Textual-Content-Based Classification of Bundles of Untranscribed Manuscript Images

Jose Ramón Prieto, Vicente Bosch, Enrique Vidal, Carlos Alonso, M. Carmen Orcero, Lourdes Marquez
joprfon@prhlt.upv.es, carlos.alonso.v@juntadeandalucia.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

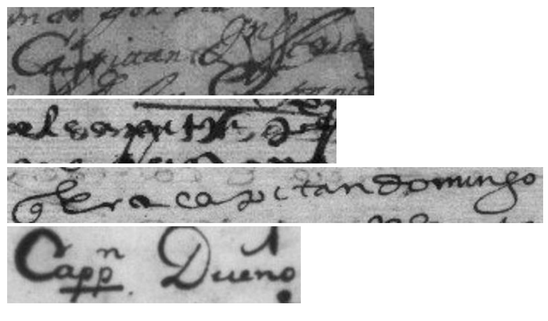
PRHLT

JUNTA DE ANDALUCIA

Instituto Andaluz del Patrimonio Histórico
CONSEJERÍA DE CULTURA Y PATRIMONIO HISTÓRICO

Motivation

- Classify bundles of handwritten images
- Not rely on transcriptions because of the difficulty of the images
- There are different contrasts, writing styles, bleed-throug, use of abridged and tangled abbreviations
- Case of study on AGI - Carabela corpus



Estimating Word and Document Frequencies from PrIx

- The Bag of Words model is assumed
- We use Information Gain to select most relevant words
  - $f(t_v)$: the number of documents in D which contain v
  - $f(c, t_v)$: the number of documents of class c which contain v
- Tf·Idf is used which is based on the following frequencies
  - $f(t_v)$ as in Information Gain
  - $f(D)$: the total number of words in D
  - $f(v, D)$: the number of times v appears in D

Relevance probability (RP), $P(R|X, v)$ of each image X for each pseudo-word v
$P(R \mid X, v) = \sum_{i,j} P(R, i, j \mid X, v) \approx \max_{i,j} P(v \mid X, i, j) \approx \max_{b \in X} P(v \mid X, b)$

The frequencies are estimated from image Relevance Probabilities as follows:

$$f(D) \equiv n(X)$$

$$f(t_v) \equiv m(v, \mathcal{X})$$

$$f(c, t_v) \equiv m(v, \mathcal{X}_c)$$

$$f(v, D) \equiv n(v, X)$$

$$E[n(X)] = \sum_{x \in X} \sum_{v} P(R \mid x, v)$$

$$E[n(v, X)] = \sum_{x \in X} P(R \mid x, v)$$

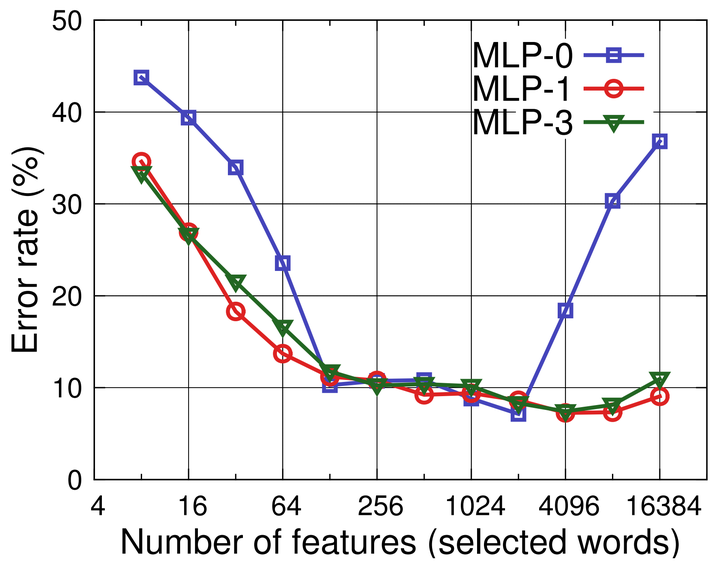
$$E[m(v, \mathcal{X})] = \sum_{X \in \mathcal{X}} \max_{x \in X} P(R \mid x, v)$$

Document Image Classification

Optimal prediction of the class of an image document X is achieved by the maximum class posterior.

- Document Image Representation: Bag of Words model based on words and documents frequencies esimated from PrIx
- Multinomial Naive Bayes: A linear classifier equivalent to a (plain) perceptron
- Multilayer Perceptrons trained with cross entropy loss
  - MLP-0: 0-hidden-layers MLP
  - MLP-1: a proper MLP including one hidden layer with 64 ReLU neurons and batch normalization
  - MLP-3: 3 hidden layers including 16, 32 and 64 ReLU neurons and batch normalization

Results



Best classification error rate (7.1%) obtained by the plain perceptron (MLP-0), for a relatively large vocabulary of the 2048 words with largest Information Gain

Conclusions

- Presented an approach that is able to perform textual-content-based document classification directly on documents of untranscribed handwritten text images
- Overcame the need to explicitly transcribe manuscripts, which is generally unfeasible for large collections

Future works

- Better representation using geometric information of pseudo-words
- Improve the term selection method to get accurate results with smaller vocabularies