class obtained to classify the data point.

- Several trees are created on different subsets of data
- Each tree further splits according to a random selection of features at each step
- The most common final label given by each tree is chosen as the answer

The main advantage of random forest classifier is that it can handle large data sets with high dimensionality and can achieve high accuracy with relatively little tuning of hyper-parameters. Additionally, it's less prone to over-fitting than individual decision trees.

H. Ensemble Methods: Soft Voting Classifier

Each of our 2 initial models, which ran as 2 different pipes, used the same pre-processing methods albeit for different hyper-parameters and were given the k-means labels for additional information.

The results of the 2 sub-models fitted with each of these pipes was then finally put through a soft-voting classifier. In Soft Voting, the predicted probabilities of each model for each class are averaged, and the class with the highest probability is chosen as the final prediction. This is in contrast to Hard Voting, where the class predicted by each model is taken as a vote, and the class with the most votes is chosen as the final prediction.

III. Results and Analysis

In addition to the above mentioned components of the classifier, we attempted several other methods, some of which were giving similar validation accuracies. However, we observed that these models were less robust and were prone to large variations in their performance. An interesting experiment to note was the naive-Bayes classifier which gave good validation results, re-affirming our earlier conclusion that the distribution of the training data is vaguely Gaussian.

Next, we would like to note the importance of soft-voting classifier: the relative stability of the logistic regression classifier helps us to balance out the large variances observed in the results of the random forest classifier. This ensemble method lends additional robustness to our approach and helps generalize it better for unseen data sets.

Further, we note that the k-means feature engineering gives us an additional feature for classification, which can help improve validation accuracies. Lastly, we acknowledge the limited size of the dataset and suggest that the same pipeline may be able to achieve higher accuracies given additional training points.

IV. Applications

Here, we explore the some of the various applications of the techniques used in the model. Local outlier factor analysis has been combined with robust classification techniques for use in micro-plastic pollution studies[5] and anomaly-based risk reduction[6] among other numerous applications. Further, we see the application of random forest based ensemble techniques in malware detection,[7] in tumor classification[8] and in neuroscience.[9] K-Means, combined with SVMs, has been used for diabetes diagnosis.[10] Logistic regression models have been used for smart-grid monitoring[11] and for ground-water potential mapping (when combined with soft computing ensemble models).[12] Lastly, we'd like to bring up an interesting application wherein a logistic regression model for diabetes detection was improved using PCA and K-Means techniques.[13]

References

[1] Scikit-learn: Machine Learning in Python,
Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss,

R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay,E.
Journal of Machine Learning Research,12,2825–2830,2011

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). Association for Computing Machinery, New York, NY, USA, 93–104. https://doi.org/10.1145/342009.335388

[3] Chapter 9, *Linear Discriminant Analysis*; Kevin P. Murphy, Probabilistic Machine Learning: An introduction.MIT Press, 2022.

[4] Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[5] Elena M. Höppener, M. (Sadegh) Shahmohammadi, Luke A. Parker, Sieger Henke, Jan Harm Urbanus, Classification of (micro)plastics using cathodoluminescence and machine learning, Talanta, Volume 253, 2023

[6] Pointner, A., Spitzer, EM., Krauss, O., Stöckl, A. (2023). Anomaly-Based Risk Detection Using Digital News Articles. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 542. Springer, Cham.

[7] Vashishtha, L.K., Chatterjee, K., Sahu, S.K., Mohapatra, D.P. (2023). A Random Forest-Based Ensemble Technique for Malware Detection. In: , et al. Information Systems and Management Science. ISMS 2021. Lecture Notes in Networks and Systems, vol 521.

[8] Tao Shi, David Seligson, Arie S Belldegrun, Aarno Palotie, Steve Horvath, Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma, Modern Pathology, Volume 18, Issue 4, 2005.

[9] Paul F. Smith, Siva Ganesh, Ping Liu, A comparison of random forest regression and multiple linear regression for prediction in neuroscience, Journal of Neuroscience Methods, Volume 220, Issue 1, 2013, Pages 85-91.

[10] T. Santhanam, M.S. Padmavathi, Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis, Procedia Computer Science, Volume 47, 2015, Pages 76-83.

[11] Manoharan H, Teekaraman Y, Kirpichnikova I, Kuppusamy R, Nikolovski S, Baghaee HR. Smart Grid Monitoring by Wireless Sensors Using Binary Logistic Regression. Energies. 2020; 13(15):3974.

[12] Nguyen PT, Ha DH, Avand M, Jaafari A, Nguyen HD, Al-Ansari N, Van Phong T, Sharma R, Kumar R, Le HV, Ho LS, Prakash I, Pham BT. Soft Computing Ensemble Models Based on Logistic Regression for Groundwater Potential Mapping. Applied Sciences. 2020; 10(7):2469.

[13] Changsheng Zhu, Christian Uwa Idemudia, Wenfang Feng, Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Informatics in Medicine Unlocked, Volume 17, 2019, 100179, ISSN 2352-9148.