Optical Character Recognition

Lecture 5

Qurat-ul-Ain (Ainie) Akram
Sarmad Hussain

Center for language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore, Pakistan

Learning Algorithms

- Supervised Learning
  - Machine learning process of predicting a function from labeled training data
  - Training process
    - Training data consists of an input object and a desired output value
    - The learning algorithm processes and analyzes the training data and produces an inferred function
  - Recognition process
    - Input example is given
    - The output value is computed using the inferred function
  - Examples
    - Decision Trees
    - HMMs
    - Bayesian Networks

Learning Algorithms

- Un-supervised Learning
  - Machine learning process of trying to predict or infer a function in unlabeled data
    - Training data consists of an input objects having no labeled information
    - The learning algorithm processes and analyzes the training data and produces an inferred function
  - Examples
    - Clustering
    - Neural network models
    - Self-organizing map (SOM)
    - Adaptive resonance theory (ART)

Supervised Learning for OCR

- Decision Trees
- Steps for training
  - Data preparation
  - Features computation
  - Training: Decision tree formation
  - Coding of decision trees

- Decision Trees
- Steps for recognition
  - Input character shape
  - Features computation
  - Recognize shape using trained decision tree

Supervised Learning for OCR

- Feature Selection
  - List features which can be used to disambiguate the following shapes

Supervised Learning for OCR

- Feature extraction and computation
  - Analyze data and extract significant features
  - Dimensional features
    1. Height
    2. Width
    3. Total number of Black Pixels
    4. Density

       Density=Total Number of Black Pixels /Height*Width
  - Morphological Features

Supervised Learning for OCR

- C45 Decision Tree Algorithm
  - Weka installation and setup (http://weka.wikispaces.com/)
  - Feature computation from images along with labeled class
  - Decision tree using weka
  - Code decision tree in java

Project Deliverables

Deliverable	Urd Group	Nep Group	Sin Group	Sin Group 2	Snd Group	Tam Group
Data Preparation 1. Document Image 2. MBs Real Data 3. Diacritics Real Data						
Binarization						
Line Segmentation						
Ligature/Syllable Segmentation						
Ligature/Syllable Segmentation						
Diacritics Training and recognition						
Main Body Classification and Recognition						
Syllable String Creation						