Image recognition with Fashion-MNIST dataset

Pavel Ianko
pavel.ianko@studenti.unipd.it

1. Introduction

This paper considers four machine learning models, applied to a multi-classification problem with fashion MNIST dataset.

The goal is to assess performance of classical machine learning models, ensemble methods and neural network for a multi-classification problem. Studied models are compared based on four metrics, with evaluation on the unseen validation set. In addition, the study is dedicated to models behaviour analysis (confusion matrices, mistaken images and learning curves).

Among the studied models (random forest, SVM, decision tree, neural network), best validation and test performance corresponds to SVM, with all metrics evaluated at 85% (precision, recall, accuracy, f1). Confusion matrices clearly explain the mistakes nature (e.g. mostly mistaken classes were T-shirt and Shirt, with over 14% misclassifications on test set).

2. Dataset

Fashion-MNIST dataset includes 60,000 train and 10,000 test data instances. Each instance is a 28×28 pixels flattened grey-scale image, representing one of ten classes — t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot.

Under conditions of high time requirements, the size of train dataset was reduced, while the test set remained untouched. Final data split represents 5,500 images for the train set, 1,100 and 10,000 images for validation and test sets respectively.

Before training, it was assured that the class distribution was maintained uniform across train, validation and test sets (Fig. 1).

2.1. Filtering and preprocessing

None of the data instances contain NAN values, hence no NAN-processing was applied. Each of the 10 classes is present in train, validation and test sets, complying with uniform distribution (around 10% of images for each type of clothes). A distribution of pixels' grey-scale value shows necessity for applying data scaling (Fig.2).

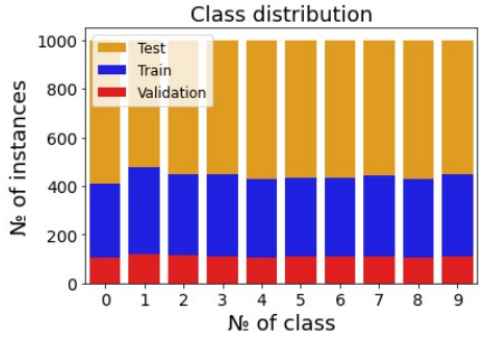

Figure 1. Class distribution in train, validation and test sets

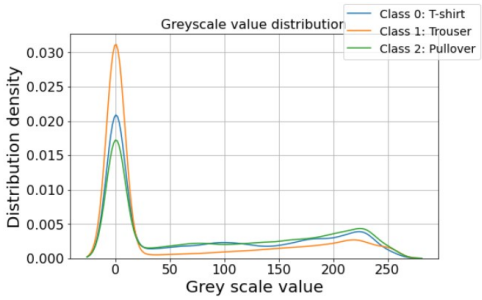

Figure 2. Pixels greyscale distribution for classes: T-shirt, trouser and pullover

In summary, a preprocessing pipeline encapsulates highlighting edges with Sobel filtering and scaling (Fig.3). Because of non-normal bimodal grey-scale distribution (Fig. 2), using min-max scaling instead of standardization raised final metrics by 5%.

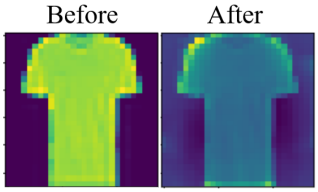

Figure 3. Example of image before and after preprocessing (T-shirt) with Sobel filtering and standard scaling

3. Method

In the study, four models are compared – decision tree, support vector machine, random forest and feed forward neural network. Every presented algorithm covers certain area in the hypothesis space. We compare performance of single model against ensemble method (decision tree versus random forest), linear model and deep learning approach, according to four multi-classification metrics.

Choosing feed-forward neural network (FFNN) is a common approach for image classification. However, neural networks are often addressed as "black-box", as mistakes interpretation for these models is challenging. Hence, an SVM model was chosen, because, with a proper kernel choice, SVM performs better than FFNN [2].

In addition, with a proper choice of random forest structure, this model performs comparably with SVM model, yet reducing training and testing costs [1]. To compare a performance of an ensemble model, a single decision tree was also studied.

4. Results and discussion

4.1. Training method and metrics

Models performance is compared based on accuracy, precision, recall, and F1 with macro averaging. F1 metrics is used to select the best model, as it merges precision and recall. FFNN model was trained with early-stopping mechanism, 16 images per mini-batch and 20% of data used for validation.

Rest of the models were trained using cross-validation with 5 folds and hyper-parameters tuning. Final model was selected, based on unseen validation set, after which retrained on train and validation data. In conclusion, all metrics of the best model are reported on the test set of 10,000 unseen images.

4.2. Cross-validation performance

After cross-validation training, Fig. 4 indicates SVM prevalence over decision tree and random forest with F1 metrics over 0.8.

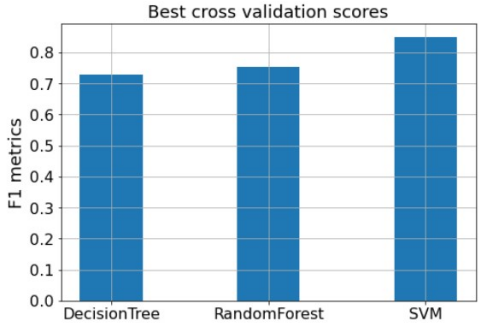

Figure 4. Models best result during cross validation training

Moreover, across all cross-validation folds SVM yields consistent results, as shown on Fig. 5

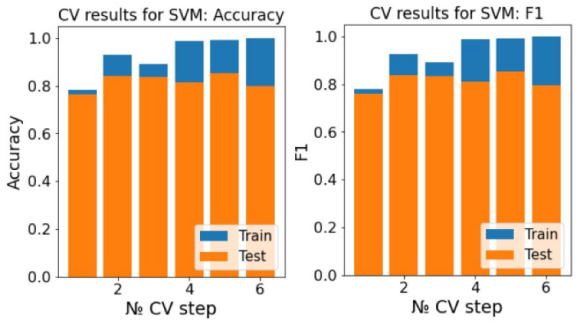

Figure 5. SVM accuracy and F1 metrics on cross-validation folds

4.3. Learning curve analysis

A subset of 2,000 data instances was used to analyse how validation metrics evolves with the train size. Figure 6 reveals an important detail about models behavior.

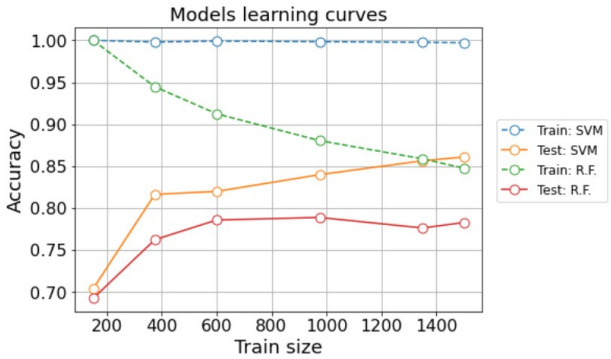

Figure 6. Learning curves comparison (R.F. stands for random forest)

While random forest demonstrates convergence, SVM shows overfitting. Despite, SVM model performs best on 10,000 unseen images. A possible explanation for overfiting is using a subset of 2,000 images. Using more data prevents overfitting, as it is proved by test performance.

On figure 7 for an FFNN learning curve, the smoothness with a low train-validation discrepancy indicate proper choice of learning rate. A possible model improvement is increasing patience of early-stopping mechanism.

4.4. Confusion matrices analysis

Each of the studied model exhibits similar mistakes (Fig.8). One of problematic classes is a shirt, with 60% correct SVM predictions, against lowest 16% for random forest model. Around 10% of mistakes correspond to shoes types (sneaker, ankle boot and sandals).

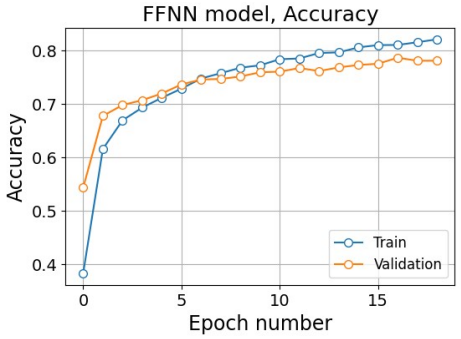

Figure 7. Learning curves for feed forward neural network

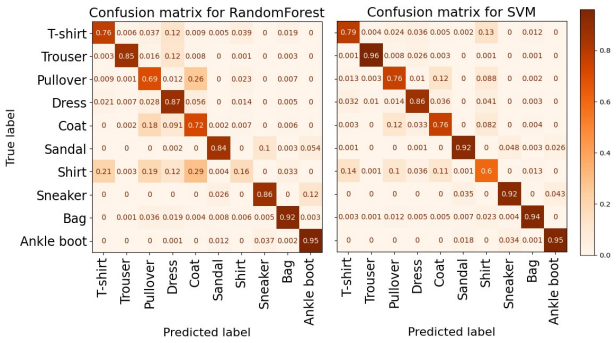

Figure 8. Confusion matrices for random forest and SVM model

Image pairs below visualize similar, mistaken classes (Fig. 9). Hence, there is an evidence for using more data on misclassified pairs. Also, using extra data about fabrication materials (e.g. cotton, leather) will increase discrimination accuracy between shirt and coat, dress and pullover, which are frequently mistaken class pairs.

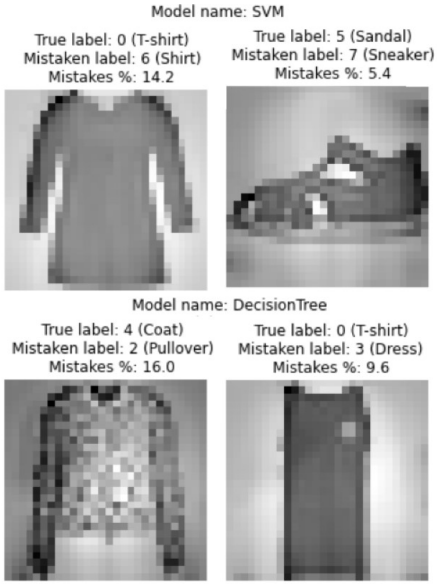

Figure 9. Mostly mistaken classes for SVM and decision tree models

4.5. Final metrics report

Figure 10 summarizes models performance on the unseen validation set of 1,100 images. Thus, SVM outperforms rest of the models with metrics above 80%. Figure 11 reports test metrics, evaluated at 84.9%.

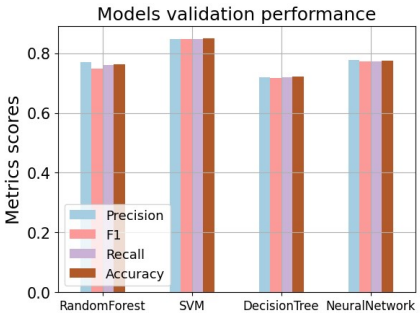

Figure 10. Models performance on the unseen validation set of 1,100 images. Studied metrics are precision, recall, accuracy and F1

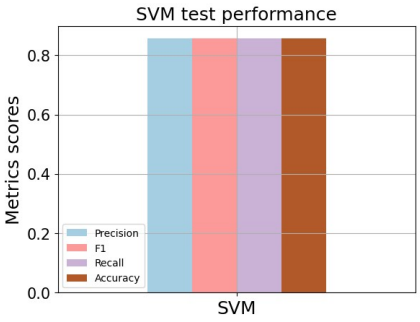

Figure 11. SVM performance on the unseen test set of 10,000 images (precision, recall, accuracy and F1) is estimated at 84.9%

5. Conclusion

Four models – SVM, decision tree, random forest and FFNN network – were applied to a multi-classification task. Each model was trained on 5,500 data instances, with cross-validation approach and hyperparameter tuning. Best model – SVM – performed with 80% accuracy on the unseen 1,100 images. Final SVM metrics (accuracy, precision, recall, F1) were evaluated on 10,000 test images, with uniform scores of around 85% each.

References

[1] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. Ieee, 2007.

[2] Le Hoang Thai, Tran Son Hai, and Nguyen Thanh Thuy. Image classification using support vector machine and artificial neural network. *International Journal of Information Technology and Computer Science*, 4(5):32–38, 2012.