Mining Enron Emails Using Data Streaming Techniques

Semih Sahin
20801301

Outline

- Problem Description
- Importance of the Problem
- Methodology
  - Stream Processing Concepts
  - Incremental Nearest Neighbor Algorithm
- Experimental Results
- Conclusion

Problem Description

- Social Network Extraction with Email Analysis
  - Relationship
  - Relationship type (personal, professional)

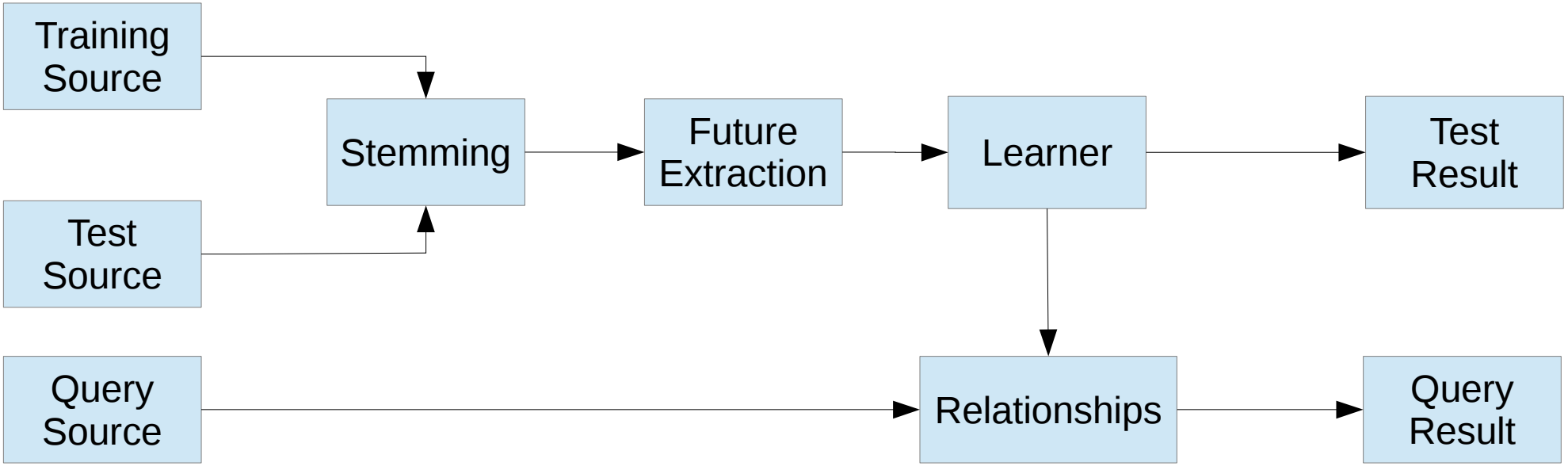
- Data Stream Mining

Importance of the Problem

- Real time analysis
- More specific community detection
  - List personal contacts
  - List professional contacts

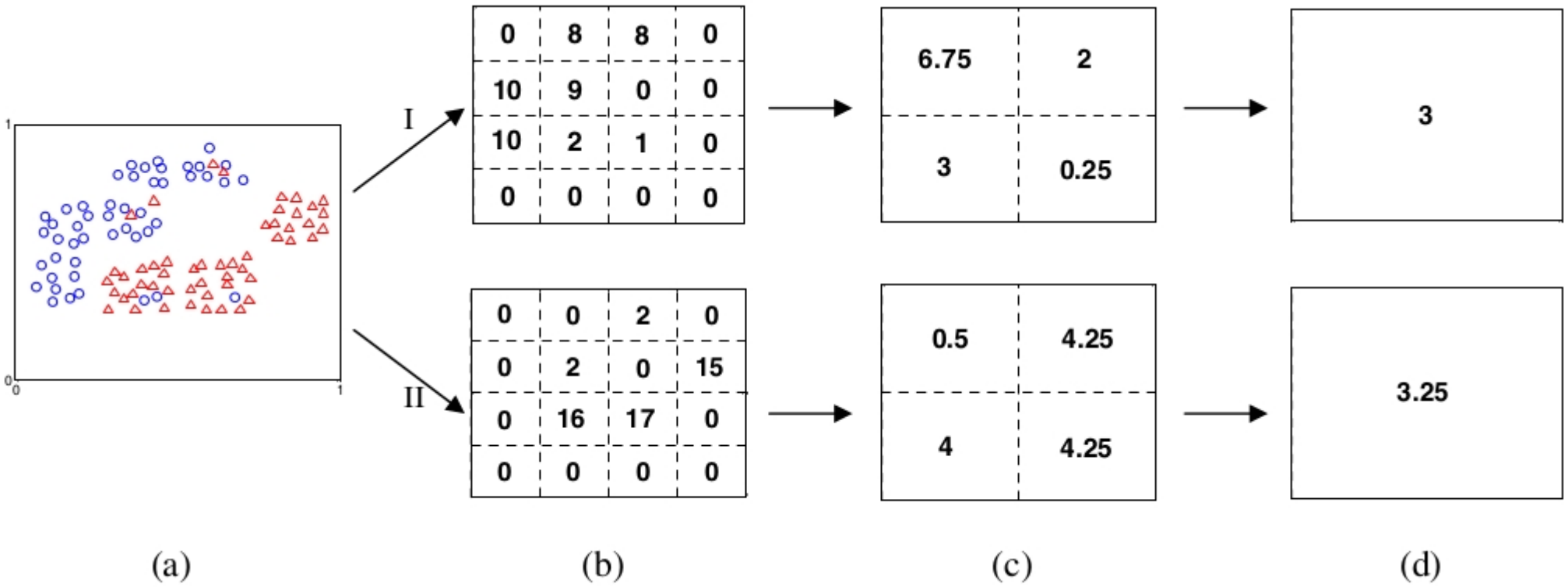
Methodology

- Stream Processing Concepts
  - Real time analysis
  - No batch processing
  - Single scan algorithms
  - Pipeline and data parallelism



Methodology

- Incremental Nearest Neighbor Algorithm
  - d-dimensional data
  - divide each dimension into g equal intervals



Methodology

- Incremental Nearest Neighbor Algorithm

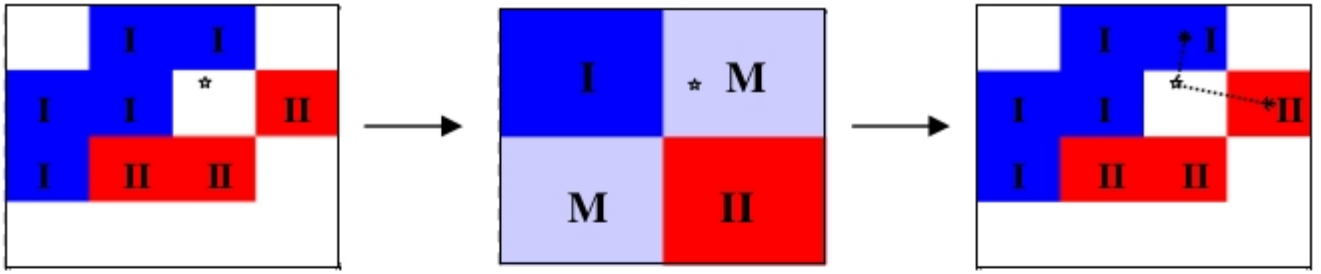


Fig. 3. Hierarchical classifier access

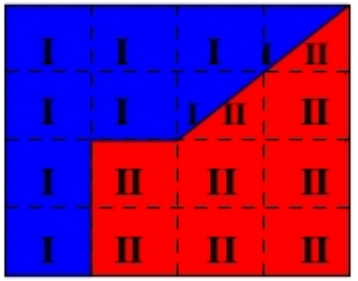


Fig. 4. The combined classifier

Experimental Results

- Semih-Bugra Stream Processing System
  - User defined operators
  - c++
- Enron Email Collection
  - ~1700 labeled emails
- Effectiveness
  - Respect to d, and g

Conclusion

- Emails can be used for social network extraction
- Incremental text classification algorithm can be used for real time analysis.

Thank you!