Less is More: Active Learning with Support Vector Machines

Greg Schohn, David Cohn

Presented by: Nikolaidou Panagiota

March 9, 2006

Outline

- Introduction
- Support Vector Machines
- A greedy optimal strategy
- A simple heuristic
- Experiments
- Conclusions

Inroduction

- labeled examples
  - obtained costly
  - presence of domain experts
- Solution: *active learning*
  - selects the training examples the most *informative*
  - increases performance by reducing the number of the training examples

Support Vector Machines

- defines a unique hyperplane that seperates positive and negative examples and for which the margin is maximized
- *soft SVM*
  - used when data are not separable
  - seperate data with a minimal number of errors
- *bound examples*
  - examples incorrectly classified
  - examples within the margin

A greedy optimal strategy

- based on *probabilities* assigned to points classified by SVM

$$P(y=1|x) = \frac{1}{1+exp(-f(x))}$$

  where f(x) is the output of SVM

- based on the *expected error* :
  sum of the expected error of each training example weighted by the distributions of test examples

A greedy optimal strategy

- algorithm:
  - for each candidate unlabeled example x, calculate $P(y = 1|x)$ and $P(y = -1|x)$
  - Add $(x, 1)$ to the training set, retrain, and calculate the new expected error $E_{(x,1)}$
  - Remove $(x, 1)$, add $(x, -1)$ to the training set, retrain, and calculate $E_{(x,-1)}$
  - Estimate expecting error as $E_x = P(y = 1|x) * E_{(x,1)} + P(y = -1|x) * E_{(x,-1)}$
  - Choose the unlabeled example x, which has the minimum E_x
- impractical: evaluating each candidate requires solving two QP problems

A simple heuristic

- example nearest to the dividing hyperplane
- for all the unlabeled examples find the distance between them and the hyperplane (dot product computation) and select the one that has the minimum distance
- reducement of the uncertainty area which is situated near the dividing hyperplane

Experiments

- two domains:
  - binary classification of 4 newsgroup pairs from the 20 Newsgroups data set
  - topic classification on a subset of five topics from Reuters
- number of examples in every iteration $= 8$
  - trade-off against the cost of re-solving a new QP problem (more examples per iteration, less QP problems) and the cost of labelling an example
- active learning performs better than random selecting

Experiments

- *stopping criterion*
    - when the margin has been exhausted ⇒ when there are no other training examples within the margin
- the performance increases up to a peak and after it starts to decrease
    - until the margin has been exhausted (until peak) ⇒ performance increases, the model remains consistent
    - when margin contains no available training data ⇒ examples that make the model inconsistent may be added (soft SVM), performance decreases

Conclusions

- reduce of the number of the training examples
- reduce in time
- give bounds for b
- accuracy decrease very soon ⇒ stop ?