The Class Imbalance Problem in Author Identification

Efstathios Stamatatos

University of the Aegean

Talk Layout

- Introduction
- Instance-based vs. profile-based author identification
- The CNG approach
- New dissimilarity functions
- Experiments
- Concluding remarks

Introduction

- Authorship identification can be seen as a single-label multi-class text categorization task

- Applications:
  - Literary research (attribution of historical texts of unknown or disputed authorship to known authors)
  - Intelligence (attribution of messages or proclamations to known terrorists)
  - Criminal law (identifying writers of harassing letters)
  - Computer forensics (identifying the authors of source code of viruses)
  - …

Text Representation

- Vocabulary richness
- Most frequent words
- Syntax-based features
- Character *n*-grams
- …
- Parameter-free approaches
  - Compression-based models

The Class Imbalance Problem

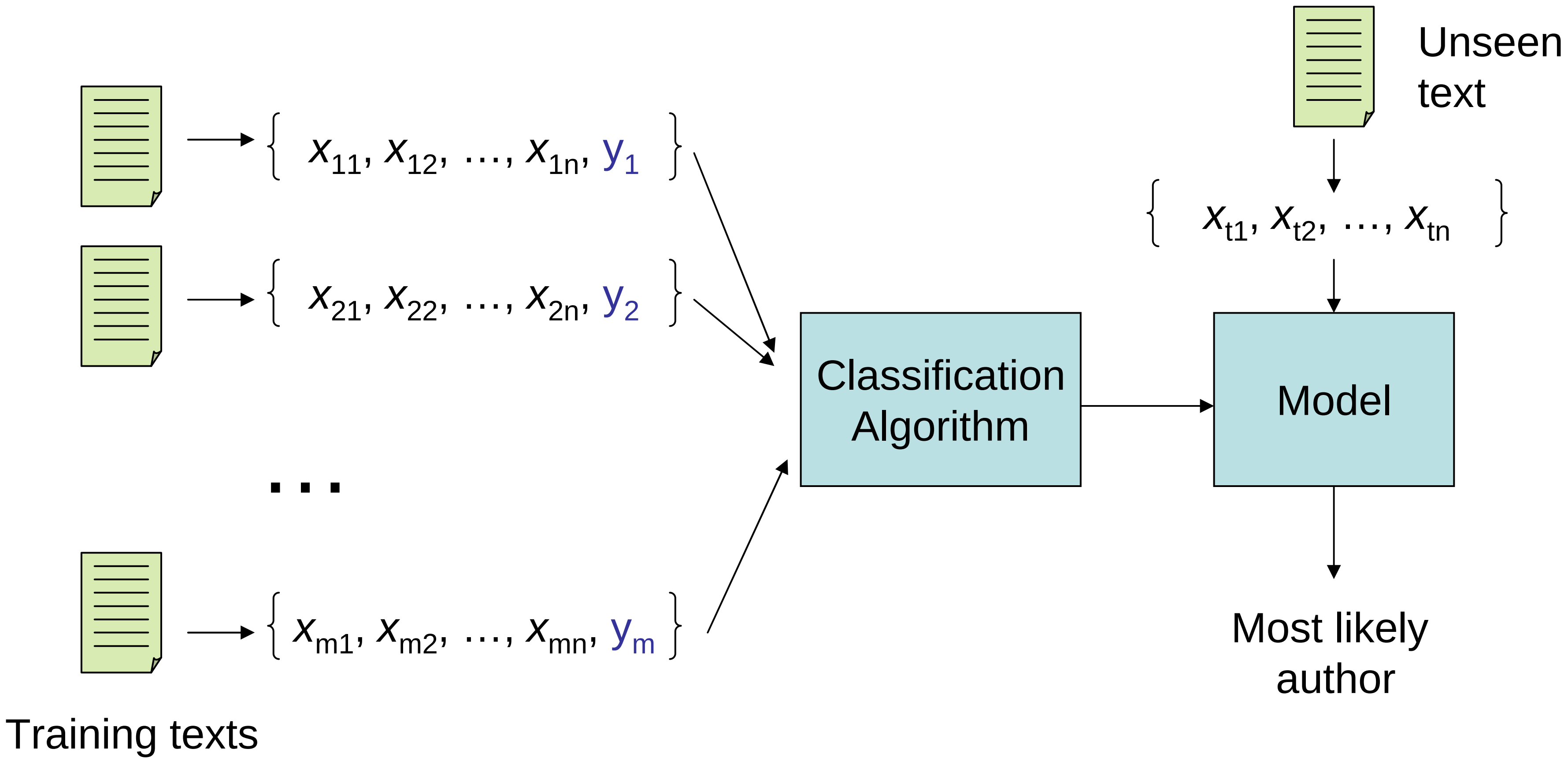
- Very often, there are extremely few training texts at least for some of the candidate authors
- Alternatively, there may be a significant variation in the text-length among the available training texts of the candidate authors
- In forensic tasks usually there is no similarity between the distribution of training and test texts over the authors
  - A basic assumption of inductive learning does not apply

Author Identification Methods

- Instance-based approaches
  - Each text of known authorship provides a training instance
    [Stamatatos, 2000; Diederich, 2003]
- Profile-based approaches
  - All the available texts of known authorship per author are concatenated
  - A profile is extracted
    [Keselj, 2003; van Halteren, 2004]

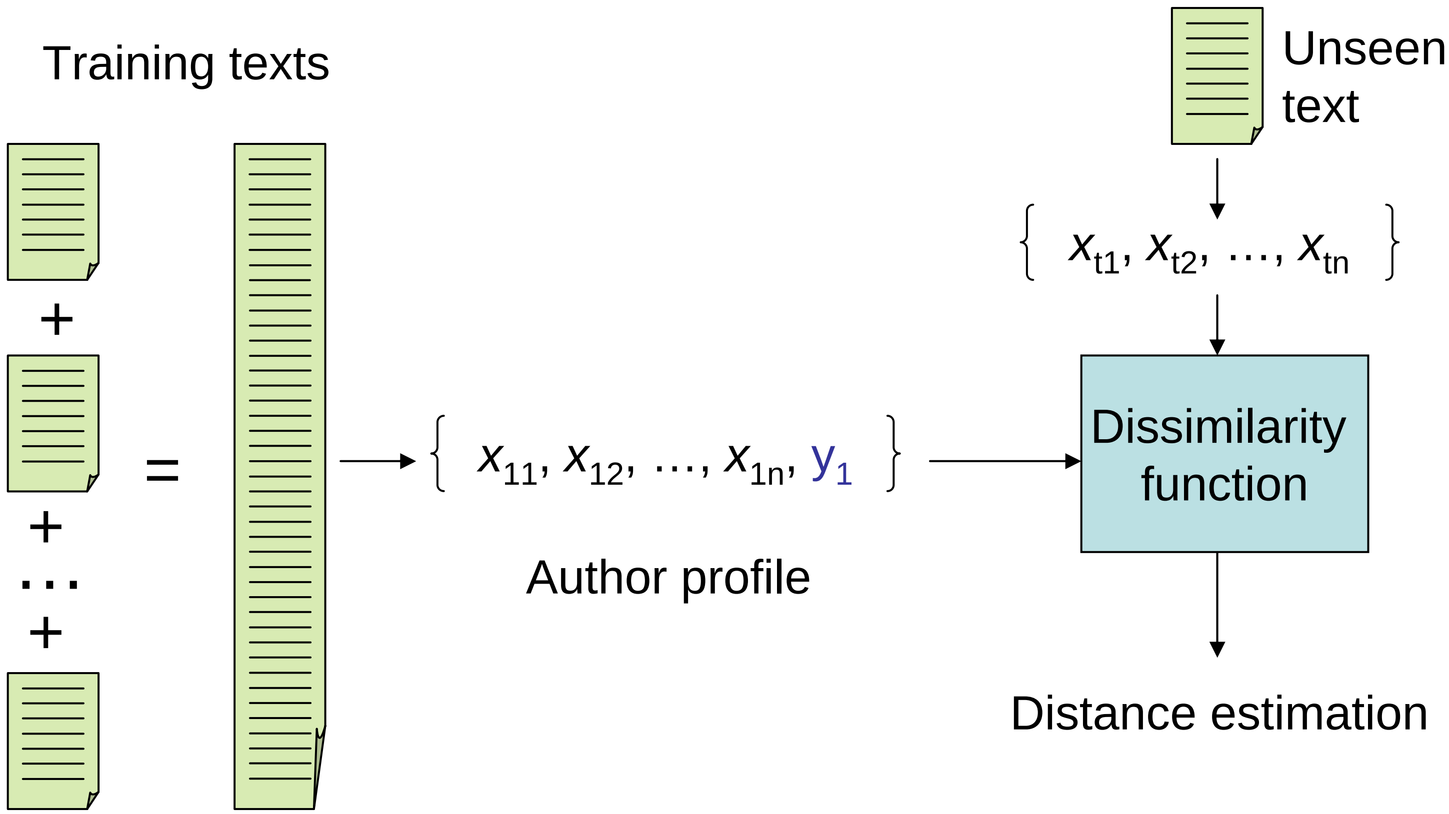
Instance-based Approach

- Given *m* texts of known authorship
- Each text is represented by *n* features



Profile-based Approach

- Given *k* texts of known authorship for a certain author
- *n* features are used to represent the style



Instance-based vs. Profile-based Author Identification

- Instance-based approaches
  - Powerful algorithms (e.g., SVM) can be used
  - Document-level features (e.g., greetings, signatures) can be included
  - Class imbalance depends on the *amount* of training texts per author
- Profile-based approaches
  - Naturally models similarities and differences between authors
  - The extracted profile can sketch out the properties of the author's style
  - Class imbalance depends on *text-length* of training texts per author

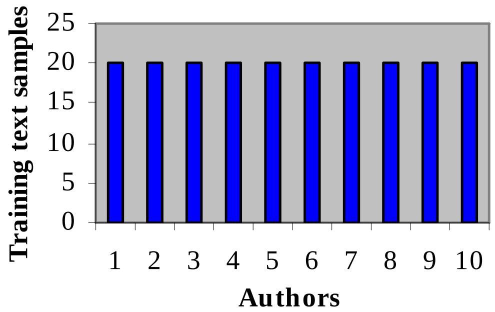
Solutions for Class Imbalance in Instance-based Approaches

[Stamatatos, 2006; Stamatatos, 2007]

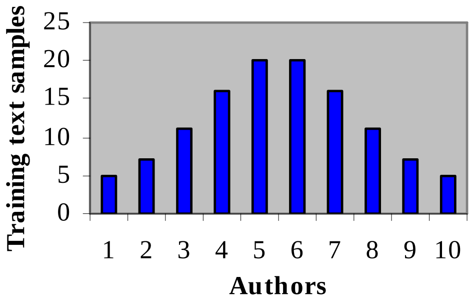
- Textual data can be handled flexibly so that to produce a variable amount of text samples of variable length
- Efficient segmentation of the training texts into sub-samples according to the size of the class
  - Many short samples for the minority classes
  - Less but longer samples for the majority classes
- Text re-sampling can easily provide new synthetic data

Solutions for Class Imbalance in Instance-based Approaches

Produced distribution

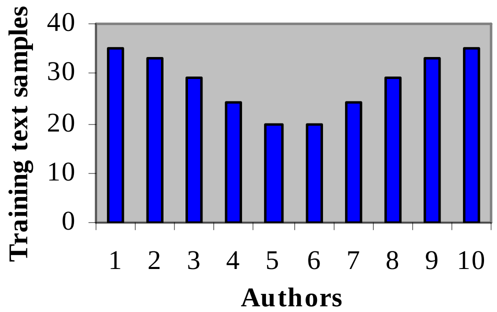


Initial distribution



- Re-balancing the dataset by variable length samples

- Re-balancing the dataset by text re-sampling



The CNG Approach

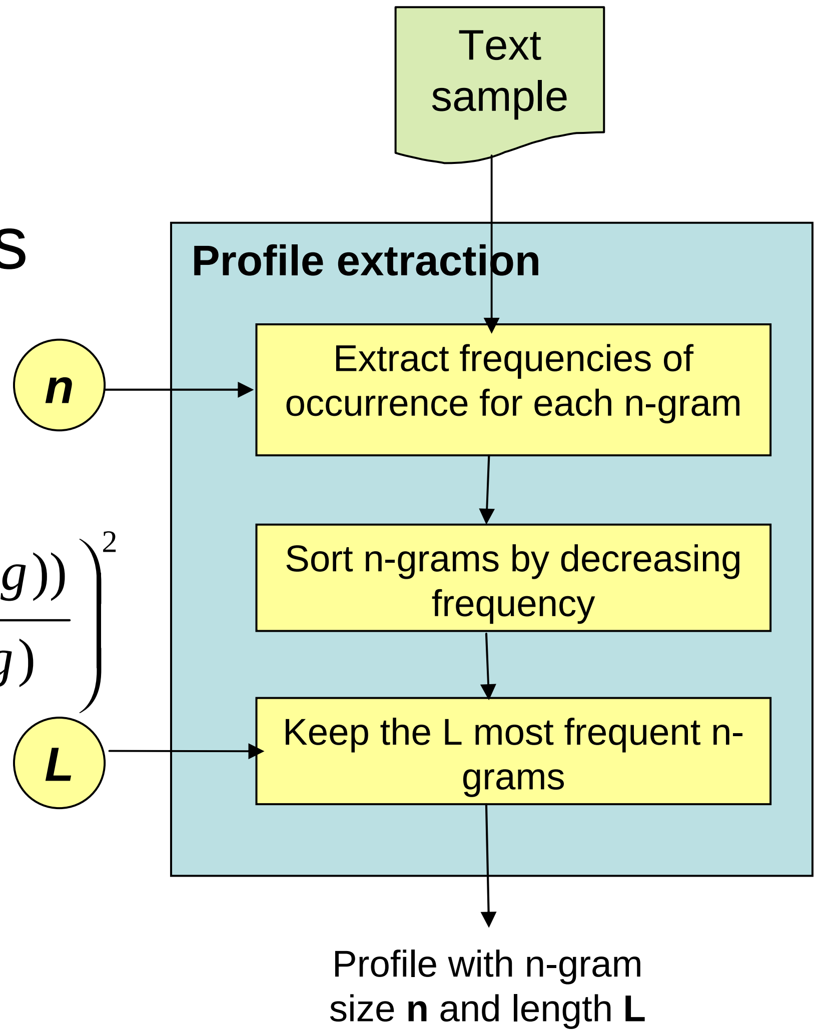
[Keselj et al., 2003]

- Profile-based
- Character *n*-gram features
- Case-sensitive
- Dissimilarity function:

$$d_0(P(x),P(T_a)) = \sum_{g\in P(x)\cup P(T_a)} \left(\frac{2(f_x(g)-f_{T_a}(g))}{f_x(g)+f_{T_a}(g)} \right)^2$$

- Classification:

$$author(x) = \arg\min_{a\in \mathbf{A}} d_0(P(x),P(T_a))$$



The CNG Approach

- Pros
  - Language-independent
  - Simple and fast
  - Able to deal with imbalanced data
  - Excellent performance [Juola, 2004]
- Cons
  - Parameters *L* and *n* have to be tuned
  - A predefined *L* may not be applicable
  - If an author profile is shorter than *L* it becomes unstable
    - It happens under class imbalance conditions

Instability of d_0

$$d_0(P(x), P(T_a)) = \sum_{g \in P(x) \cup P(T_a)} \left(\frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)} \right)^2$$

- d_0 favours authors with less training texts when L is higher than the profile length of the author
- A realistic scenario in author identification tasks
- [Frantzeskou et al., 2006] propose an alternative metric, Simplified Profile Intersection (SPI):

$$d_{spi}(SP_x, SP_{T_a}) = \left| SP_x \cap SP_{T_a} \right|$$

  - The frequency of occurence of n-grams is not taken into account
  - Similarity function (while d_0 is dissimilarity function)
  - Good results in source code author identification experiments

New Dissimilarity Functions: d_1

- d_0 is a symmetrical function:

$$d_0(P(x), P(T_a)) = \sum_{g \in P(x) \cup P(T_a)} \left(\frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)} \right)^2$$

- d_1 is not symmetrical
- It ensures that the distance of the test profile from the author profile will be calculated based on the same amount of terms
- It is not affected by short profiles (shorter than L)

$$d_1(P(x), P(T_a)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)} \right)^2$$

New Dissimilarity Functions: d_2

- d_2 is an extension of d_1
- It takes into account the *corpus norm*
  - Concatenation of all available files from all the authors
- The more an *n*-gram deviates from its 'normal' frequency, the more contributes to the model

$$d_2(P(x), P(T_a), P(N)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_{T_a}(g))}{f_x(g) + f_{T_a}(g)} \right)^2 \cdot \left(\frac{2(f_x(g) - f_N(g))}{f_x(g) + f_N(g)} \right)^2$$

Experiments

- Corpus:
  - Texts taken from RCV1
  - 50 authors with texts on the topic CCAT
- Model:
  - $n = 3$
  - $L = 1{,}000 – 10{,}000$
- Distances:
  - d_0
  - d_1
  - d_2
  - SPI

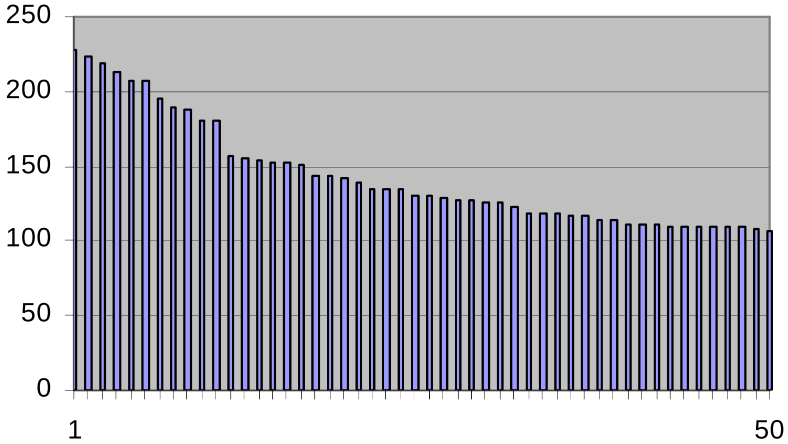
Experiments: Corpus

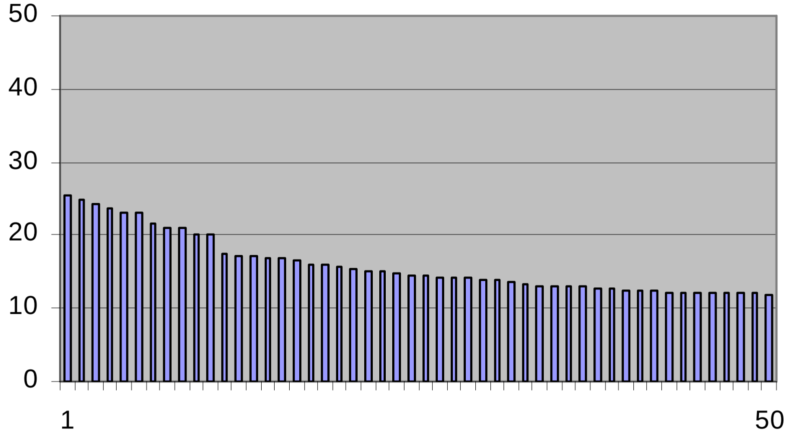
	C50ir	C50ig	C50b50	C50b10
Training corpus	Imbalanced	Imbalanced	Balanced	Balanced
Test corpus	Imbalanced	Balanced	Balanced	Balanced
Training corpus (text samples)	7,962	1,234	2,500	500
Test corpus (text samples)	883	2,500	2,500	2,500
Longest training text (KB)	812	170	179	43
Shortest training text (KB)	288	6	100	18
Longest training profile (3-grams)	11,817	7,326	7,955	4,504
Shortest training profile (3-grams)	8,244	1,807	5,956	2,890

Distribution of C50ir and C50ig

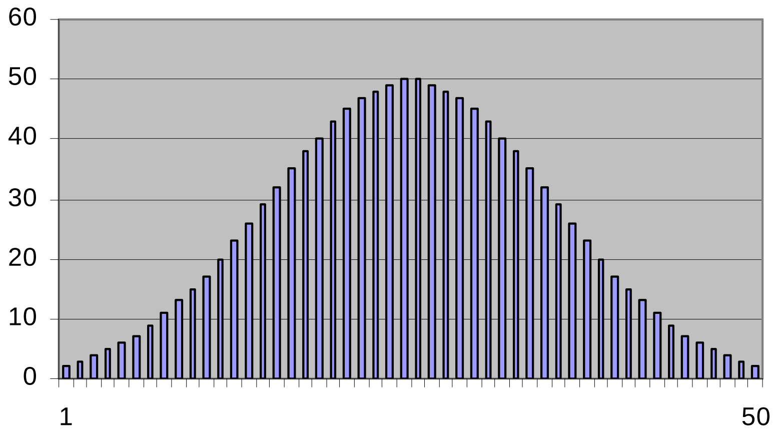
Train | Test

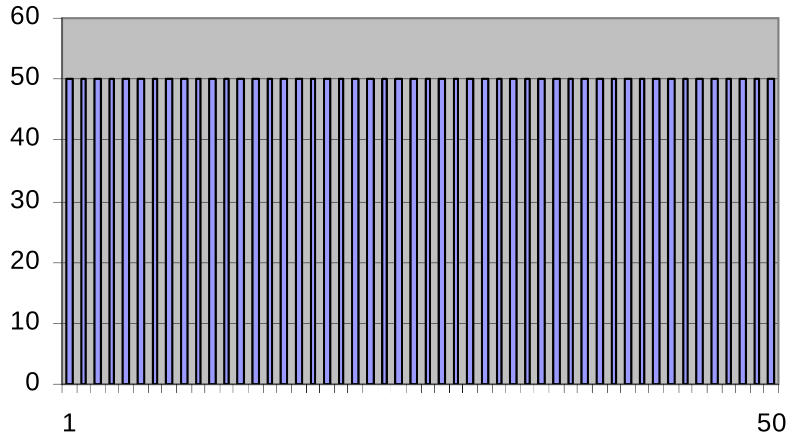
C50ir



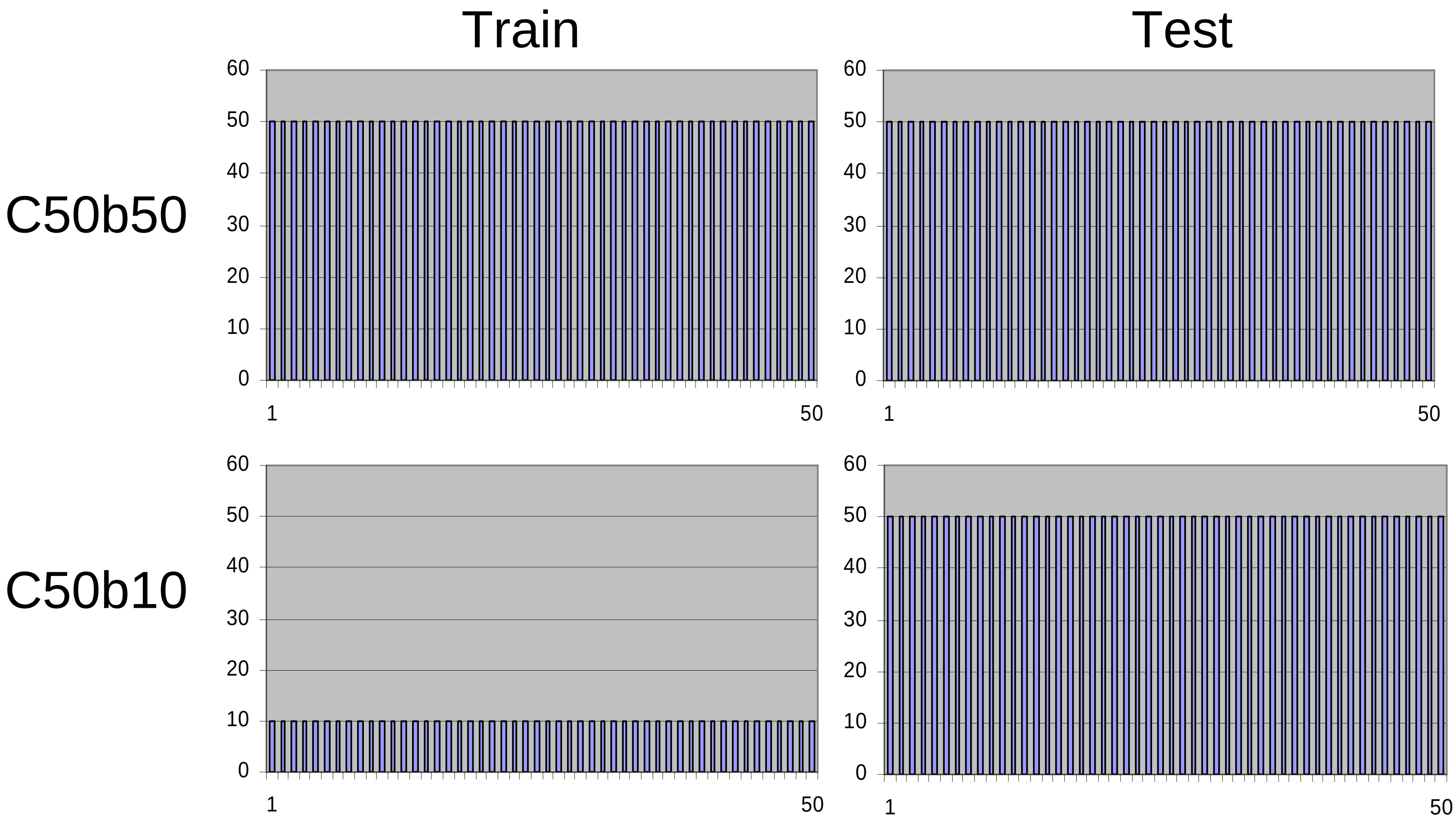


C50ig

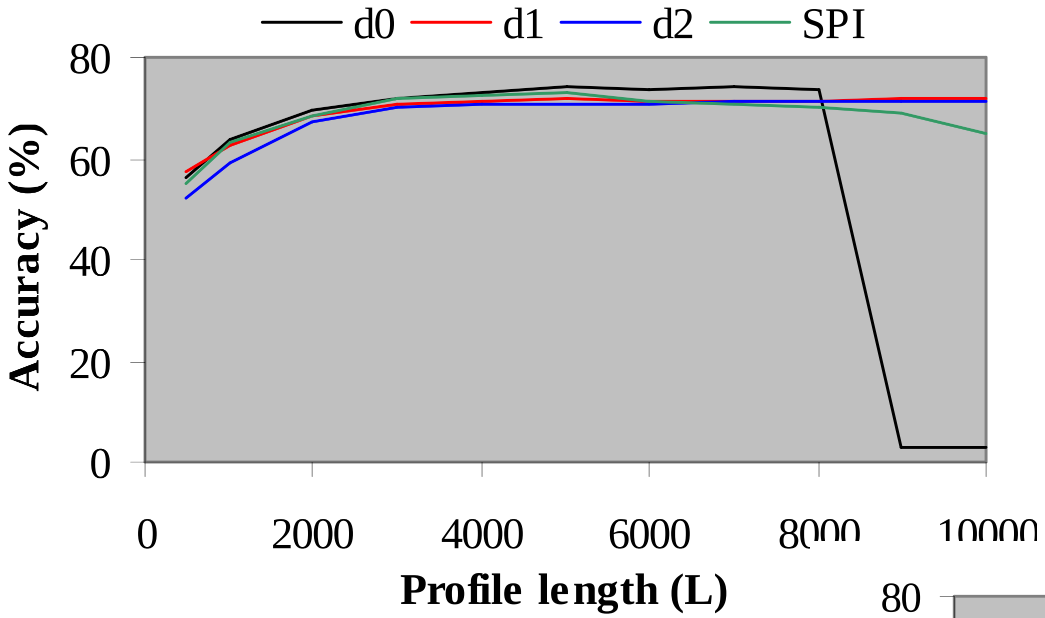


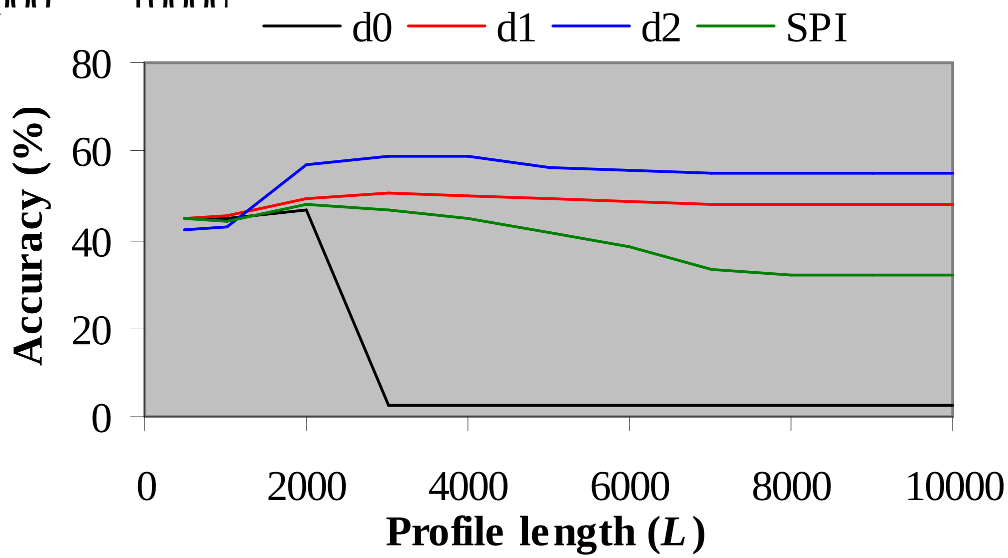


Distribution of C50b50 and C50b10

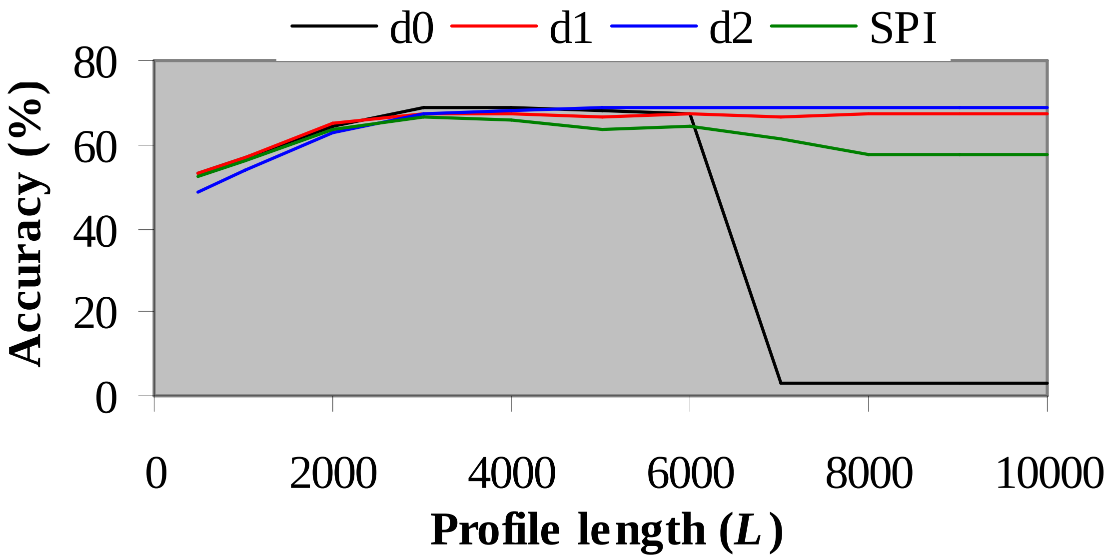


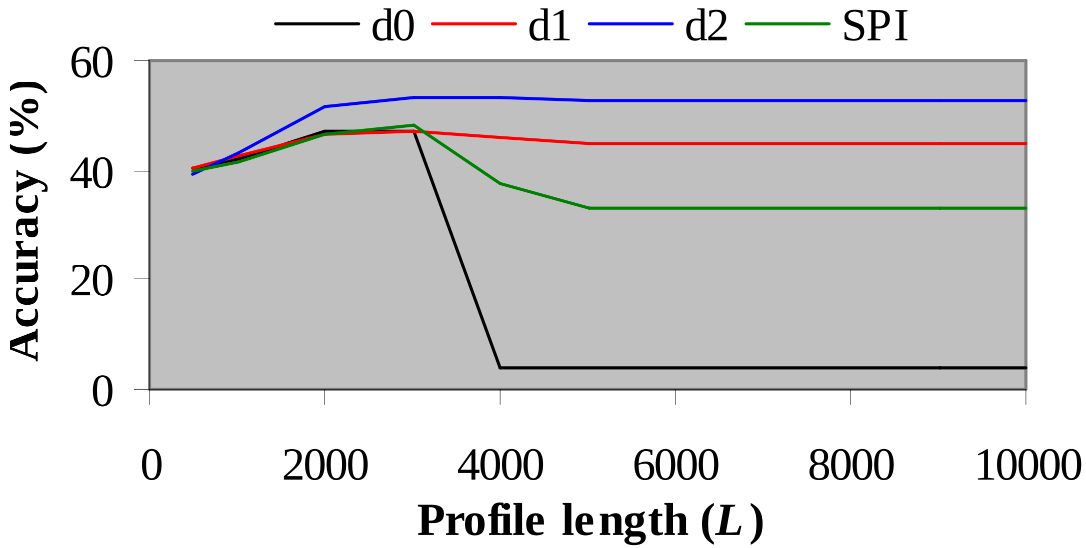
Results: C50ir and C50ig





Results: C50b50 and C50b10





Comparison with Other Approaches

	C50ir	C50ig	C50b50	C50b10
RAR	71.35	16.68	66.08	50.64
SVM	**84.60**	52.24	**73.60**	50.80
CNG-d_0	73.61	46.68	69.04	47.16
CNG-d_2	71.23	**58.68**	68.52	**53.16**

- SVM model based on 10,000 most frequent character 3-grams
- RAR is a parameter-free, compression-based model
- SVM and CNG-d_0 are better when many training texts are available
- CNG-d_2 is superior when limited and imbalanced training texts are available

Conclusions

- The class imbalance is an important problem for author identification tasks
  - Instance-based approaches
  - Profile-based approaches
- The proposed distance measures provide robust solutions for imbalanced and limited training sets
- The corpus norm factor enhances the distance estimation
- Based on d_2, practically we don't care about L
  - We still have to predefine n

THANK YOU!

Further information:
http://www.icsd.aegean.gr/lecturers/Stamatatos

Contact: stamatatos@aegean.gr