Scribe Classification on UCI Avila Dataset using Naïve Bayes, KNN, SVM, Random Forest

Rajat Verma

Department of Computer Science
Stevens Institute of Technology
Hoboken, New Jersey, 07307, USA

rverma2@stevens.edu

I. Introduction

In the field of manuscript studies (paleography and codicology), a particularly interesting case is the study of highly standardized handwriting and book typologies. In such cases, the analysis of some basic layout features, mainly related to the organization of the page and to the exploitation of the available space, may be very helpful for distinguishing similar scribal hands. In this framework the system considers a set of features typically used by palaeographers, which are directly derived from the analysis of the page layout. I tried for identifying the scribes who collaborated to the transcription of a single medieval Latin book. Using Supervised machine learning algorithm on Avila Dataset for model training and to observe the prediction result based on different parameters of Algorithm.

The paleographic analysis of the manuscript has individuated the presence of 12 copyists. The pages written by each copyist are not equally numerous. Each pattern contains 10 features and corresponds to a group of 4 consecutive rows. The prediction task consists in associating each pattern to one of the 12 copyists (labeled as: A, B, C, D, E, F, G, H, I, W, X, Y). There is a total of 20860 samples. The pages written by each copyist are not equally numerous and there are cases in which parts of the same page are written by different copyists.

In this project, focused on learning a classifier which is accurately able to classify handwritten alphabets of the manuscript. The project compare the performance of commonly used machine learning algorithms specifically K-Nearest Neighbours, Naïve Bayes, Support Vector machines, Random forest classifiers on the Avila data set which has been extracted from 800 images of the "Avila Bible", a giant Latin copy of the whole Bible produced during the XII century between Italy and Spain. Further, continuing to understanding the parameters and feature of dataset and models. The prediction result can be improved by using more rigorous and advanced models like ANN, CNN or Deep learning to further exploit the information about the classification reliability.

II. Approach

To approach this problem, I tried to use four different model to learn about the accuracy and performance of classification, every model is tuned to understand about the performance metrics. A series of operation is performed before feeding the data to the model.

In the pre-processing step noisy pixels, such as those corresponding to stains or holes onto the page or those included in the frame of the image, were already detected and removed which would have otherwise aided to the faulty input data and therefore, incorrect training and prediction. Data transformation and scaling is done using standard library tools to help visualize the data and to speed up the algorithm from denser dimension to a lower dimension using PCA. Since, PCA is affected by scale so the features space of the data is standardized before applying PCA. Dataset's features onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance. The next step, performs the recognition task, which has the effect of identifying the rows in each page written by the same copyist. In our study, we have assumed that the manuscript has been produced by N different copyists, previously identified through the traditional palaeographical analysis. Based on assumption that each single pattern to be classified is formed by a group of M consecutive rows, described by using the previously defined features.

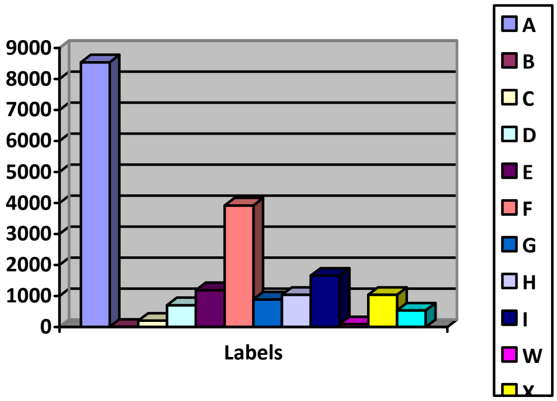
TABLE I
Feature Overview

1	intercolumnar distance	6	modular ratio
2	upper margin	7	interlinear spacing
3	lower margin	8	Weight
4	Exploitation	9	peak number
5	row number	10	Modular ratio /interlinear spacing

III. Experimental Design

Avila Data Set: https://archive.ics.uci.edu/ml/datasets/Avila

The Avila data set has been extracted from 800 images of the "Avila Bible". The pages written by each copyist are not equally numerous. Each pattern contains 10 features and corresponds to a group of 4 consecutive rows. The prediction task consists in associating each pattern to one of the 12 copyists. The data have been normalized, containing the 20874 samples.



Distribution of Instance

Tools and libraries used in implementation are- NumPy, Pandas, Sklearn, Matplot, jupyter Notebook and Python. The training and testing data are obtained by randomly shuffling the dataset and splitting it into two part of which 10-50% of the data is used for training whereas other for testing.

K Nearest Neighbour is a nonparametric approach that is based on feature similarity techniques that consider the out of samples and training sets for comparison to make the prediction by examining the majority vote from K nearest neighbour selected on the basis of similarity feature. A positive integer k is specified along with new sample. We find the K - nearest samples to the new sample which by considering the distance. We do majority polls among the K nearest samples to select the class of new sample. This gives us the class of the new sample. The K component is of the prime importance in the algorithm. When k was fairly small the model was sensitive to the noise. On the other hand, if k taken fairly large the performance decreases because more samples are involved in polling. Therefore, in order to best approximate the k_component different values were tried and the mean accuracy was recorded before finalising the value.

Naïve Bayes multi-class labelling is trivial. It's a simple classifier that applies Bayes' rule with strong (naive) independence assumptions also known as the "independent feature model". Model performs reasonably well despite simplicity in many cases. For each test example ii, and each class kk you want to find:

$$\arg\max_{\theta} \Sigma_i \log P(x_i|\theta)$$

In other words, you compute the probability of each class label in the usual way, then pick the class with the largest probability.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

However, if categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. Therefore, it was ensured that training data has enough variation. Best approximate the accuracy using naïve Bayes was recorded on randomly training the model for definite iterations and the mean accuracy was recorded before finalising the value.

SVM multiclass classification is very effective in high dimensional spaces. Two terms regularization parameter and gamma, also one major parameter called as kernel, which defines the kind of separation between classes. The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with view to making it linear dataset. Usually linear and polynomial kernels are less time consuming and provides less accuracy than the rbf or Gaussian kernels. So, RBF kernel is used with gamma being tuned multiple times to obtain a suitable curve. Moreover, for large values of C (Regularization), the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

$$K(X_1, X_2) = exponent(-\gamma \|X_1 - X_2\|^2)$$

||X1 — X2 || = Euclidean distance between X1 & X2

Random Forest is a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between models is the key. Decisions trees are very sensitive to the data they are trained on, small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees known as bagging.

IV. Experimental Result

As anticipated in the Introduction, after testing our system on a large dataset of digital images obtained from a giant Latin copy of the whole Bible, called “Avila Bible” normalized, by using the Z-normalization method, and divided in two subsets: the first one has been used as training set for the various models, while the second one, distributed from range of .1 to .5 time of size, has been used for testing the system. For each class, the samples have been randomly extracted from the database, following are observed-

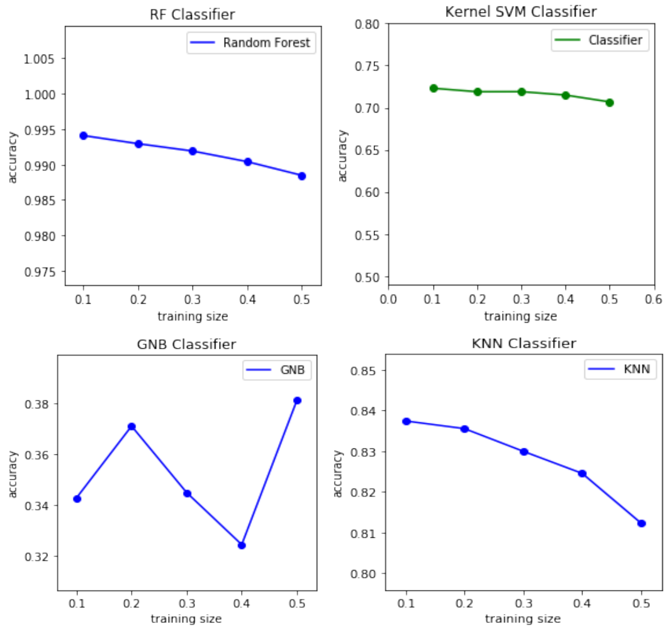
Classifier Name	Mean Accuracy on training	Mean Accuracy on testing	Mean Train Accuracy std Dev	Mean Test Accuracy std Dev
KNN	82.79	73.313	00.102	01.649
Naïve Bayes	35.285	34.688	02.305	1.968
Random Forest	99.156	95.847	00.220	0.889
SVM	71.643	70.727	00.615	01.083

Random forest performs the best out of given classifiers with the mean accuracy of about 96.6% on test samples. However, Random Forest did not perform well when features are monotonic transformation of other features using PCA because the trees of the forest were less independent from each other which probably overfit and had a poor performance of 65% on testing from 95% on training with default estimator set to 10.

KNN performs good with neighbours set to optimal odd number to avoid extravagant computation of irrelevant polling and provides and accuracy of around 77% on test data at one point.

Naïve Bayes, since it’s a simple classifier that applies Bayes' rule with strong (naive) independence assumptions did not performed good even when tested on low train size as it usually does on low train size and produces an accuracy of just 33%.

SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible. RBF kernel is used with gamma being tuned multiple times to obtain a suitable curve. One interesting fact that worth observing about this model is the performance which is almost same on training and testing data. The four graph shows the relationship of train accuracy on train size for all 4 models-



Conclusion

The task as stated in the problem introduction has been accomplished using various models to gain insight into statistical models of machine learning. The experimental investigation has regarded some main aspects. The one was intended to test the effectiveness in discriminating different scribes and understanding the potential and discriminant power of each model using principle component analysis. The experimental results clearly state the effectiveness of the proposed approach and different models in test. Future work will include exploiting the information about the classification reliability. Information would allow palaeographers to find further confirmation of their hypothesis and to concentrate their attention on those sections of the manuscript which have not been reliably classified.

References

[1] Maniaci, M., Ornato, G.: Prime considerazioni sulla genesi e la storia della bibbia di avila. Miscellanea F. Magistrale, Spoleto, in press (2010)

[2] Quinlan, J.R.: C4.5: Programs for Machine Learning Morgan Kaufmann (1993)

[3] Stokes, P.: Computer-aided palaeography, present and future. In: Rehbein, M., Sahle, P., Schaan, T. (eds.) Kodikologie und Palographie im digitalen Zeitalter / Palaeography in the Digital Age. pp. 309–338 (2009)

[4] Claudio De Stefano, destefano '@' unicas.it, University of Cassino and Southern Lazio (ITALY)

[5] Francesco Fontanella, fontanella '@' unicas.it, University of Cassino and Southern Lazio (ITALY)

[6] Marilena Maniaci, m.maniaci '@' unicas.it, University of Cassino and Southern Lazio (ITALY)

[7] Alessandra Scotto di Freca, a.scotto '@' unicas.it, University of Cassino and Southern Lazio (ITALY)