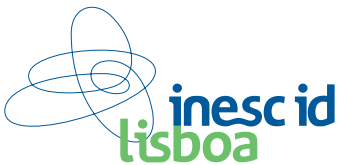
Semi-supervised Single-label Text Categorization using Centroid-based Classifiers

Ana Cardoso-Cachopo Arlindo Oliveira

Instituto Superior Técnico — Technical University of Lisbon / INESC-ID

SAC-IAR 2007, March 12th



Outline

Problem Description

- Text Categorization
- Single-label
- Datasets
  - Reuters 21578 - R8
  - 20 Newsgroups - 20Ng
  - Web Knowledge Base - Web4
  - Cade - Cade12

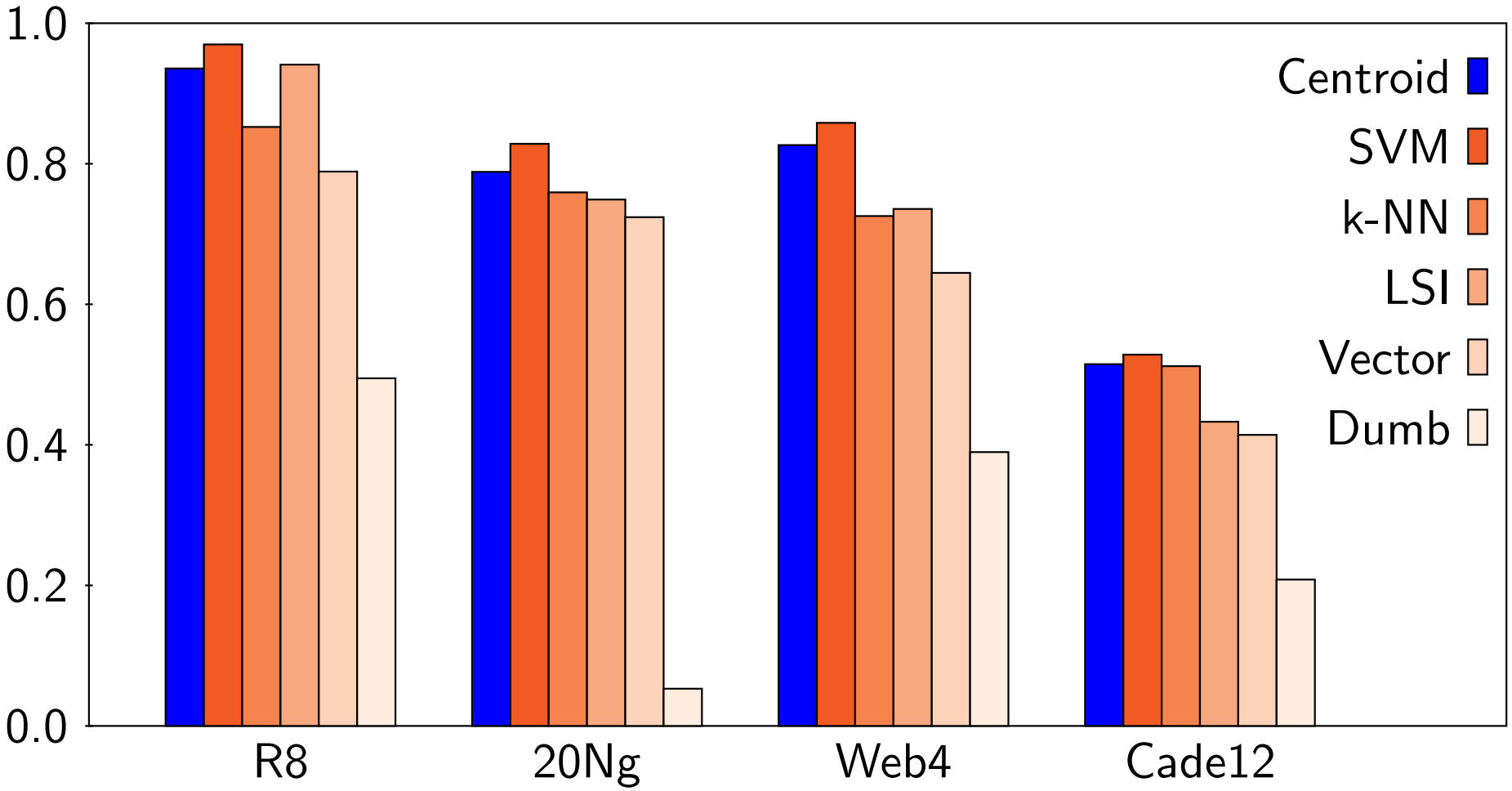
Characteristics of the Datasets

	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
R8	5485	2189	7674	51	3923
20Ng	11293	7528	18821	628	999
Web4	2803	1396	4199	504	1641
Cade12	27322	13661	40983	625	8473

Numbers of documents for the datasets: number of training documents, number of test documents, total number of documents, number of documents in the smallest class, and number of documents in the largest class.

Why use Centroid-based Methods

- Very fast
- Good Accuracy



Why use Unlabeled Data

- Small amounts of labeled data available
- Large amounts of unlabeled data available
- Hard or expensive to label new data

Incorporate Unlabeled Data using EM

If the entire dataset is available from the start, like in a library.

Inputs: A set of labeled document vectors, L, and a set of unlabeled document vectors U.
Initialization step: For each class c_j appearing in L, determine the class´s centroid $\overrightarrow{c_j}$, using one of the formulas for the centroids and considering only the labeled documents.
Estimation step: For each unlabeled document $d_j \in U$, classify it according to the available centroids.
Maximization step: For each class c_j, update its centroid $\overrightarrow{c_{j_{new}}}$, considering the labeled documents and the labels for the unlabeled documents obtained in the previous step.
Iterate: Until the centroids do not change in two consecutive iterations.
Outputs: For each class c_j, the centroid $\overrightarrow{c_j}$.

Incrementally Incorporate Unlabeled Data

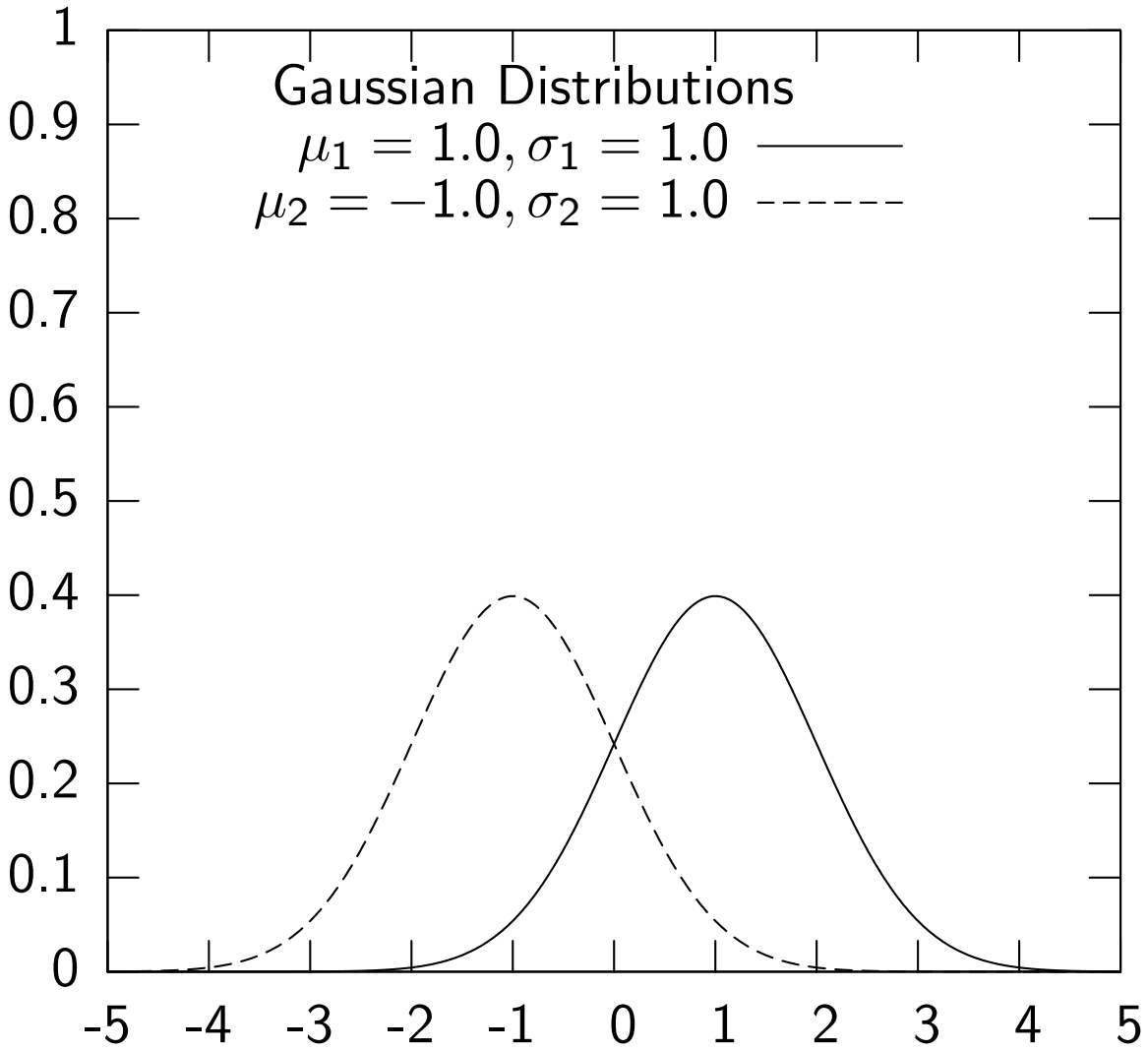
If the dataset changes over time, like a news feed or the web.

Inputs: A set of labeled document vectors, L, and a set of unlabeled document vectors U.
Initialization step: For each class c_j appearing in L, determine the class´s centroid $\overrightarrow{c_j}$, using one of the formulas for the centroids and considering only the labeled documents.
Iterate: For each unlabeled document $d_j \in U$:

- Classify d_j according to its similarity to each of the centroids.
- Update the centroids with the new document d_j classified in the previous step.

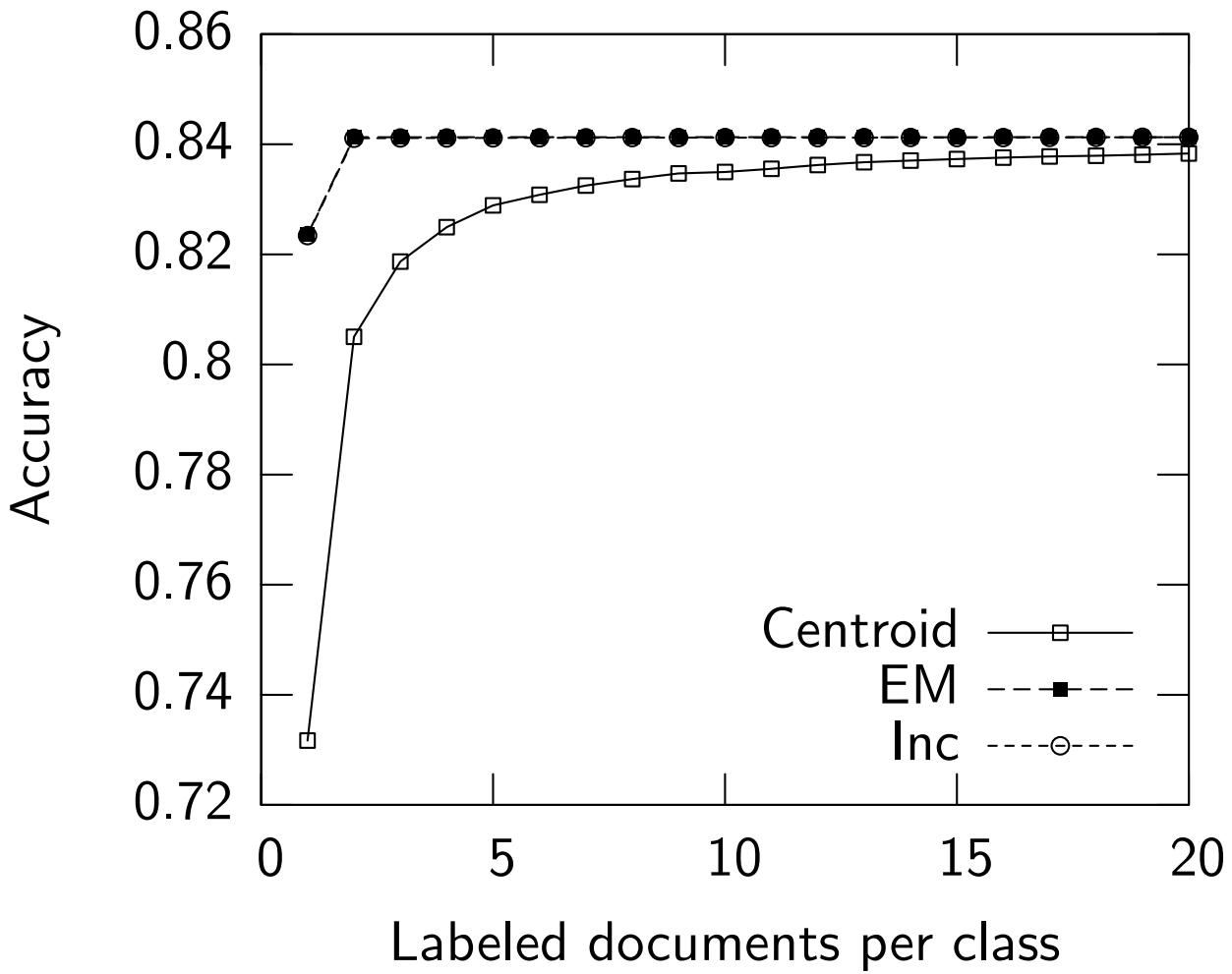
Outputs: For each class c_j, the centroid $\overrightarrow{c_j}$.

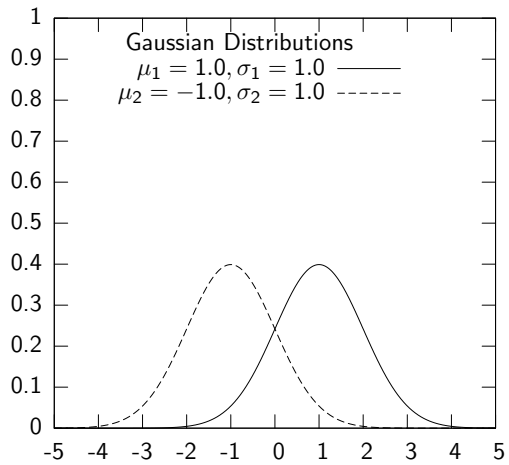
Experimental Results - Synthetic Dataset



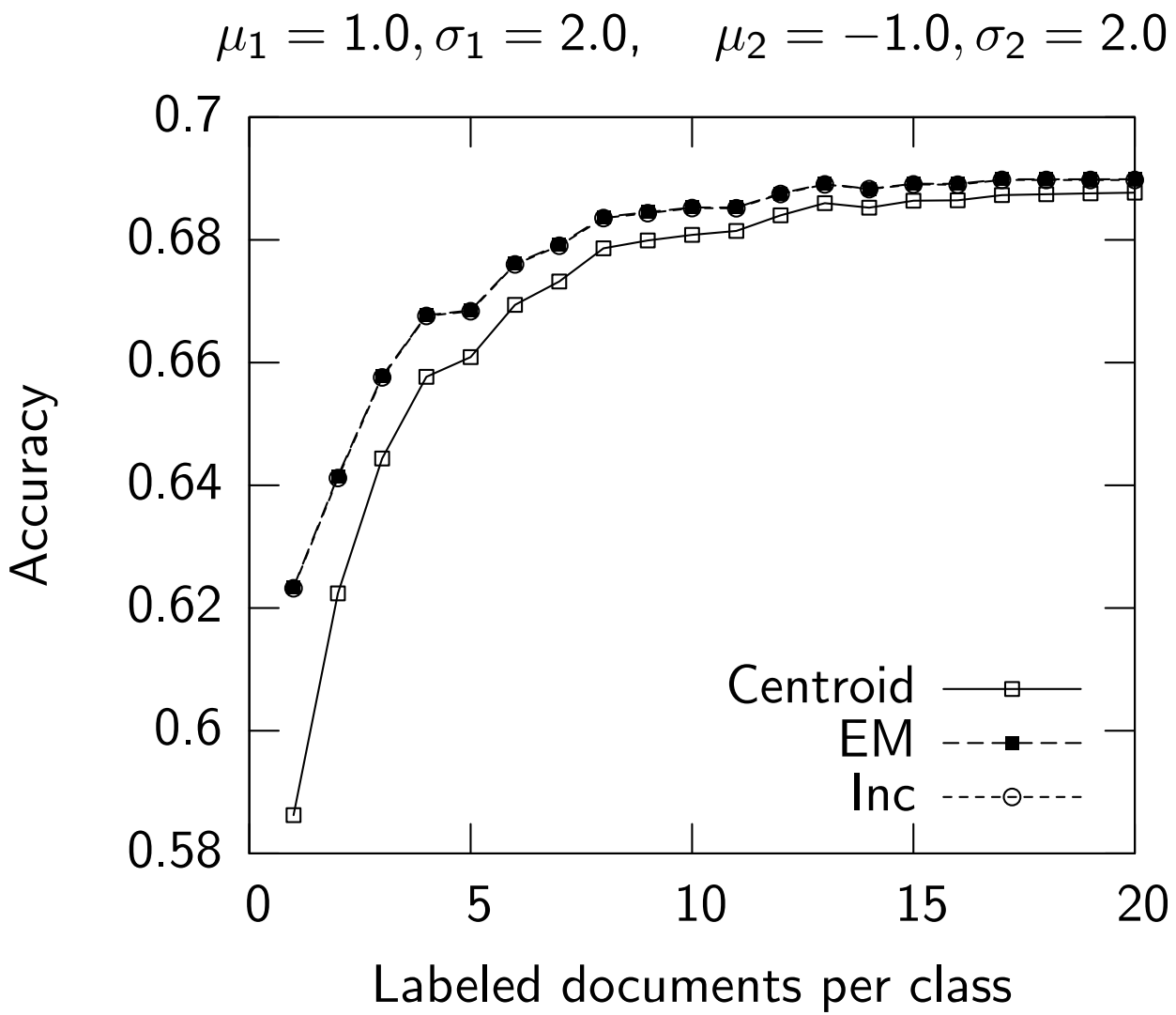
Experimental Results - Synthetic Dataset

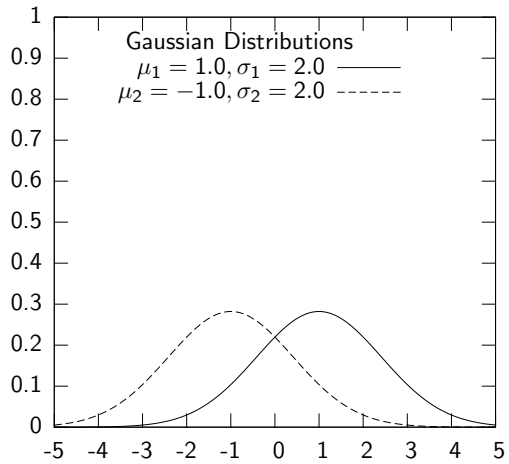
$\mu_1 = 1.0, \sigma_1 = 1.0, \quad \mu_2 = -1.0, \sigma_2 = 1.0$





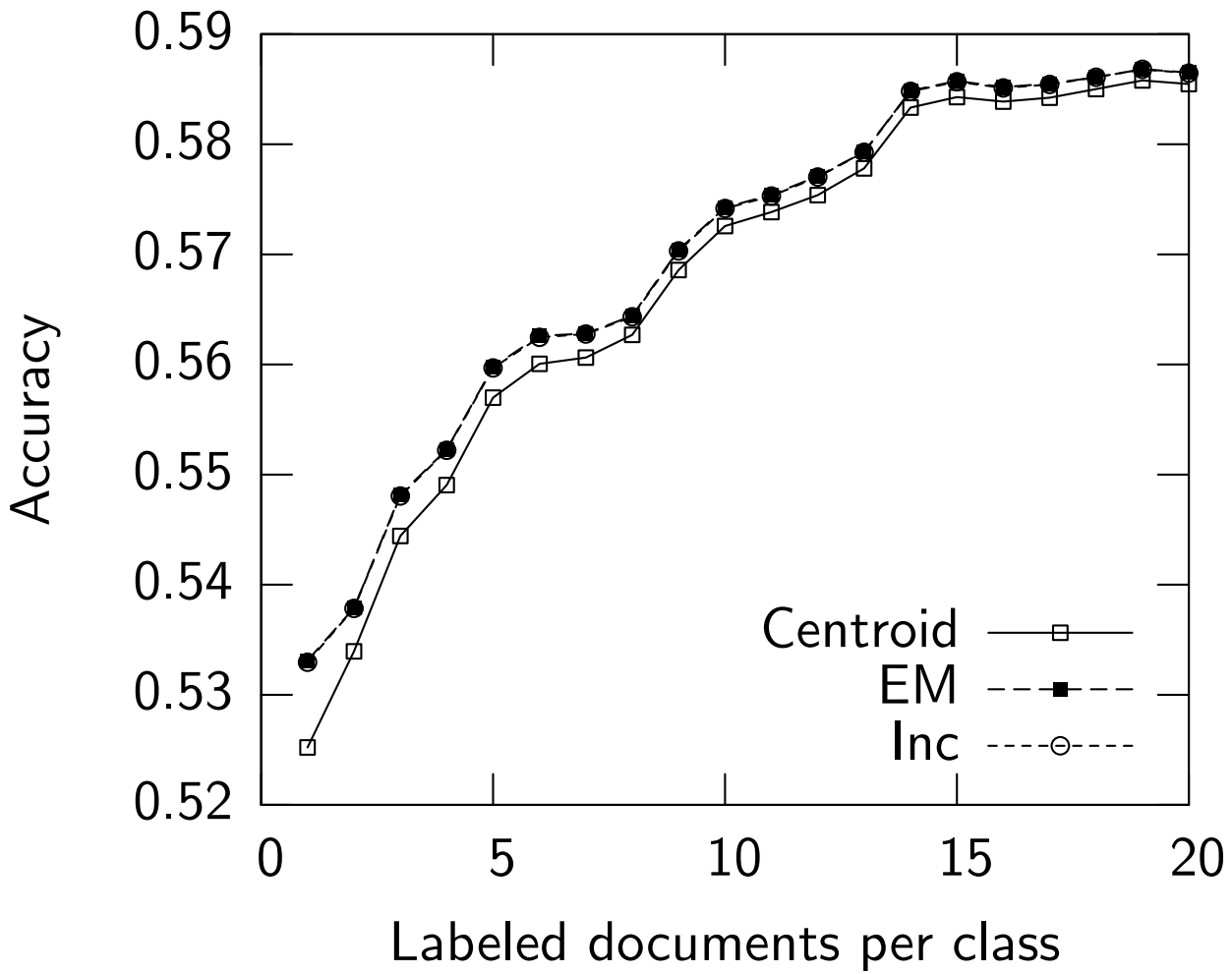
Experimental Results - Synthetic Dataset

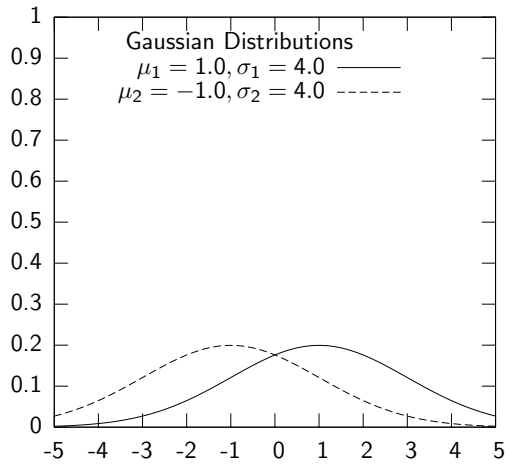




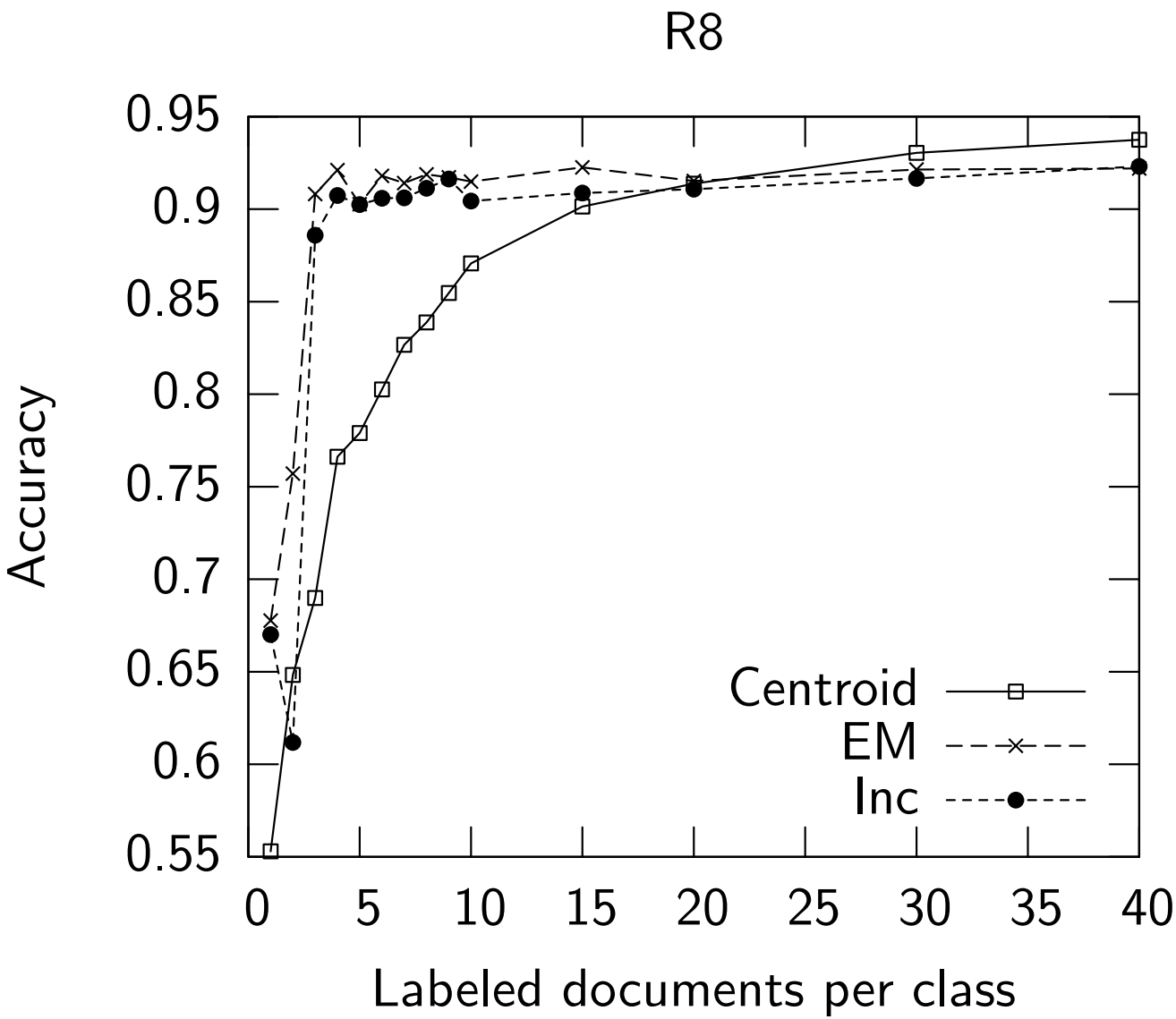
Experimental Results - Synthetic Dataset

$\mu_1 = 1.0, \sigma_1 = 4.0, \quad \mu_2 = -1.0, \sigma_2 = 4.0$

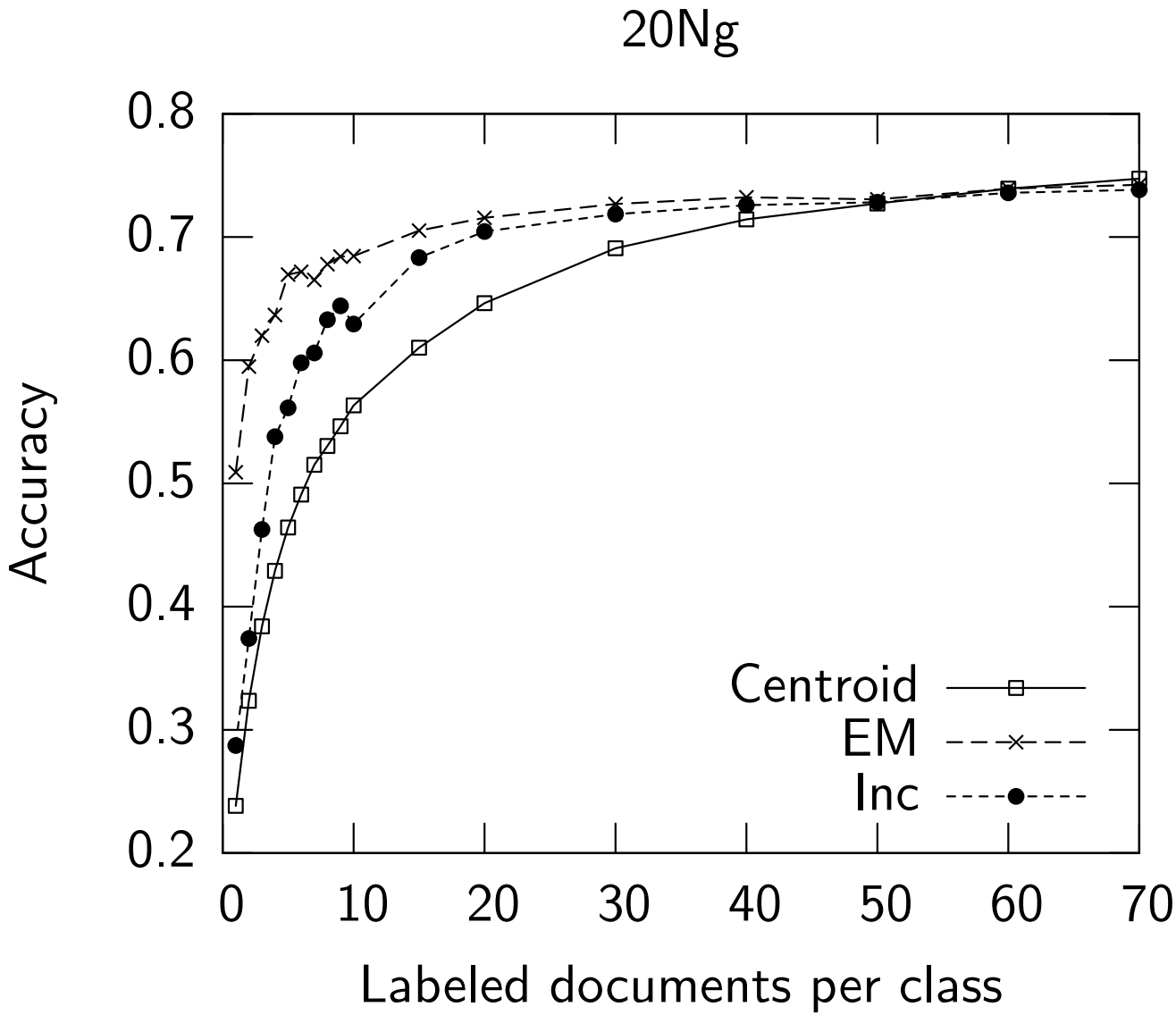




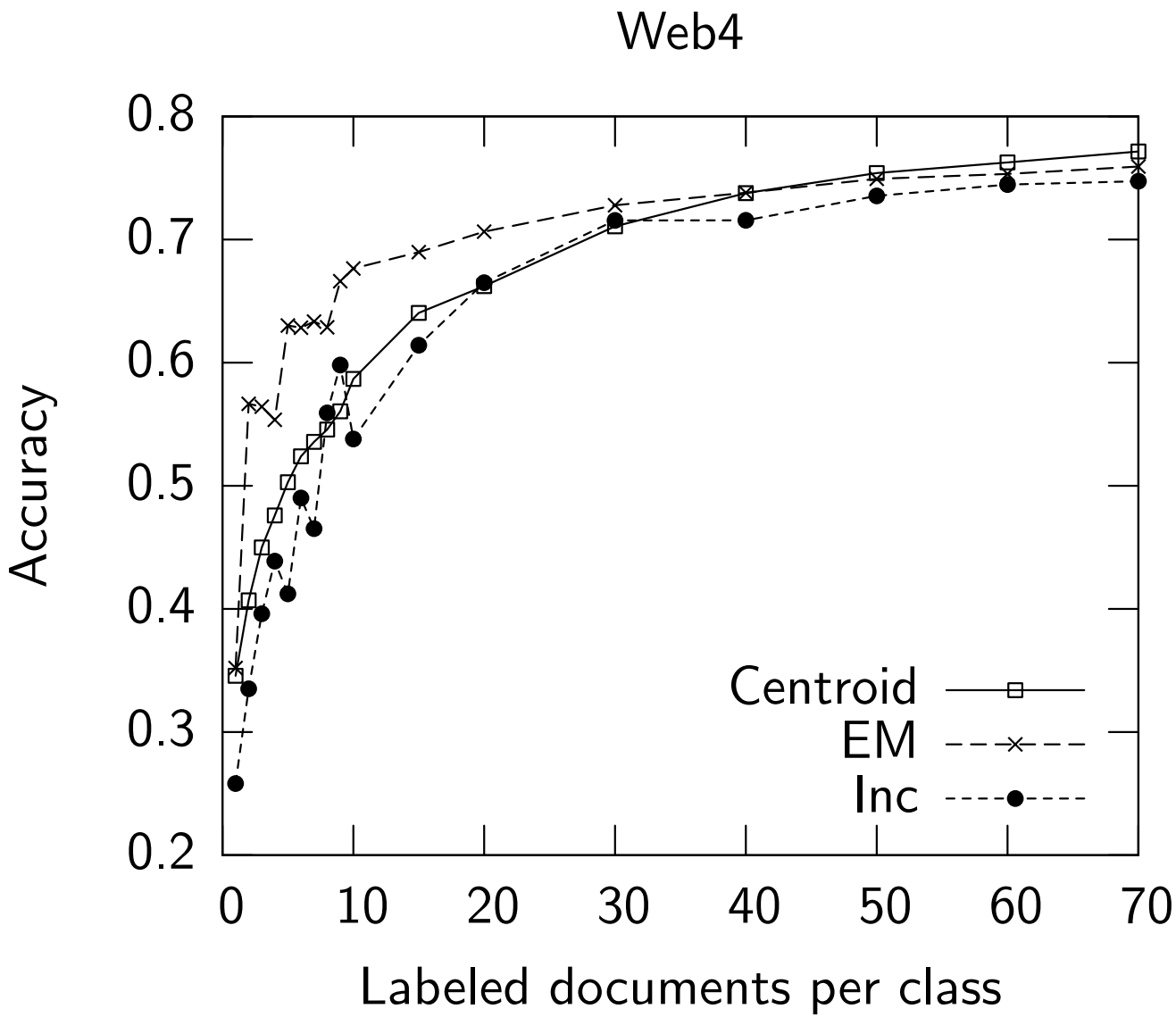
Experimental Results - Real World Datasets



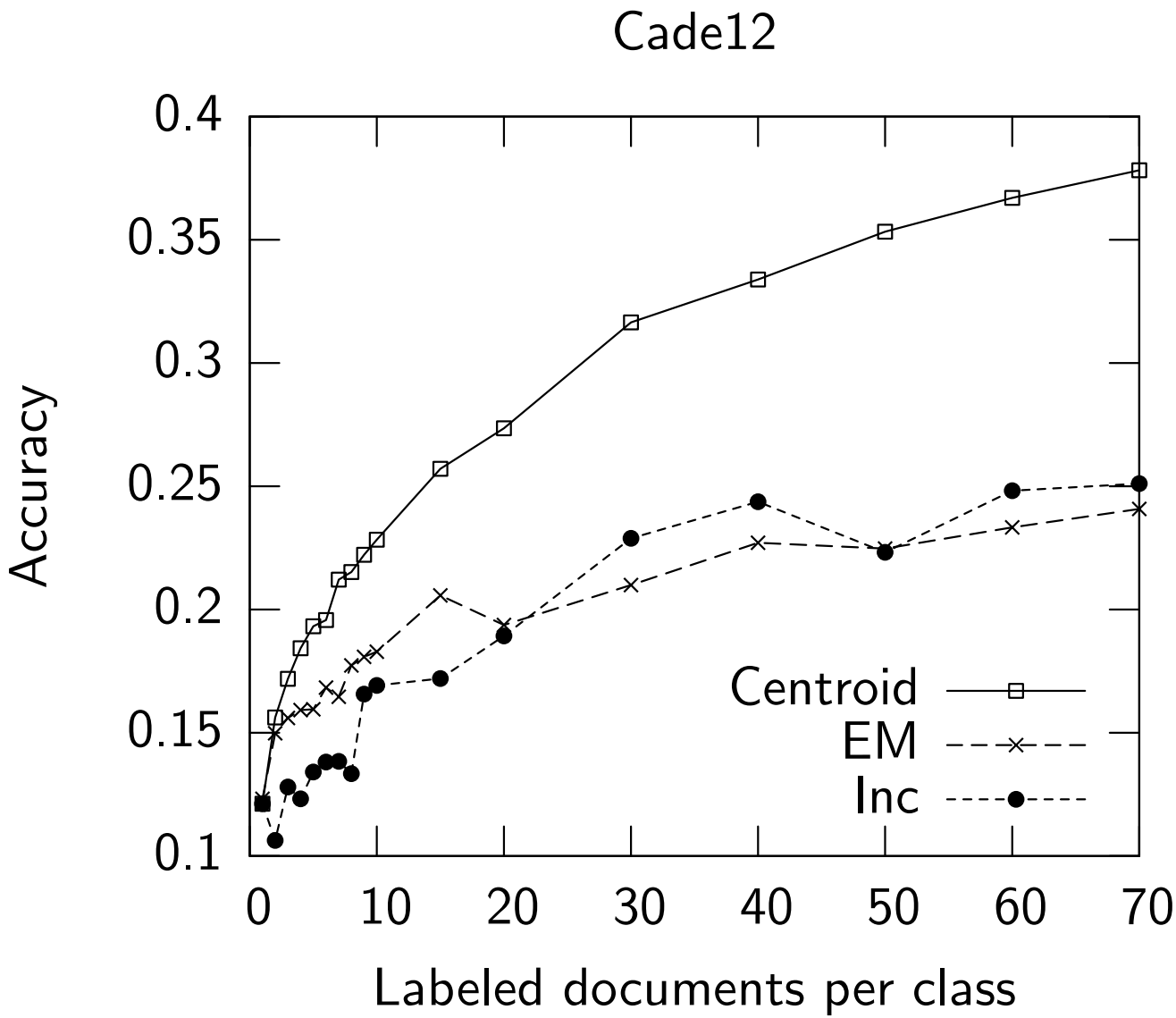
Experimental Results - Real World Datasets



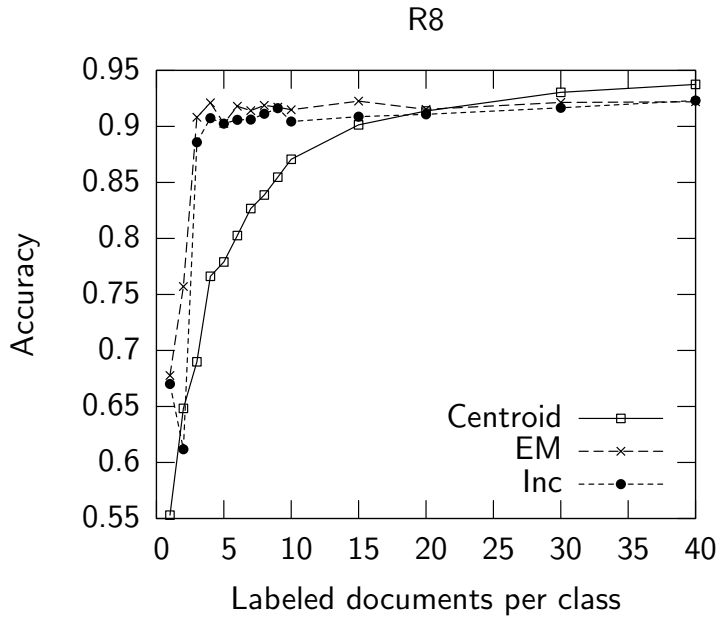
Experimental Results - Real World Datasets

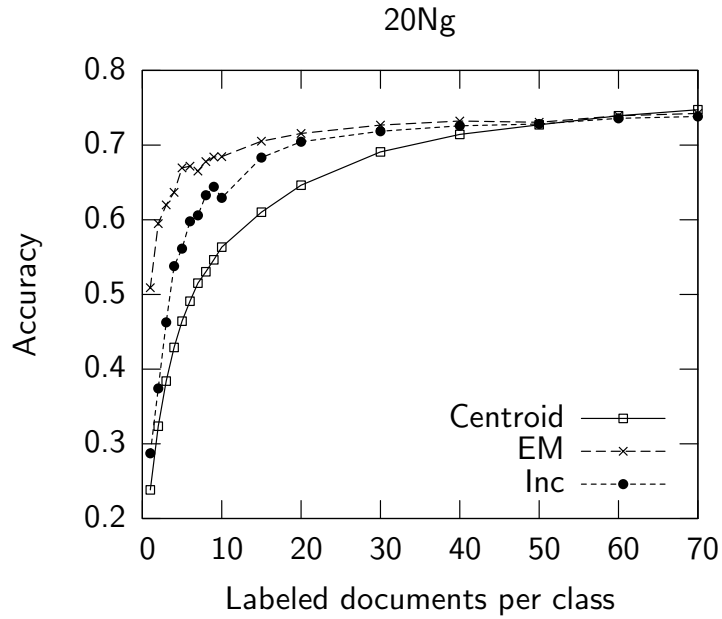


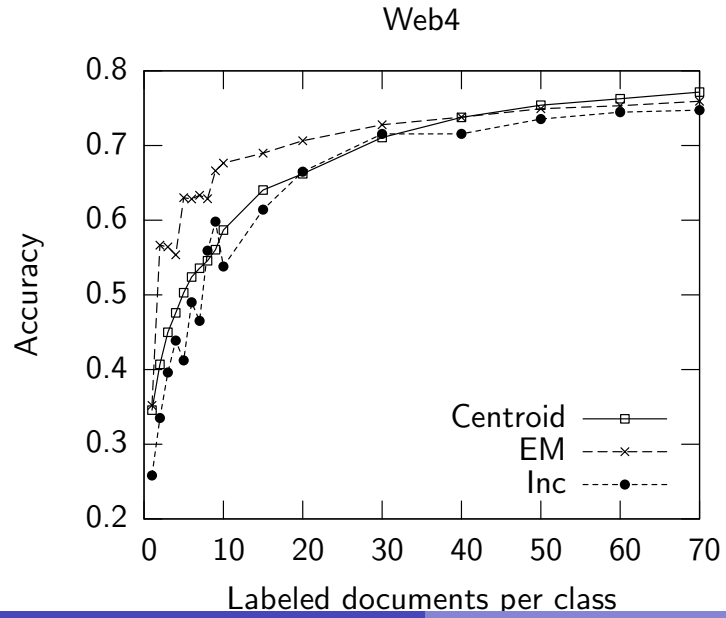
Experimental Results - Real World Datasets

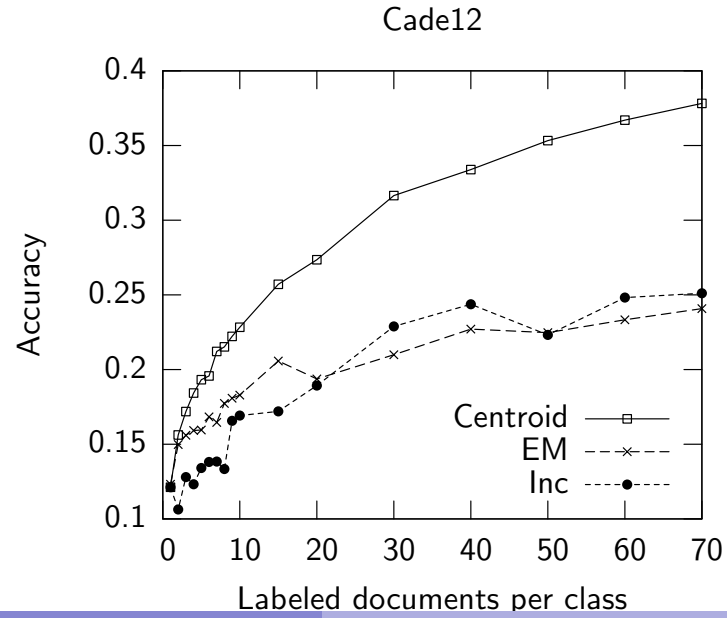


Experimental Results - Real World Datasets









Conclusions and Future Work

- If the initial model of the data is sufficiently precise, using unlabeled data improves performance.
- Using unlabeled data degrades performance if the initial model is not precise enough.
- As future work, we plan to extend this approach to multi-label datasets.

Thank You.

Any Questions?