Adaptive handwriting recognition: Adaptive classifiers and an adaptive committee

Matti Aksela

Neural Networks Research Centre
Laboratory of Computer and Information Science
Helsinki University of Technology

8. December 2005

Introduction

- Handwriting recognition
  - vast amount of intrinsic variation
  - limited amount of models
  - how to take into account all variability in the data?

- Additionally, situation may change in time
  - for example in handwriting recognition writing style may change due to
    1. position (standing vs. sitting)
    2. movement (in a car, train, bus,...)
    3. speed (in a hurry vs. taking ones time)
    4. entirely different writers
    5. many, many different reasons

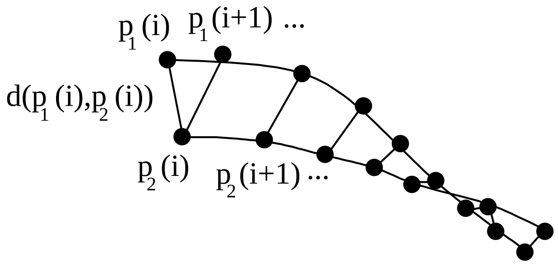
- Adaptation
  - a classifier can attempt to adapt to a particular writer to obtain optimal performance
  - classifier adaptation or committee adaptation possible
  - classifier adaptation can be more efficient due to the adaptation taking place on for example the prototype set
  - committee adaptation can use a variety of classifiers without need of detailed information on the problem at hand and still provide significant improvements

- Combining classifiers
  - take the outputs of a set of member classifiers
  - attempt combine the results in a way that improves performance
- Members have a significant effect on performance
  - classifiers' individual error rates
  - correlatedness of the errors made by the classifiers
- The more different the mistakes made by the classifiers are, the more beneficial the combination of the classifiers can be

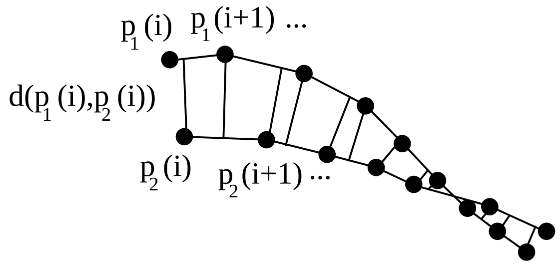
- Adaptive combination of adaptive classifiers
  - when striving for the best possible performance, combining adaptive classifiers in an adaptive fashion could be interesting
  - problem: predicting the behavior of adaptive classifiers is very difficult, as their behavior by definition changes in time
  - solution: balance between learning and robustness

Adaptive Classifier

- Classification based on using the k-NN rule on matching the input sample to a prototype set
- Member classifiers use Dynamic Time Warping distances; PP, NPP, PL



Point-to-point



Point-to-line

- When using a prototype based classifier, the prototype set may be modified:
  - adding new prototypes
  - remove bad or unused prototypes
  - adjust prototypes (LVQ variant)
  - hybrid approach: add new prototypes if none of the k nearest neighbors correct, otherwise adjust

Adaptive Committee Classifier

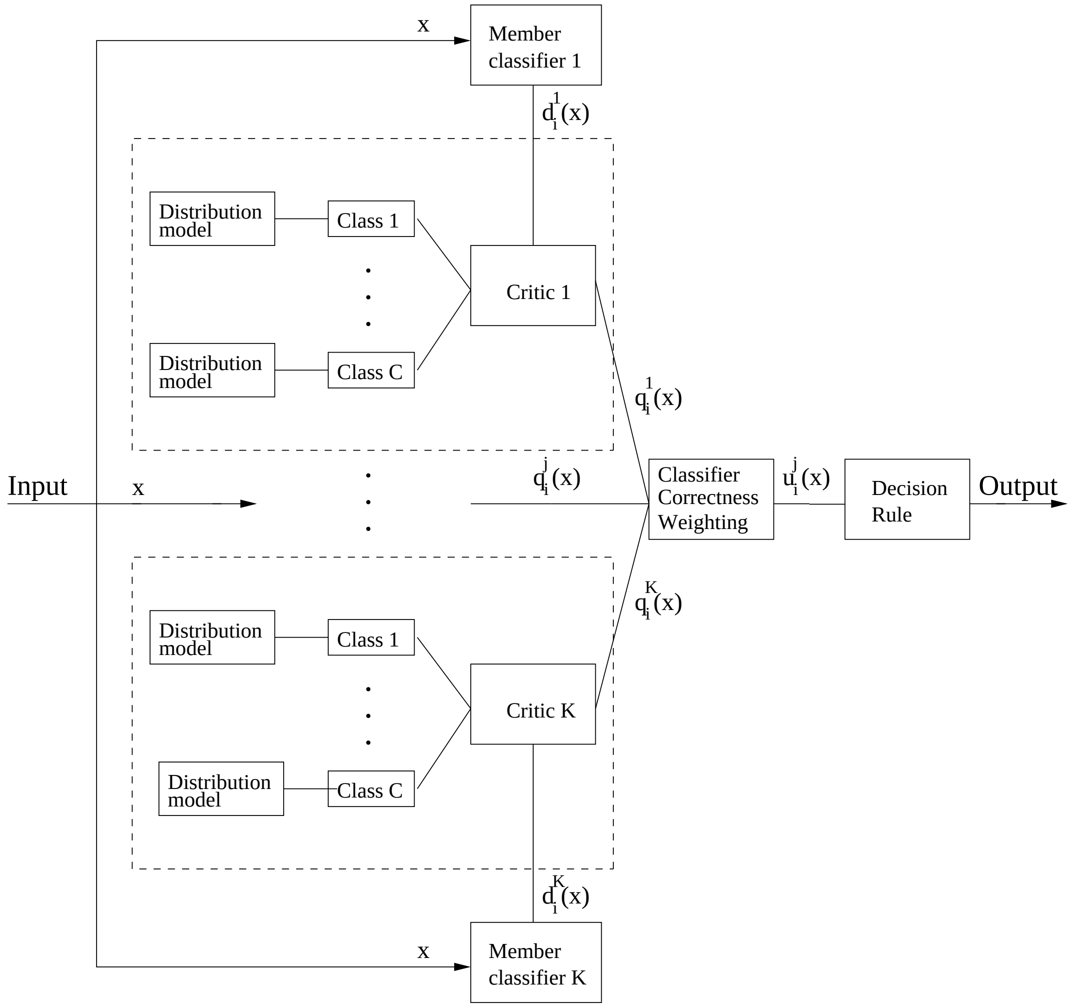
- A way to estimate confidence in classifier decisions
  - separate classification units that makes decisions on the correctness of the members can be used, called critic-driven schemes
  - an estimate can also be based on prior data

- The presented scheme:
  1. confidence evaluation based on information on previous decisions
  2. a balance between the impact of the older and more recent samples
     - use a weighting scheme to focus on more recent samples

Adaptive Committee Classifier

- Class-Confidence Critic Combining (CCCC)
  - experts assess member classifier correctness
  - confidence in decisions estimated from previous behavior in same class
  - critics produce confidence values used in classification

- Committee structure
  - one critic for each classifier
  - one distance distribution per class in each critic
  - combination scheme based on the evaluated confidences
  - added weighting scheme to balance prior sample impact


Member classifier 1
x
$d_i^1(x)$
Distribution model
Class 1
Distribution model
Class C
Critic 1
$q_i^1(x)$
Input
x
$q_i^j(x)$
Classifier Correctness Weighting
$u_i^j(x)$
Decision Rule
Output
$q_i^K(x)$
Distribution model
Class 1
Distribution model
Class C
Critic K
$d_i^K(x)$
x
Member classifier K

Operation overview

1. Member classifiers produce classification results and distances
2. Distances are normalized
3. Confidences for classifications are calculated in critics based on the normalized distances and prior data
4. Committee decision based on confidences
5. Two-phased committee adaptation

CCCC step 1: Member classifiers

- Member classifiers produce classification results
  - a sample x is input
  - the input is classified by all K member classifiers who produce for every one of C possible classes a distance-indicating value $d_c^k(x)$
  - if the classifiers doesn't work with distances, transform measure to be used into a distance (for example, for a confidence measure $t \in [0 - 1]$ use $1 - t$)
  - $d_c^k(x) \in [0, \inf]$; distance to the nearest prototype of class c from classifier k
  - an infinite distance may be produced if matching is not possible

CCCC step 2: Distance normalization

- The normalized distance is defined as

$$q_c^k(x) = \begin{cases} \frac{d_c^k(x)}{\sum_{i=1}^{C} \hat{d}_i^k(x)} & \text{, if } d_c^k(x) \text{ is finite} \\ 1 & \text{, otherwise} \end{cases}$$

  where $\hat{d}_c^k(x)$ equals $d_c^k(x)$ if $d_c^k(x)$ is finite and is otherwise zero

- If the distance to only one class is finite, the normalized distance for that class is defined to be zero

CCCC step 3: Confidence calculation

- The received $d^k(x)$ values are modeled by gathering previous values into distributions from which the value for the confidence can be obtained. To shorten the notation, $p^i(q_c^k(x)) = p_c^i(z)$.

- Exponential kernel distribution estimate

$$p_c^k(q_c^k(x)) = \frac{1}{\sum_{j=1}^{N_i} w_i(z_j^i)} \sum_{j=1}^{N_i} w_i(z_j^i) e^{-\frac{|z - z_j^i|}{b}}$$

CCCC step 4: Decision

- Overall confidence is calculated from the confidence obtained from the critics and the corresponding classifiers correctness rate

$$u_c^k(x) = p_c^k(q_c^k(x)) \cdot q_c^k(x) \cdot p(\text{classifier } k \text{ correct})$$

- Final decision using the Sum rule:

$$c(x) = \arg\max_{j=1}^{C} \sum_{k=1}^{K} u_j^k(x)$$

CCCC step 5: Weights and adaptation

- Information of the correctness of the classification is assumed to be known
- Two-phased adaptation
  1. classified samples $d^k(x)$ values are inserted into the critics' distribution model whenever that particular critics' classifier was correct
  2. each sample in the distribution models is assigned a weight
- weights in the distribution are recalculated to decrease in accordance with a decay constant λ

$$w_i(z_i^j) = \max\{0, 1 - \lambda(N_i - n_i(z_j^i))\}$$

Committee operation example

- Example: K classifiers and C classes

1. Each classifier classifies the input, resulting in K vectors of C distance values $d_c^k(x)$
2. Distances are normalized to obtain values $q_c^k(x)$
3. Critics produce confidence values $p_c^k(q_c^k(x))$ from the distribution models of previous classification results
4. Confidence values are adjusted with the corresponding classifiers correctness rate to produce final confidences $u_c^k(x)$
5. Final output is selected using the sum rule
6. Distributions updated by appending the $q_c^k(x)$ values for the classifiers that were correct to the distribution models
7. Weights are recalculated according to the weighting scheme

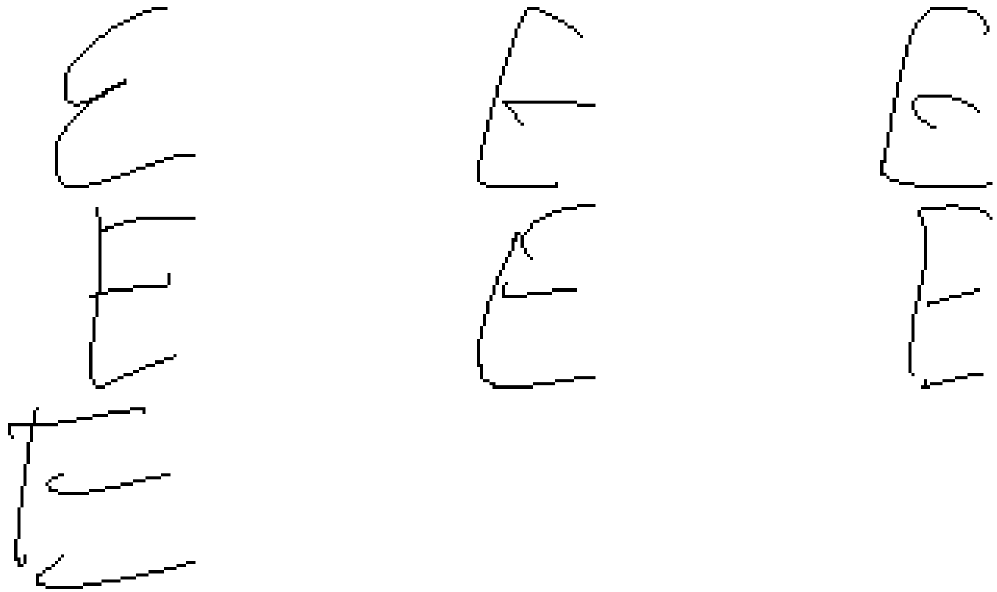
Experiments

- Handwritten character recognition, isolated on-line characters
  - collected on a Wacom Artpad II Tablet
  - stored in UNIPEN format
  - upper-case and lower-case letters and digits used

- Three databases

Database	Writers	Characters	Usage
DB1	22	9961	creating member classifiers
DB2	8	8077	evaluating parameters, ordering
DB3	8	8046	testing

- Character examples; some samples of the character 'E'



Adaptive classifier results

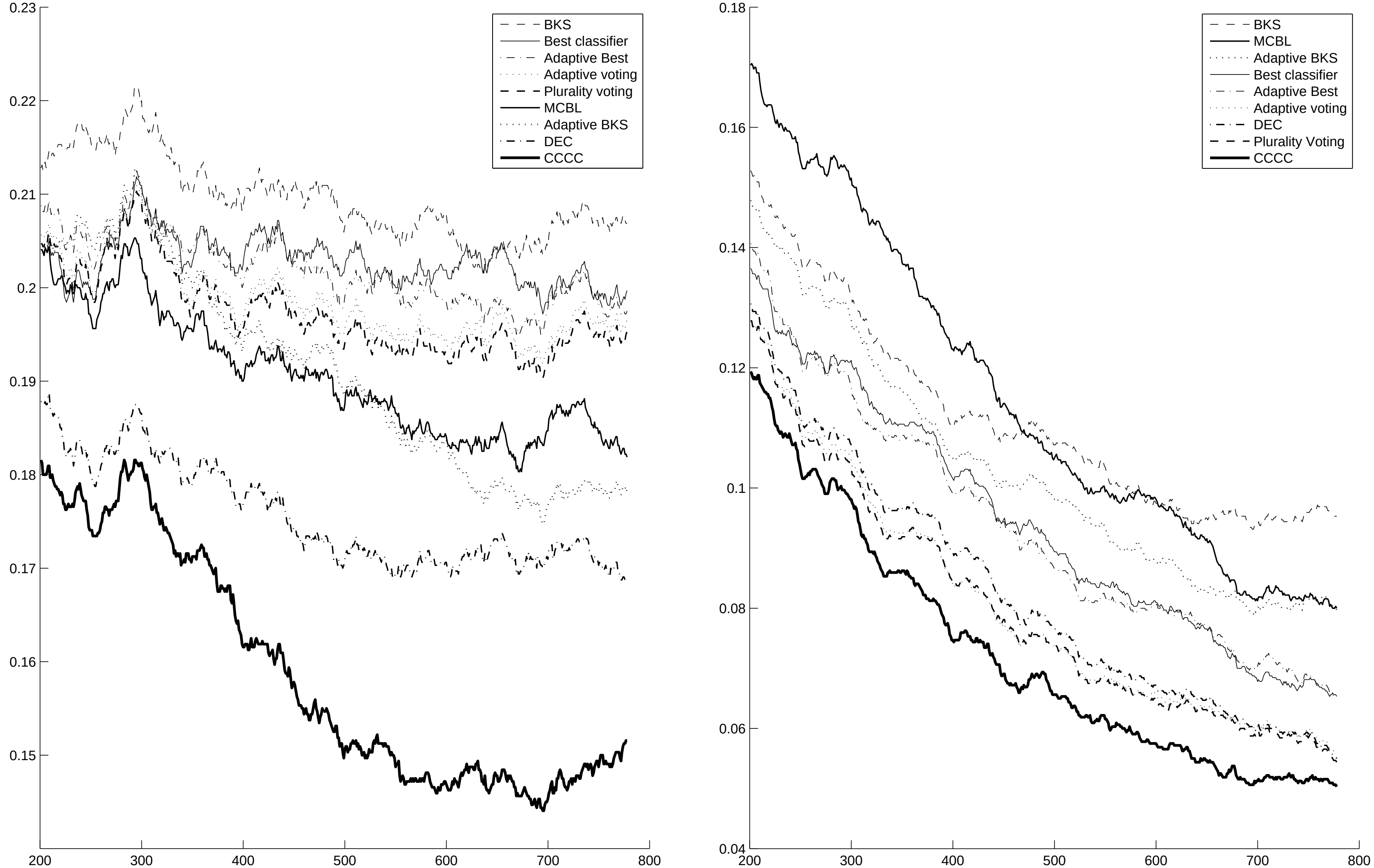
- **Dynamic Time Warping (DTW)** - based distances
  - point-to-point (PP), point-to-line (PL) or normalized point-to-point (NPP)
  - scaled using either mass center (MC) or bounding box center (BBC)
  - non-adaptive or adaptive using the hybrid adaptation approach

Member classifier	Error rate (non-adaptive)	Error rate (adaptive)
PP-MC	20.02%	9.87%
PP-BBC	21.18%	9.90%
NPP-MC	20.93%	10.24%
NPP-BBC	21.18%	10.70%
PL-MC	20.77%	15.56%
PL-BBC	22.28%	16.27%

Adaptive committee results

- CCCC Committee
  - compared with simple plurality voting and best individual classifier
  - combination of both adaptive and non-adaptive member classifiers

Committee	**Error rate (non-adaptive members)**	**Error rate (adaptive members)**
CCCC	15.53%	7.85%
Plurality voting	19.68%	8.69%
Best member	20.02%	9.87%


BKS
Best classifier
Adaptive Best
Adaptive voting
Plurality voting
MCBL
Adaptive BKS
DEC
CCCC
BKS
MCBL
Adaptive BKS
Best classifier
Adaptive Best
Adaptive voting
DEC
Plurality Voting
CCCC

Conclusions

- When applicable, adapting an individual classifier may produce most gain
  - however, not all classifiers equally suitable for adaptation (PP vs PL)
  - adaptation generally increases performance for one subject at the cost of generalization ability, ie. poor performance for other subjects
- Committee adaptation can also produce significant gain
  - as less information on the task at hand is available, adaptation is performed on a more abstract level and hence drastic changes can cause problems
  - robustness for also other subjects is easier to maintain
- Clearly the doubly adaptive strategy of an adaptive combination of adaptive member classifiers provided the best results