Comparing and Contrasting Machine Learning Tools

Edward Chen, Michael Cirrito,

Jainam Shah, Efthimios Vlahos

11/28/2022

Abstract

Does machine learning have a one size fits all technique? Does data coming in all different modalities make any difference? In this experimental study, several machine learning tools, more specifically classification tools, are investigated. Two gender classification datasets with different data types, i.e., image and numerical, are used to compare the performance of those tools. The efficiency of the features and the misclassification cases are analyzed.

Related Works

[1] presented a comprehensive view on these machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application. The paper explained the principles of different machine learning techniques and their applicability in various real-world application domains. Based on the study, the challenges and potential research directions were also highlighted.

[2].A number of machine learning tools were discussed, as well as how they are applied to different tasks. The paper gave a brief overview of those machine learning tools and their key features. Those tools that can be used to solve real-world problems were also presented. Different parameters and highlights features were examined. Specific frameworks that can be used with the processing platforms were provided as well.

[3].A variety of performance criteria to evaluate the learning methods were used. The authors noted that 1) it is possible that learning methods perform well on one metric but poorly on another, 2) performance by metric has found that calibration has significant impact on tool performance, 3) performance by model has found that there is no universally best learning algorithm. Even the best models perform poorly on some problems, and models that have poor average performance perform well on a few problems or metrics.

Problem Statement

- **Motivated by the aforementioned papers, in this study, we aim to find out which tools [6] out of SVM (Support Vector Machines), Random Forests, and ANN (Artificial Neural Networks) work best for classification on different data modalities. More specifically, the goal is**
- **1) To find the best performance of each classification tool on the dataset**
- **2) To compare the performance of each machine learning tool to classify the same modality and the different modalities**
- **3) To investigate the misclassification samples**
- **4) To analyze the efficiency of those features**

Image Dataset

- The image dataset [4] is of cropped images of around 28,500 male and 28,500 female faces. Som sample images are shown below.

Numerical Dataset

The numerical dataset [5] contains a label and seven descriptive features for each of the 5,001 subjects. The label is the gender of a subject, which is either ”Male” or ”Female”. The features:

- **longhair - 1 if the subject has “long hair”, and 0 otherwise.**
- **foreheadwidthcm - Width of the subject’s forehead in centimeters.**
- **foreheadheightcm - Height of the subject’s forehead in centimeters.**
- **nosewide - 1 if the subject has a “wide nose”, and 0 otherwise.**
- **noselong - 1 if the subject has a “long nose”, and 0 otherwise.**
- **lipsthin - 1 if the subject has “thin lips”, and 0 otherwise.**
- **distancenosetoliplong - 1 if “long distance between nose and lips” and 0 otherwise.**

Our Approach

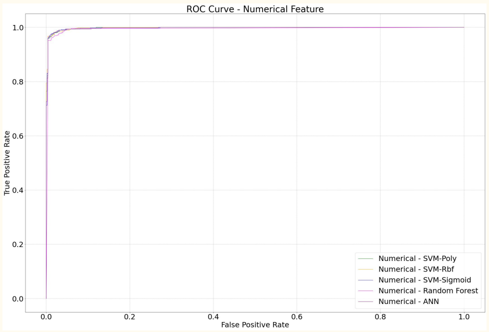
- **We use grid search to find the best parameters for each of those classification tools.**
- **Three performance metrics Accuracy, F1-score [7], and AUC-ROC curve [8] are used to gauge the comparison of the above classification tools.**
- **We focus on the analysis of misclassification and the analysis of the efficiency of those features after we have those test results.**

Numerical Results

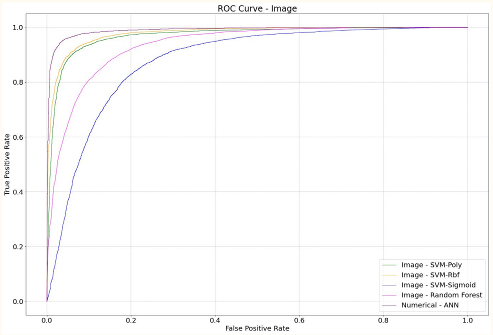
		SVM-Poly	SVM-Rbf	SVM-Sig	ANN	R Forest
Numerical Feature	**Accuracy**	**0.9770**	**0.9780**	**0.9790**	**0.9800**	**0.9740**
	F1- Score	**0.9766**	**0.9777**	**0.9790**	**0.9799**	**0.9736**
	AUC/ROC	**0.9981**	**0.9981**	**0.9977**	**0.9976**	**0.9943**
	Time (Milisec)	**62.13**	**95.40**	**73.48**	**1,770.24**	**749.40**
Image	**Accuracy**	**0.9196**	**0.8778**	**0.8143**	**0.9574**	**0.8646**
	F1- Score	**0.9213**	**0.8800**	**0.8187**	**0.9576**	**0.8684**
	AUC/ROC	**0.9696**	**0.9408**	**0.8788**	**0.9903**	**0.9362**
	Time (Sec)	**3,739.26**	**6,936.98**	**1,028.12**	**1,695.29(*)**	**10,224.12**

* - GPU time. Others are CPU time.

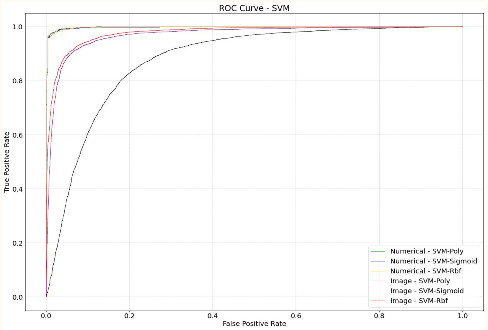
ROC Curves - Numerical Feature



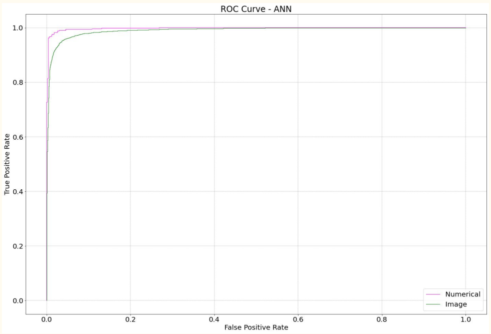
ROC Curves - Images



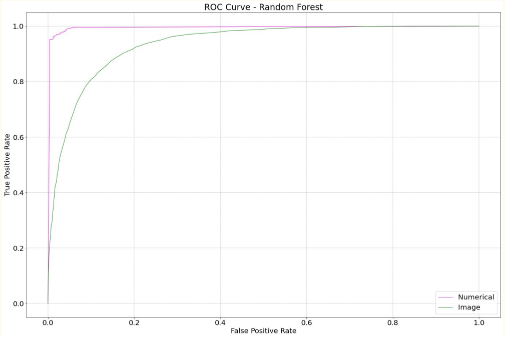
ROC Curves - SVM



ROC Curves - ANN



ROC Curves - Random Forest



Results and Analysis

- For numerical dataset, the performances of all the tools are very comparable. For image dataset, ANN has better performance than all other classification tools. Apparently, ANN takes much longer time than all the other tools. It is also clear that time elapsed on image dataset is way longer than that on numerical dataset.
- It has been found that long hair is very insignificant to determine if a subject is female or not and the four most significant features are wide nose, long nose, thin lips, and long distance nose to lip. When a female is at least three out of those four significant features, the female is misclassified as a male. If a male either has at most one of those four significant features, or at most two of those features plus a forehead width between 12.8cm and 12.9cm, the male is misclassified as a female.
- It has been observed that a lot of images belong to people displaying an emotion. It is also easy to be misclassified if certain features of a face are obstructed. Males with thinner eyebrows seem to be misclassified as female. Females with wider noses seem to be misclassified as male. It is worthy noticing that there are also a few faces on which we as human cannot determine their gender.

Conclusions and Future Work

- Machine learning has been speeding up its advancement in recent years, due to great advancement of computational power, such as GPU. ANNs have become the backbone of machine learning tools nowadays. We have seen in this experimental study that ANNs have outperformed other tools in classification, though with limited data. However, the elapsed time is still a concern of ANNs.
- From this experimental study, it has been found that the tools performs much better on the numerical dataset than on the image dataset, in terms of the performance metrics and also elapsed time. This is because of human interaction with the machine, where the human tells the machine which features may be important and should be looked at. Therefore, when we combine human intelligence and artificial intelligence we get the best results. This could become a future research such that more human intelligence is involved in machine learning.
- Moreover, there is still room for the numerical features to improve. Although there are four significant numerical features, other numerical features are still in the play. Some other relevant features can be added in the future, though they might be less significant. One such example is the measure of the eyebrows.

References

[1]. I. H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science 2.3 (2021): 1-21.

[2]. S. Sarumathi, M. Vaishnavi, S. Geetha, P. Ranjetha, Comparative Analysis of Machine Learning Tools: A Review, International Journal of Computer and Information Engineering 15.6 (2021): 354-363.

[3]. R. Caruana, A. Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[4]. https://www.kaggle.com/datasets/cashutosh/gender-classification-dataset.

[5]. https://www.kaggle.com/datasets/elakiricoder/gender-classification-dataset.

[6]. https://www.wikipedia.org.

[7]. J Korstanje, The F1 score. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

[8]. R Draelos, Measuring Performance: AUC (AUROC). https://glassboxmedicine.com/2019/02/23/measuringperformance-auc-auroc/.