Committee Based Approaches to Active Learning

Jerzy Stefanowski

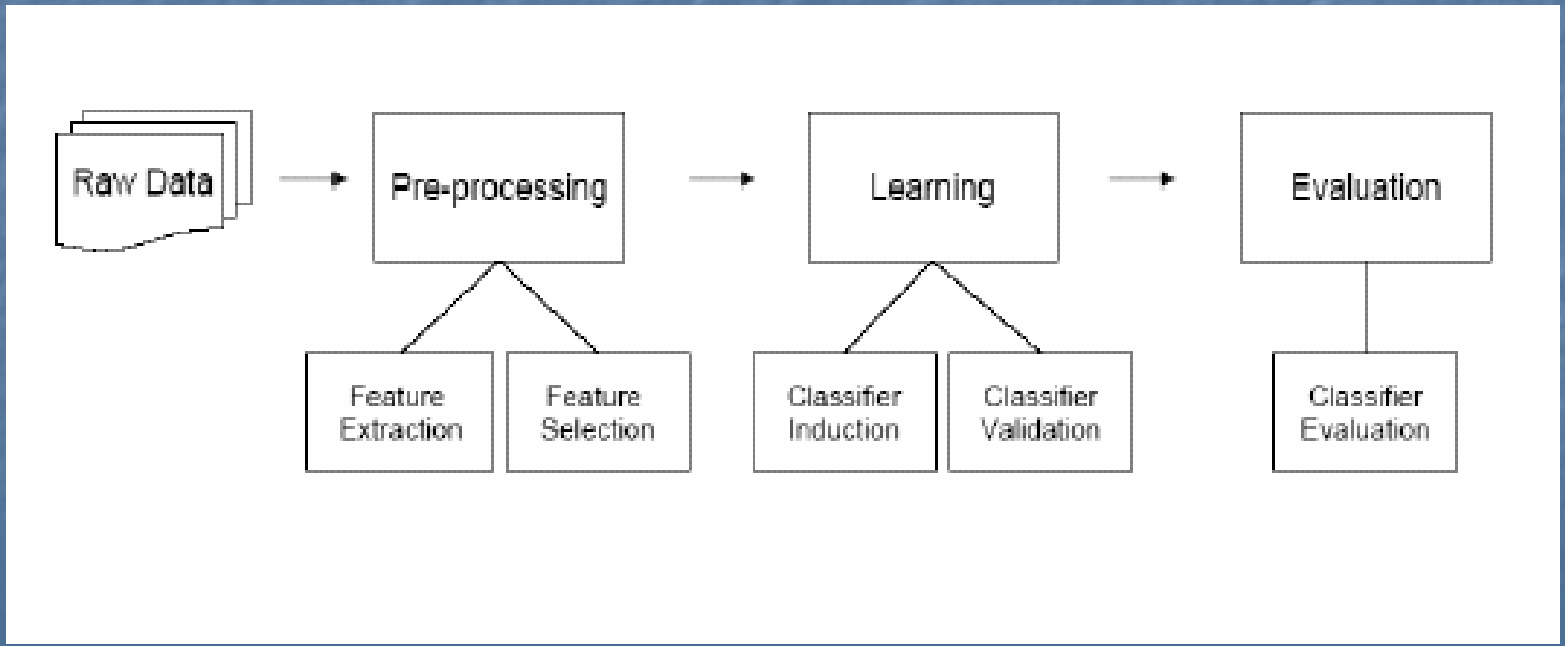
+ support of Marcin Pachocki

Institute of Computing Science
Poznań University of Technology

TPD students – Advanced Data Mining
Poznan, 2009/2010

A typical approach to supervised learning

- Construct data representation (objects x attributes) and label examples
- Possibly pre-process (feature construction)
- Learn from all labeled examples



Motivations

- Limited number of labeled examples; Unlabeled examples are easily available
- Labeling costly
- Examples:
  - Classification of Web pages, email filtering, text categorization.
- Aims
  - An efficient classifier with a minimal number of additional labeling

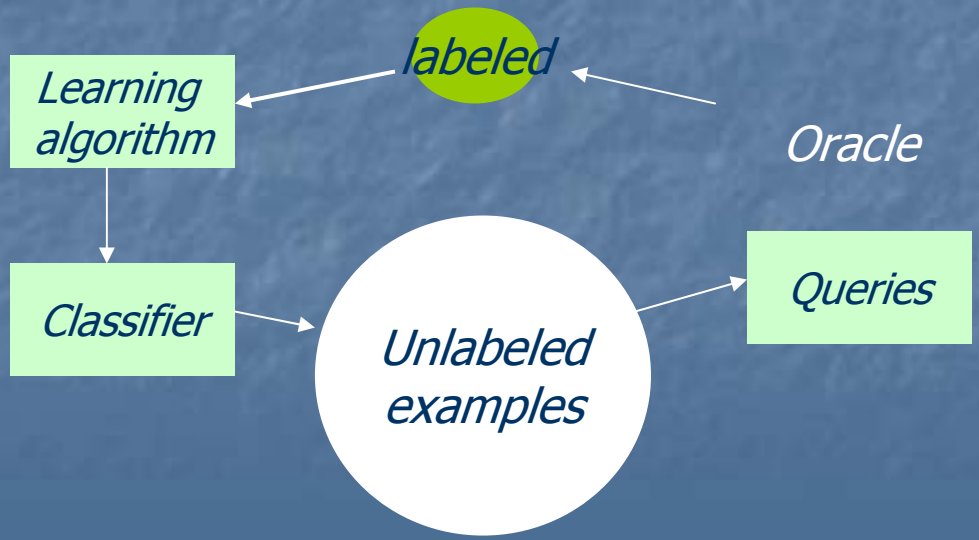
David Cohn, Les Atlas, Richard Ladner - *Improving Generalization with Active Learning*, Machine Learning, 1994.

Cele aktywnego uczenia (1)

- Osiągnięcie jak najwyższej trafności klasyfikowania przy jak najmniejszej liczbie przykładów uczących potrzebnych na starcie algorytmu.
- Minimalizacja odwołań do eksperta w celu pozyskania etykiety dla przykładów.
- Minimalizacja czasu konstrukcji złożonego klasyfikatora (komitetu) oraz czasu klasyfikacji nowych przykładów.

Active Learning

- Passive vs. Active Learning:
  - An algorithm controls input examples
- It is able to query (oracle / teacher) and receives a response (label) before outputting a final classifier
- How to select queries?



Active Learning Structure

- "wise learner is one which will classify easy cases by itself and reserve difficult cases for the teacher" [Turney]
- Who is an oracle?

Given:
Training dataset L of labelled examples
Oracle T
Source of unlabeled examples U
Stopping criteria M.

While stopping criteria not met

- Obtain query x for Teacher to label.
- Have the Teacher label the query $y = T(x)$.
- Add example to set of labelled examples $L = L \bigcup \{(x, y)\}$.
- Induce new classifier from training dataset $C_i = I(L)$.

Output a final classifier C

Figure 2: Generic Active Learning Structure

Do not ask too many questions!

- Query vs. random sampling

If only 1 in 1000 texts are class members and only 500 texts can be labelled, then a random sample will usually contain 500 negative examples and no positive.

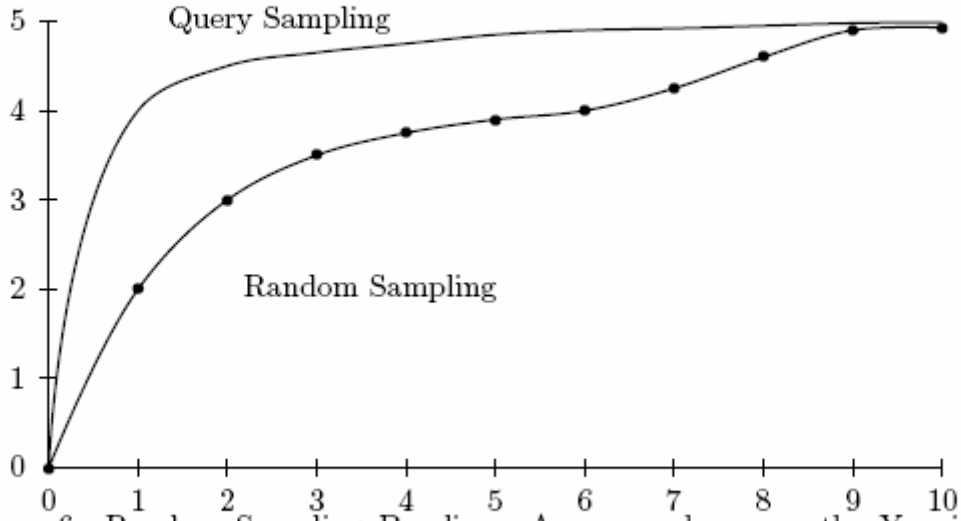


Figure 6: Random Sampling Baseline. Accuracy shown on the Y axis and increments of Active Learning are shown on X axis

Previous Research

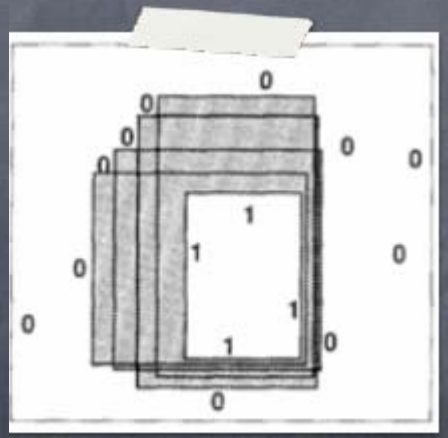
- Selective sampling [Cohn et al. 94]
- Uncertainty sampling [Lewis,Catlett 94]
- ...
- Ensembles
  - Query by Committee of Two [Sueng et al.; Freund et al. 97]
  - Sampling committees
  - Query by Committee [Abe, Mamitsuka 98]
  - QBC and Active Decorate [Melvile, Mooney 04]

Selective Sampling

- David Cohn, Les Atlas, Richard Ladner - *Improving Generalization with Active Learning*, Machine Learning, 1994.
- Bazuje na pojęciu regionu niepewności. Jest to wybrany podzbiór z dziedziny danych wejściowych, w którym spodziewamy się znaleźć interesujące przykłady.
- Zakłada znajomość rozkładu danych wejściowych (lub jego przybliżenia).

Selective Sampling - podejście naiwne

- Oparte o pojedynczą sieć neuronową. Stosowane jedynie dla binarnych problemów klasyfikacji.
- Dla znormalizowanej wartości na wyjściu sieci neuronowej z przedziału <0; 1> region niepewności można zdefiniować w następujący sposób:
  - "1" dla wartości na wyjściu z przedziału <0.9; 1>
  - "0" dla wartości na wyjściu z przedziału <0; 0.1>
  - niepewny dla wartości na wyjściu z przedziału (0.1; 0.9)
- Niepewność jest mierzona tylko w obrębie danej konfiguracji sieci, a nie we wszystkich możliwych konfiguracjach danej architektury sieci.



Query by Committee

- H. S. Seung, Manfred Opper, Haim Sompolinsky. *Query by committee*. 1992.
- Po raz pierwszy pojawia się pojęcie komitetu złożonego z co najmniej dwóch klasyfikatorów.
- Algorytm uczący traktowany był jako algorytm działający *on-line* i analizowano w nim spadek błędu klasyfikowania w funkcji liczby zapytań do eksperta.
- Yoav Freund, H. S. Seung, Eli Shamir, Naftali Tishby. *Selective sampling using the query by committee algorithm*. 1997.

Query by Committee

L - set of labeled examples
U - set of unlabeled examples
A - base learning agorithm
k - number of act iterations
m - size of each sample
Repeat *k* times

1. Generate a committee of classifiers *C* = EnsembleMethod(A, L)*
2. For each *x* in *U* compute Info_val*(C*,x)*, based on the current committee
3. Select a subset *S* of *m* examples with max Info_val
4. Obtain Labels from Oracle for examples in *S*
5. Remove examples in *S* from *U* and add to *L*

Return *Ensemble*

Active vs. Passive Learning

	Passive	Active
Number of training examples	Large	Relatively Small
Number of classifiers induced	One (Batch)	Many (Iterative)
Choice over training examples	None	Some
Stopping criteria	Simple	Complex

Table 3: A simple, high level comparison between passive and active learning

Research Questions

- Algorithms for constructing committees
- Selection of examples to queries
  - Disagreement measures
  - How many examples to query
- Influence of creating the starting labeled set
- An experimental evaluation of different approaches

Constructing Committees

- Considered approaches
  - Bagging [Abe,Mamitsuka 98]
  - Boosting [Abe,Mamitsuka 98]
  - Decorate [Melville,Mooney 03]
    $\rightarrow$ Active_Decorate [MM 04]
    - Iterative and adaptive multiple classifier
    - Additional artificial training examples to get more diversified component classifiers
  - Random Forests

Zastosowane podejście (2): algorytm *Decorate*

Given:

T - set of m learning examples $<(x_1, y_1), \ldots, (x_n, y_n)>$ with labels y_i

C_{size} - desired ensemble size

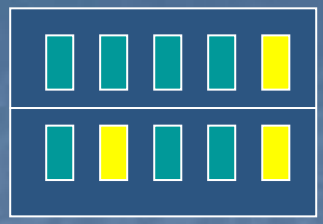
I_{max} - maximum number of iterations to build an ensemble

R_{size} - factor that determines number of artificial examples to generate

A - base learning agorithm

1. $i = 1$
2. $trials = 1$
3. $C_i = A(T)$
4. Initialize ensemble, $C^* = \{C_i\}$
5. Compute ensemble error, e
6. While $i < C_{size}$ and $trials < I_{max}$
7. Generate $R_{size} \times |T|$ training examples, R, based on distribution of training data
8. Label examples in R with probability of class labels inversely proportional to C^*s predictions
9. $T = T \cup R$, add the artificial data
10. $C' = A(T)$
11. $C^* = C^* \cup \{C'\}$
12. $T = T - R$, remove the artificial data
13. Compute training error, e', of C^* as in step 5
14. if $e' < e$
15. $i = i + 1$
16. $e = e'$
17. else
18. $C^* = C^* - \{C'\}$
19. $trials = trials + 1$

Disagreement measures



Analysing predictions of base classifiers:

- Margins of the classified example
  - Difference between number of votes in the committee for the most and the second predicted class
- Probability vectors instead of single predictions
  - Generalization of margins – difference between probabilities

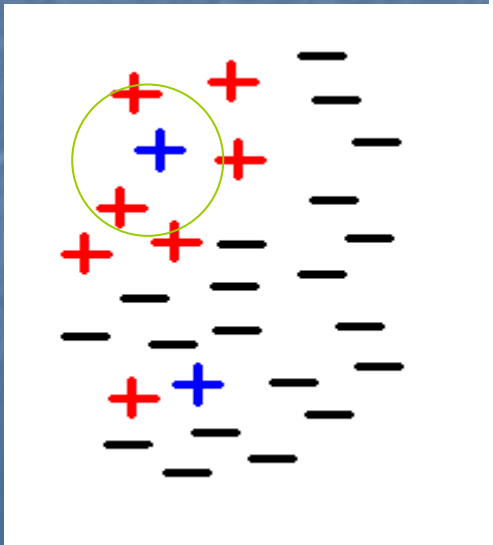
$$P_y(x) = \frac{\sum_{C_i \in C^*} P_{C_i,y}(x)}{size(C^*)}$$

  - Distance between component classifiers and final answer
  - Median distance
- Jensen Shannon divergence

$$JS(P_1, P_2, \ldots, P_n) = H(\sum_{i=1}^{n} w_i P_i) - \sum_{i=1}^{n} w_i H(P_i)$$

Constructing the starting training set?

- Influence of choosing the set L (starting in AL)
- In controlled experiments
  - Random sample vs. focused selection
- Edited k-NN [Stefanowski,Wilk 07]
  - Use Wilson's edited nearest neighbor rule:
  - Compare example's label with its neighbors,
  - Safe $\rightarrow$ correctly classified by its k nearest neighbors,
  - Choosing the most safe examples from given classes

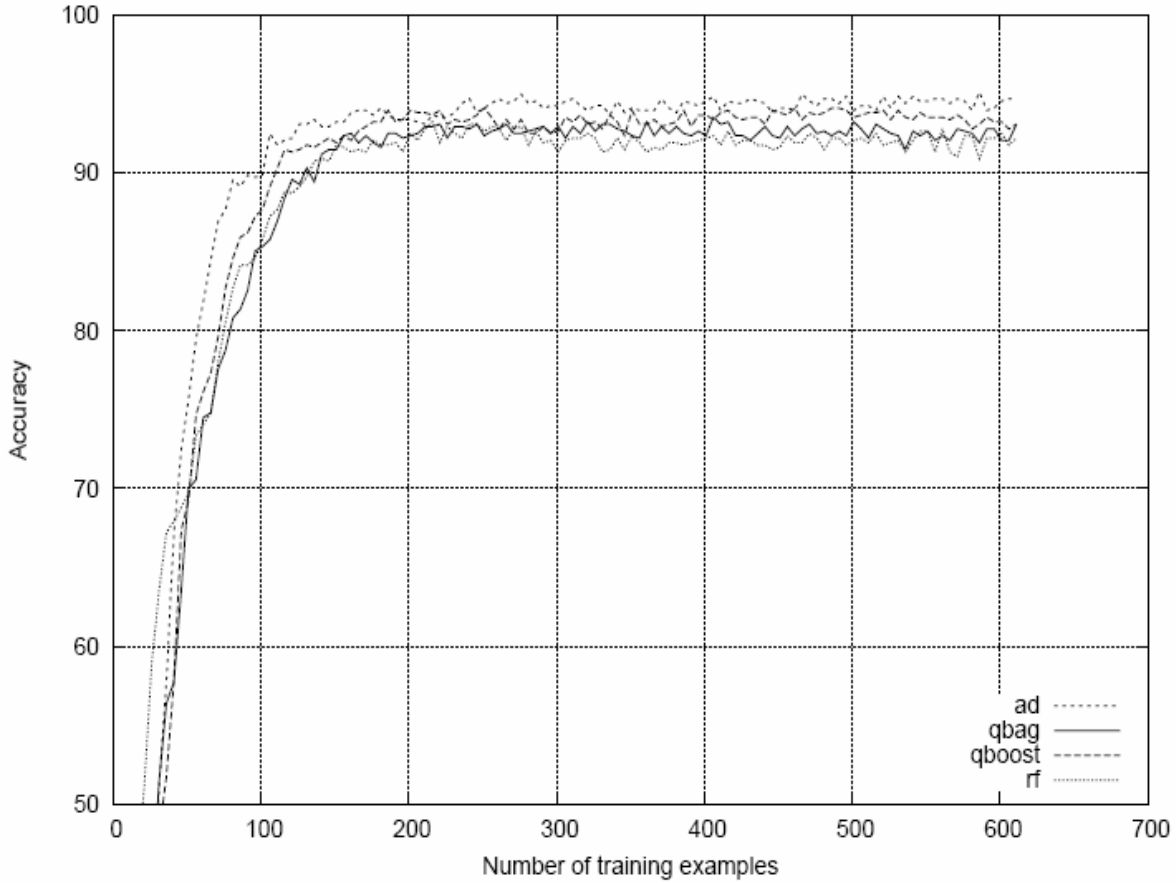


Usefullness of differenct QBC

- An experimental comparative study
  - Performance of different approaches to QBC
  - Using 4 disagreement measures
  - Selecting the starting training set
    - Random vs. edited k-NN
    - Choosing single or more examples to query
  - 6 benchmark data sets (UCI)
  - Learning curves
    - Performance passive vs. active learners
  - Implementations in WEKA

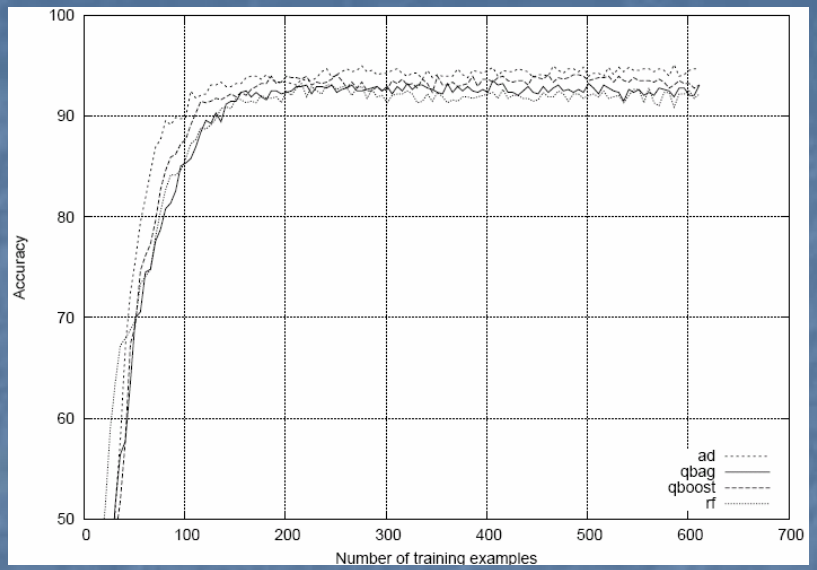
Different committees in AL

- Comparing different active learners on Soybean data

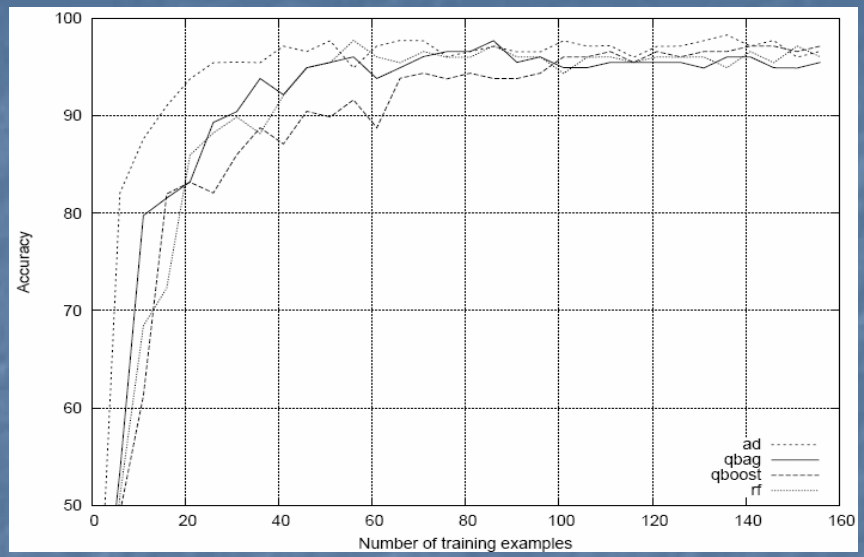


Different committees in AL

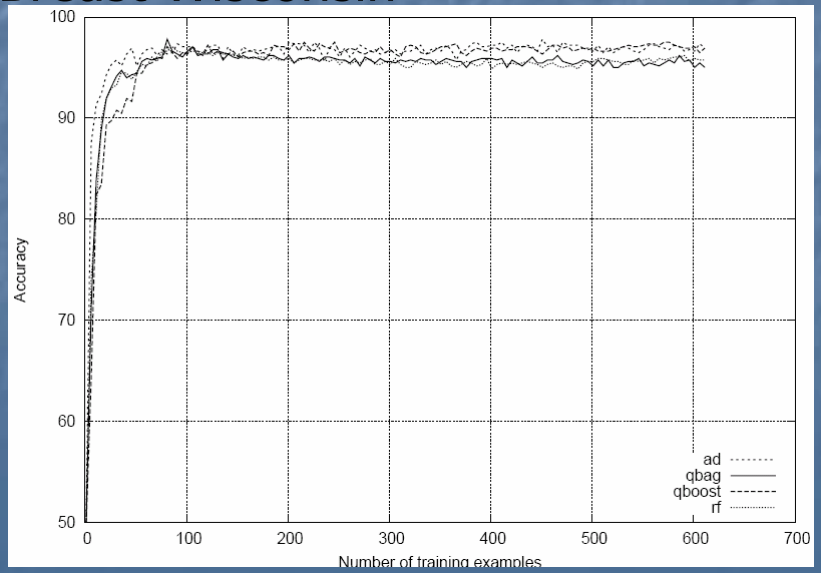
Soybean



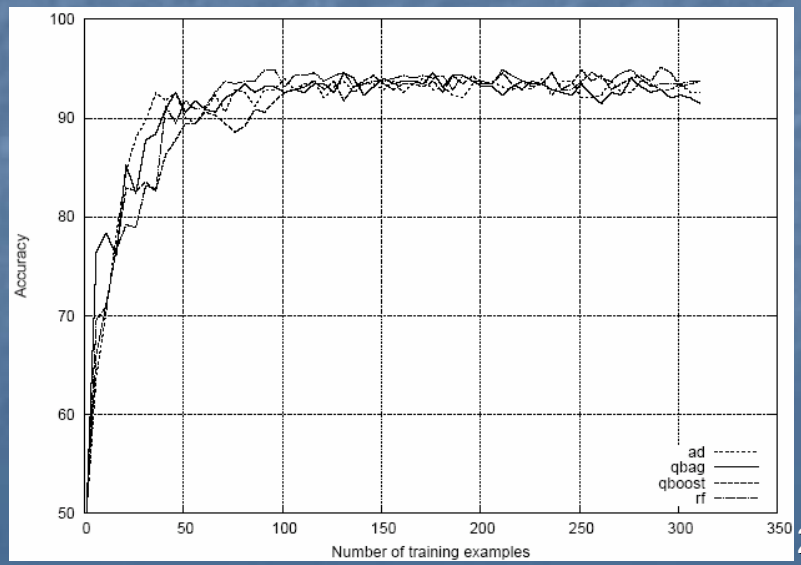
Wine



Breast Wisconsin

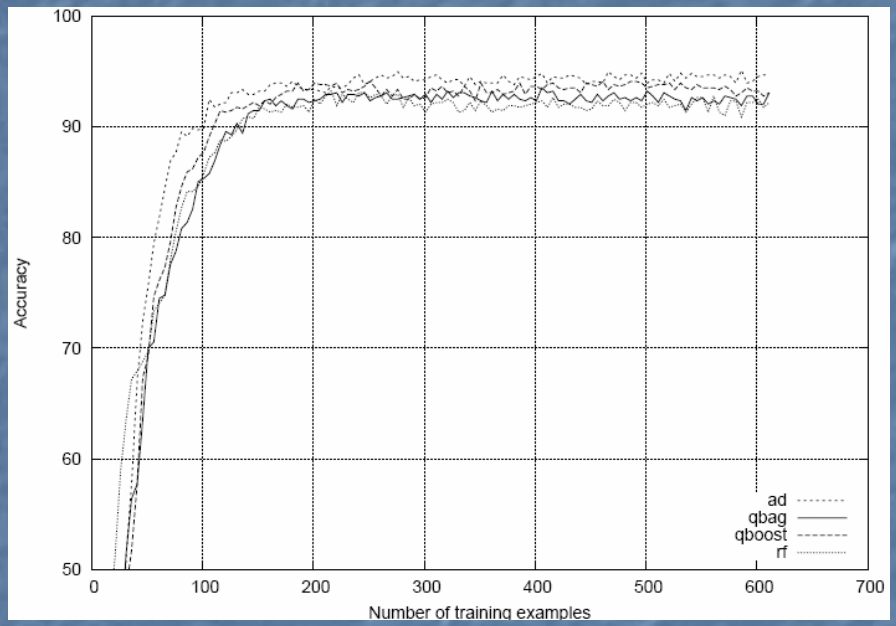


Ionosphere

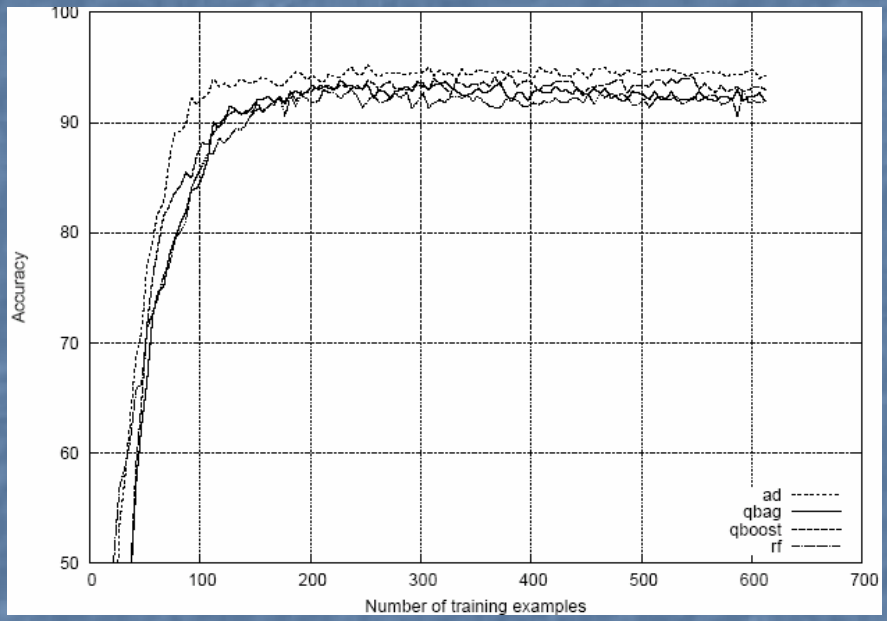


Selecting the starting set in QBC

Soybean random selection



edited k-NN



Reduction of labeled examples Active vs Passive

TABLE 2: Reduction of the number of training examples to achieve the target accuracy - for Random forests results are presented for 50 trees and for 15.

Approach	wine	ionosphere	breast	soybean	diabetes	credit-g
Decorate	37 (0.23)	56 (0.18)	35 (0.06)	347 (0.57)	172 (0.25)	174 (0.19)
AD	27 (0.17)	36 (0.11)	37 (0.06)	125 (0.20)	410 (0.59)	267 (0.30)
AD k-nn	**22 (0.14)**	**32 (0.10)**	**30 (0.05)**	**105 (0.17)**	**48 (0.07)**	**161 (0.18)**
Bagging	122 (0.76)	206 (0.65)	388 (0.63)	373 (0.61)	126 (0.18)	297 (0.33)
QBag	52 (0.33)	**46 (0.15)**	53 (0.09)	144 (0.23)	**98 (0.14)**	**215 (0.24)**
QBag k-nn	**40 (0.25)**	48 (0.15)	**53 (0.09)**	**126 (0.21)**	118 (0.17)	293 (0.33)
Boosting	111 (0.69)	162 (0.51)	60 (0.10)	269 (0.44)	201 (0.29)	251 (0.28)
QBoost	**75 (0.47)**	**102 (0.32)**	62 (0.10)	149 (0.24)	199 (0.29)	203 (0.23)
Qboost k-nn	103 (0.64)	130 (0.41)	**54 (0.09)**	**145 (0.24)**	**38 (0.05)**	**170 (0.19)**
RF (50)	158 (0.99)	167 (0.53)	105 (0.17)	176 (0.25)	418 (0.68)	479 (0.53)
ARF (50)	**56 (0.35)**	71 (0.23)	56 (0.09)	76 (0.11)	**151 (0.25)**	272 (0.30)
ARF k-nn	128 (0.80)	**67 (0.21)**	**48 (0.08)**	**76 (0.11)**	152 (0.25)	**205 (0.22)**
RF (15)	115 (0.72)	251 (0.80)	105 (0.17)	411 (0.67)	176 (0.25)	377 (0.42)
ARF (15)	**58 (0.36)**	118 (0.37)	65 (0.11)	212 (0.35)	139 (0.20)	**170 (0.19)**
ARF k-nn	63 (0.39)	**86 (0.27)**	**48 (0.08)**	**190 (0.31)**	**111 (0.16)**	250 (0.28)

Computational costs

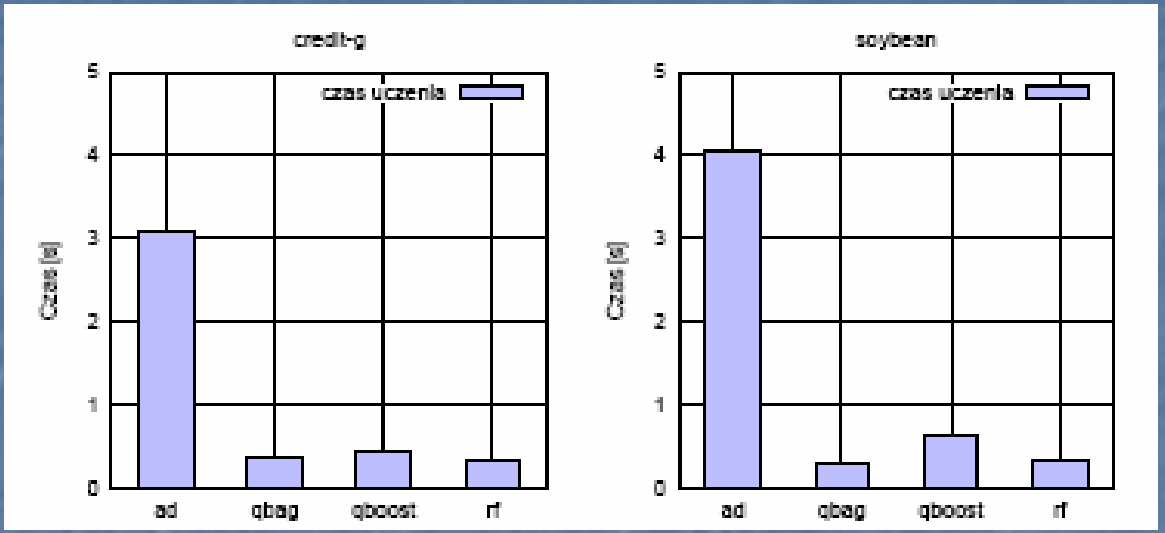
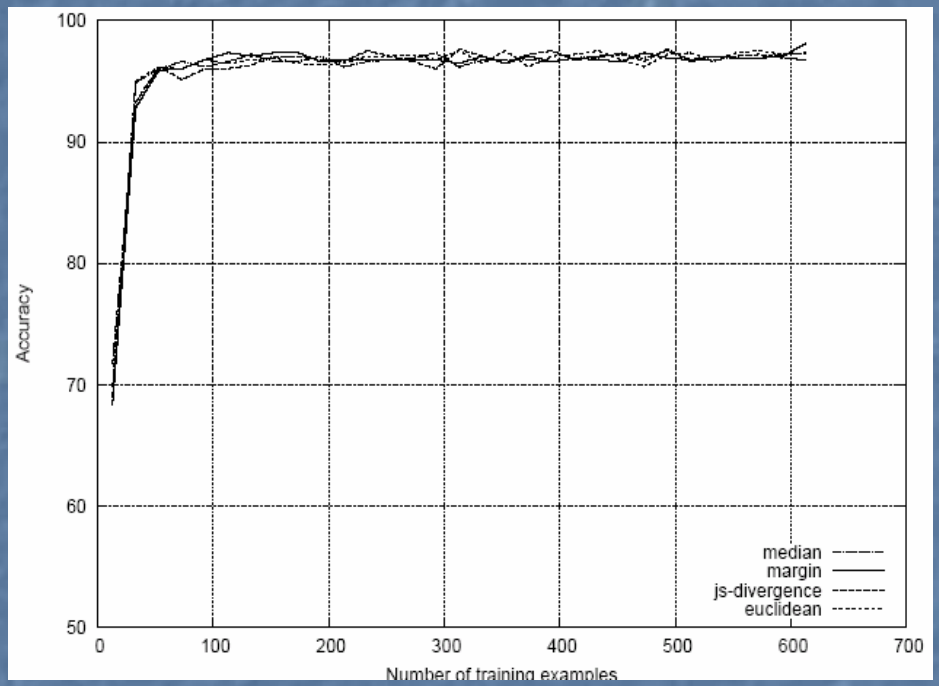


TABLE 3: Committee training time (in seconds) of different active learners.

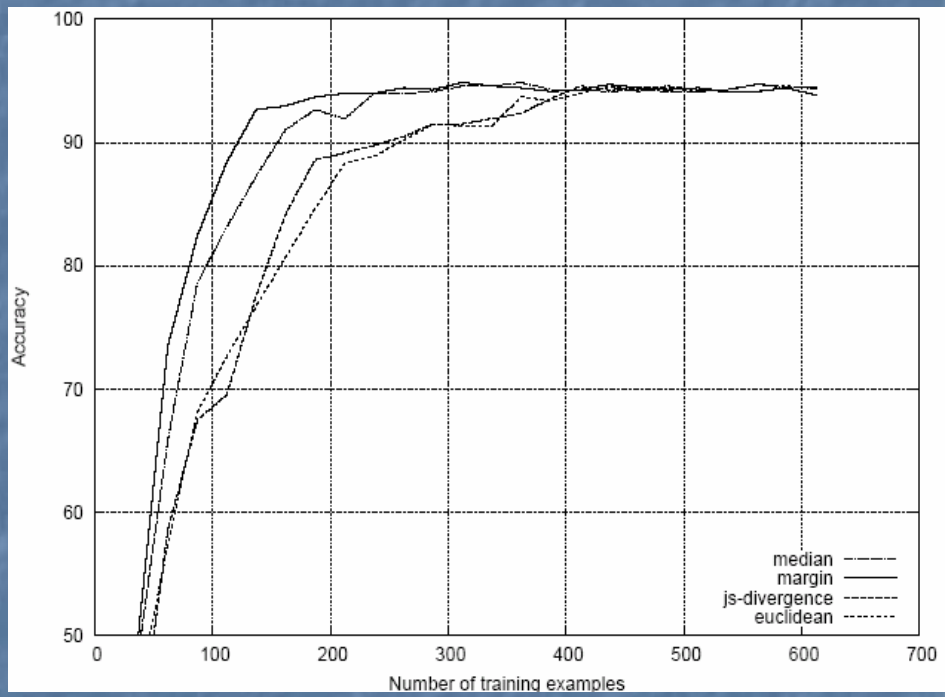
Approach	wine	ionosphere	breast	soybean	diabetes	credit-g
Active Decorate	0.42	1.8	0.7	4.0	3.1	3.08
Query by Bagging	0.04	0.2	0.01	0.28	0.4	0.38
Query by Boosting	0.07	0.36	0.11	0.63	0.14	0.44
Random Forests	0.02	0.1	0.01	0.33	0.19	0.33

Comparing disagreement measures

Breast



Soybean



Conclusions

- QBC in AL → accuracy comparable to passive versions with much smaller number of examples
- The best reduction ratio → Active Decorate (4 of 6 data)
- Trade off with computational costs
  - Active Decorate ~ 10 times more
  - Random Forests → the fastest
- Selection of the training set → edited k-NN improves all approaches
  - The best reduction ratio → Active Decorate
- Increasing the number of added queries → not too much
- Choice of disagreement measures
  - No big influence, except multi-class data (Soybean)
  - Generalized margins → JS divergence

Powiązane zagadnienia

Co-Training

- Avrim Blum, Tom Mitchell - *Combining Labeled and Unlabeled Data with Co-Training*.
- Wykorzystanie przykładów bez etykiety do polepszenia jakości klasyfikatorów otrzymanych na małym zbiorze przykładów uczących.
- Ekstrakcja *n* różnych podzbiorów cech z danego problemu klasyfikacji (widoków).
- Uczenie *n* klasyfikatorów na poszczególnych podzbiorach cech.
- Wykorzystanie klasyfikatorów do klasyfikacji przykładów bez etykiety i dodawania ich do aktualnego zbioru uczącego.

Co-training and Decomposition

- Co-training is a method which can be applied to machine learning problems with multiple views.
- By this we mean that the problem has a natural way in which to divide their features into subsets which we call views.
- There is sufficient redundant information in the description of the examples that a number of distinct sets of features can be formed - each of which is sufficient for describing the target function.
- Blum and Mitchell 98: 2 views for classifying Web pages $\rightarrow$ the body of the page and the anchor text of the links that pointed to the web page.
- Kititchenko and Matwin [KM01] an application to classify e-mail $\rightarrow$ the body and subject of the email

Algorytm Co-Training

Given:

L - set of labeled training examples

U - set of unlabeled examples

Create a pool of U' examples by choosing u examples at random from U

Loop for k iterations:

1. Use L to train classifier h_1 that considers only the x_1 portion of x
2. Use L to train classifier h_2 that considers only the x_2 portion of x
3. Allow h_1 to label p positive examples and n negative examples from U'
4. Allow h_2 to label p positive examples and n negative examples from U'
5. Add these self-labeled examples to L
6. Randomly choose $2p + 2n$ examples from U to replenish U'

References

- Naoki Abe and Hiroshi Mamitsuka (1998), Query learning strategies using boosting and bagging, in Proceedings of the Fifteenth International Conference on Machine Learning'98, 1--9.
- A. Blum, T. Mitchell (1998), Combining labeled and unlabeled data with co-training, in Proceedings of the Workshop on Computational Learning Theory
- D. Cohen, L. Atlas, R. Ladner (1994),Improving generalization with active learning. Machine Learning, 15(2), 201–-221.
- Michael Davy (2005), A Review of Active Learning and Co-Training in Text Classification. Dep. of Computer Science, Trinity College Dublin, Research Report, TCD-CS-2005-64, 39 pp.
- Kiritchenko and Matwin(2001), Email classification withco-training, in Proceedings of the CASCON '01 Conference
- Melvile and Monney(2004), Diverse ensembles for active learning, In Proceedings of the 21st Int. Conference on Machine Learning, 584--591.
- J.Stefanowski, M.Pachocki (2009), Comparing Performance of Committee based Approaches to Active Learning. In: M.Kłopotek, A.Przepiórkowski, S.Wierzchoń, K.Trojanowski (red.) Recent Advances in Intelligent Information Systems, Wydawnictwo EXIT, Warszawa, 2009, 457-470.

Thank you for your attention

Questions and remarks, please!

Contact, remarks:
Jerzy.Stefanowski@cs.put.poznan.pl