CORPORATE TECHNOLOGY

The Bayesian Committee Machine

Volker Tresp
Siemens AG

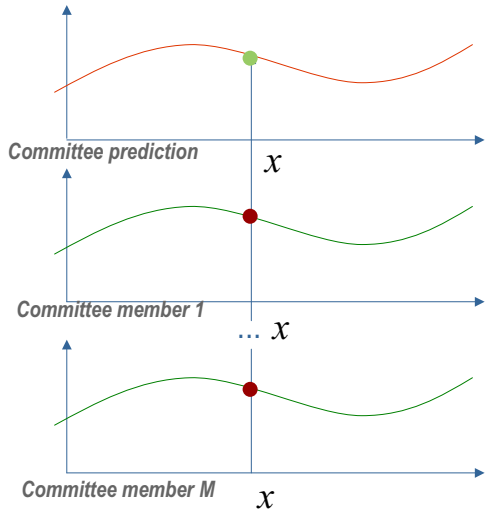
with: Anton Schwaighofer
Univ. of Graz, Siemens AG

SIEMENS

Workshop: Ensemble Methods 2004

Motivation

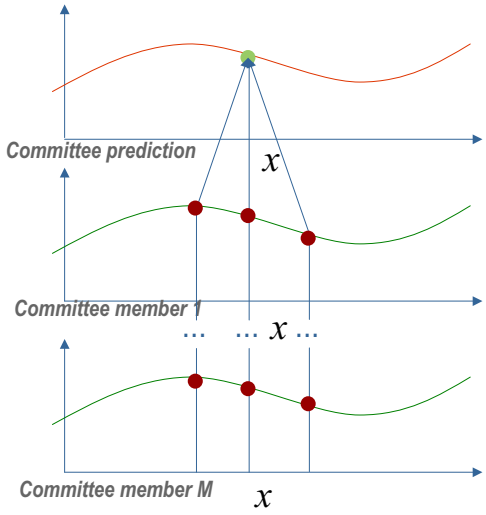
If I query a committee for the prediction at input x, only the predictions of each committee member at the same x is used



Workshop: Ensemble Methods 2004

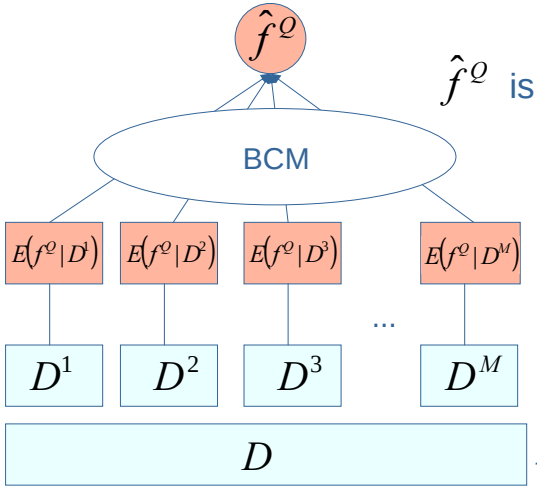
BCM

It turns out that the predictions of the committee members at other inputs can improve the prediction of the committee at x!



Workshop: Ensemble Methods 2004

BCM Architecture



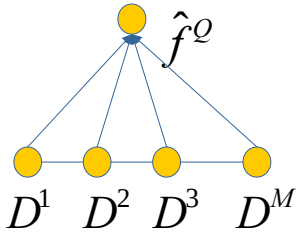
$\hat{f}^Q$ is Q-dimensional

Tresp, V. (2000). Neural Computation, Vol. 12.

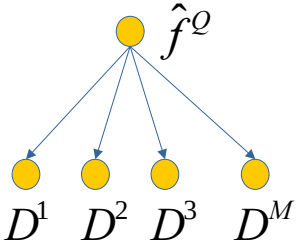
BCM: Intuition

- The prediction of the committee should exploit the information contained in all data
- The prediction and the variance of the prediction at a given x does not represent the information a committee member has learned about the data
- (In a Gaussian approximation) if Q are the degrees of freedom of a model, the information is contained in Q model parameter estimates and their posterior covariance
- The same information is contained in Q model predictions and their covariance structure!

BCM as Naive Bayes

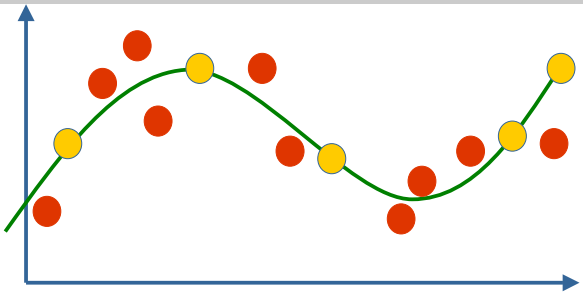


$$P(f^Q)\prod P(D^i \mid f^Q)$$



$$P(f^Q \mid D) \approx \frac{\prod_{i=1}^{M} P(f^Q \mid D^i)}{P(f^Q)^{M-1}}$$

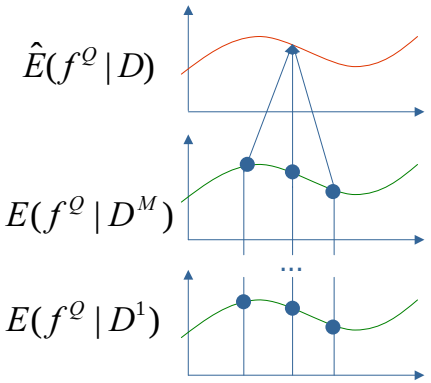
BCM: Visualization



If I select a sufficient number of test points ●
then those determine the function (green)
and make the data ● independent

BCM: Gaussian Approximation

$$\hat{E}(f^Q \mid D) = C\sum_{i=1}^{M} \mathrm{cov}(f^Q \mid D^i)^{-1} E(f^Q \mid D^i)$$

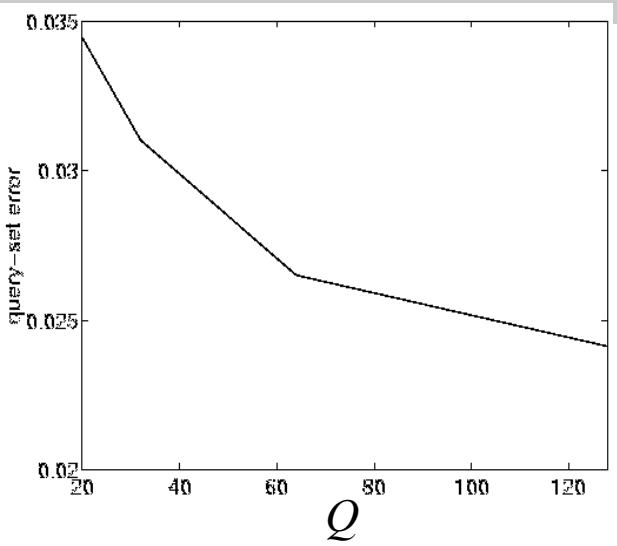


BCM for GPR

- The computational complexity of Gauss process regression (~kriging, smooting splines) scales as N^3
- The BCM version of GPR scales as K^2N (where K = N/M = fixed)
- In a benchmark test, BCM-GPR gave best results of all linearly scaling GPR approximations (Nystroem, SRM)

Workshop: Ensemble Methods 2004

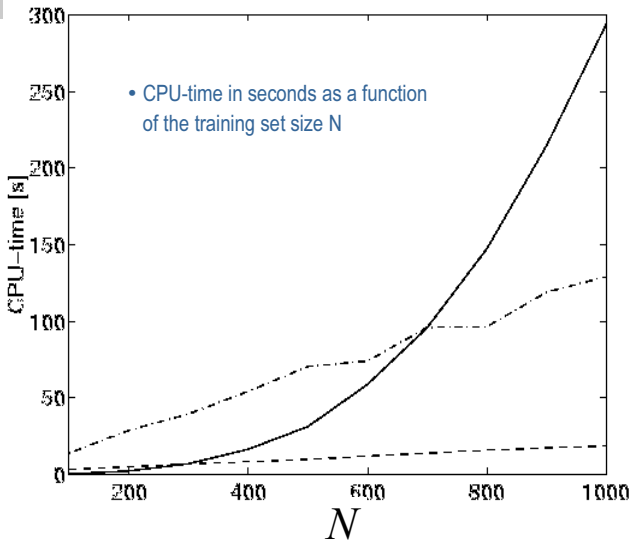
Test set error as a function of Q



Workshop: Ensemble Methods 2004

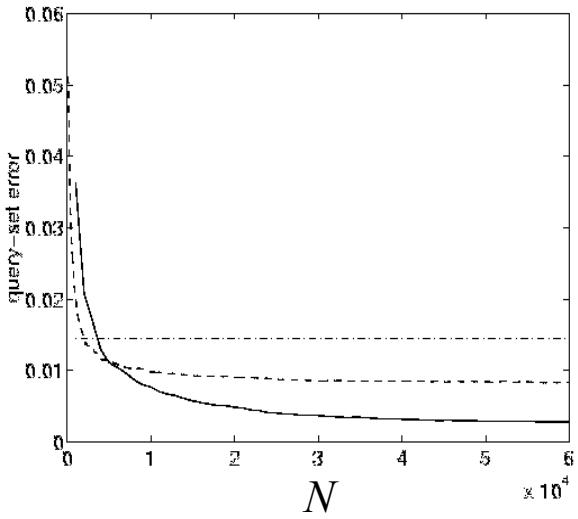
CPU-time as a function of N

- CPU-time in seconds as a function of the training set size N



Workshop: Ensemble Methods 2004

Test set error as a function of N



dule size, query-set size)

	BUPA	DIABETES	WAVEFORM	HOUSING
ze)	6	8	21	13
	200	600	600	400
	$0.828 + 0.028$	$0.681 + 0.032$	$0.379 + 0.019$	$0.1075 + 0.0065$
50)	$(-2.6 + 1.8) \times 10^{-1}$	$0.0095 + 0.0030$	$0.0566 + 0.0243$	$0.338 + 0.132$
50)	$(1.6 + 1.8) \times 10^{-1}$	$-0.0027 + 0.0023$	$0.0295 + 0.0213$	$0.117 + 0.038$
00)	$(0.1 + 0.2) \times 10^{-1}$	$-0.0001 + 0.0015$	$0.0037 + 0.0095$	$0.196 + 0.053$
	$(0.0 + 0.1) \times 10^{-1}$	$-0.0011 + 0.0010$	$0.0163 + 0.0086$	$0.1138 + 0.0568$
mods.	$0.299 + 0.07$	$0.2476 + 0.0236$	$0.639 + 0.0624$	$2.85 + 0.22$
ging	$0.0676 + 0.0094$	$0.1426 + 0.0193$	$0.4914 + 0.0576$	$1.541 + 0.157$
	159	600	600	400
	16	43	223	138

Real World Data

CORPORATE TECHNOLOGY

Hierarchical User Profiling with Nonparametric Bayesian Models

(with Applications to Information Filtering)

with Kai Yu
Siemens

SIEMENS

Workshop: Ensemble Methods 2004

Content-based Filtering: Exploring Similarity between Items

book 1: like
book 2: like
book 3: ?

feature 1: fashion:y/n
feature 2: love stories:y/n
feature 3: action:y/n

Does Belle like book 3?

Predictive Models:
SVM, NN, GLM, …,

$$P(y \mid x, w)$$
$$P(y \mid x, D)$$

Workshop: Ensemble Methods 2004

Collaborative Filtering: Exploring Similarity between People

Book 1:	dislike	like	like
Book 2:	like	like	like
Book 3:	dislike	?	like

It seems that Belle agrees more with Beast!

CF can work even without features!

The Problems

Content-based Filtering: small-sample problem, deficiency of descriptive content features, …

Collaborative Filtering: sparsity of user annotations, no means to handle new items, …

Hybrid Filtering: combines the strengths of both and overcomes the weakness of each

How should we combine content filtering with collaborative filtering?

Workshop: Ensemble Methods 2004

Our Work

- Insight into the problem
- Principled solution to hybrid filtering
- Unifying framework for information filtering

Workshop: Ensemble Methods 2004

Hierarchical Bayesian Modeling

- All users are the same:

  use one model trained on all data
- Each user is a complete individual:

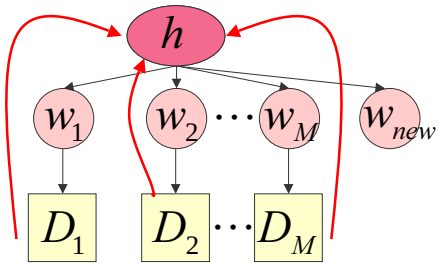
  train a separate model for each user
- Learn from one another:

  each user has her/his own model generated from a common prior distribution, which is learned from data and shared between user models.

Workshop: Ensemble Methods 2004

Hierarchical Bayesian Modeling

$$P(h)\prod_{j=1}^{M} P(w_j \mid h)P(D_j \mid w_j)$$



Hierarchical Bayes

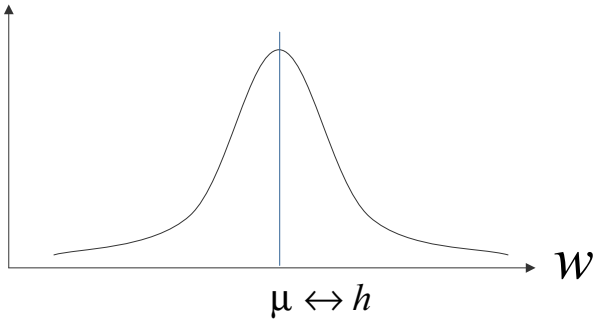
- Essentially, we learn the ***informative prior*** $P(w \mid D)$
- Two problems with Hierarchical Bayes:
  - Solution has to be approximated (e.g. Monte Carlo sampling)
  - The functional form of $P(w \mid h)$ must not only be able to encode my prior belief about w but must also be flexible enough to represent the ***learned prior distribution***

$$P(w \mid D_1,...,D_M) \underset{M \to \infty}{\rightarrow} P(w \mid h_{ML})$$

Workshop: Ensemble Methods 2004

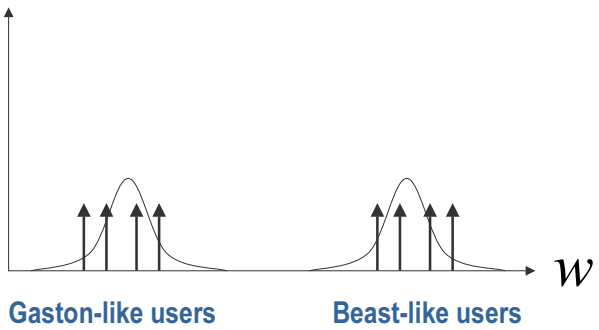
Prior Belief

$P(w \mid h)$



Workshop: Ensemble Methods 2004

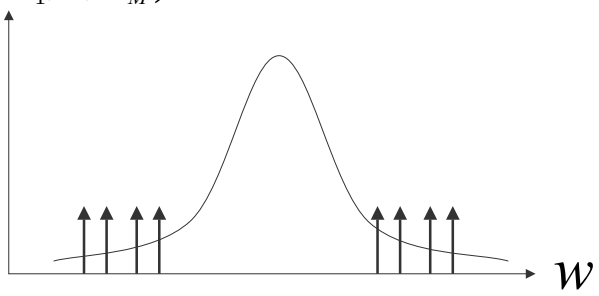
Actual Data Distribution



Workshop: Ensemble Methods 2004

HB learns …

$P(w \mid D_1,...,D_M)$



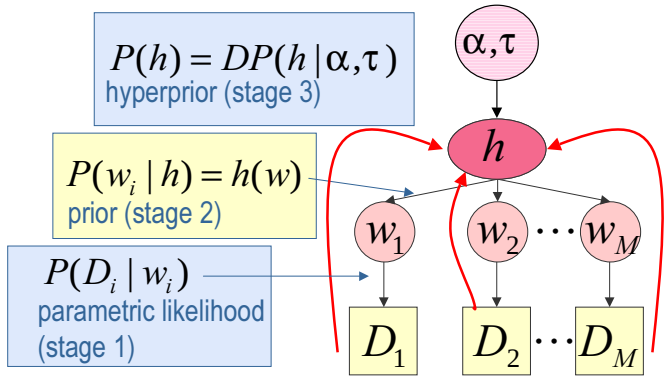
… nothing

Idea

If the world is complicated, why should we limit the form of distribution?

Workshop: Ensemble Methods 2004

Solution: Dirichlet Enhancement (Nonparametric Bayesian Modeling)



$\alpha(w)$ base distribution (subjective prior): $P(w) = \alpha(w)$

Workshop: Ensemble Methods 2004

Mixture of Dirichlet Processes (MDP)

- A posteriori:

$$P(w_i \mid D_i, \{w_j\}_{j \neq i})$$

$$\propto \tau P(D_i) P(w_i \mid D_i) + \sum_{j \neq i} P(D_i \mid w_j) \delta(w_i - w_j)$$

Suitable for Gibbs sampling!

(M. Escobar, M. West, P. Mueller, T. Ferguson, R. Neal)

Workshop: Ensemble Methods 2004

EM-Learning

- We use the data to estimate the *mode* of the hyperparameters, where the model parameters w take the role of latent variables

$$h_{\text{est}} = \arg\max_h P(h \mid D, \alpha, \tau)$$

where $D = \{D_1, \ldots D_M\}$

-- E-step $$P(w \mid D_j, h_{est}^{(t)}) = \frac{P(D_j \mid w) P(w \mid h_{est}^{(t)})}{\int P(D_j \mid w) P(w \mid h_{est}^{(t)}) dw}$$

-- M-step $$h_{\text{est}}^{(t+1)}(w) = \frac{1}{\tau + M} \left(\tau \alpha(w) + \sum_{j=1}^{M} P(w \mid D_j, h_{est}^{(t)}) \right)$$

EM-Learning – Bad News

In general, EM-learning is not suitable since $h_{\text{est}}(w)$ can not be parameterized (due to explosion of terms). Thus we have to adopt approximations

- One-step EM (UAI, 2003)
- Variational approximation (Today)

Variational Approximation: Motivation

$$h_{\text{est}}^{(t+1)}(w) = \frac{1}{\tau + M}\left(\tau\alpha(w) + \sum_{j=1}^{M} P\left(w \mid D_j, h_{est}^{(t)}\right)\right)$$

$$P\left(w \mid D_j, h_{\text{est}}^{(t)}\right) \propto P(D_j \mid w) P(w \mid h_{\text{est}}^{(t)})$$

$$P(w \mid D_j, h_{\text{est}}^{\infty}) \propto \frac{P(D_j \mid w)}{P(D_j \mid w_j^{ML}) - P(D_j \mid w)}$$

Variational Approximation (2)

Thus one might want to approximate (*even with overlapping*):

$$P(w \mid D_j, h_{est}^{\infty}) = \sum_{i=1}^{M} \xi_{i,j}\, \delta(w - w_i^{ML})$$

where $1 = \sum_{i=1}^{M} \xi_{i,j}$

Variational EM

E-step: $\xi_{i,j}^{(t)} \propto P(w_i^{ML} \mid D_j, h_{est}^{(t)})$

$$\xi_i^{(t)} = \sum_{j=1}^{M} \xi_{i,j}^{(t)}$$

M-step: $$h_{\text{est}}^{(t+1)}(w) = \frac{1}{\tau + M}\left(\tau\alpha(w) + \sum_{i=1}^{M} \xi_i^{(t)} \delta(w - w_i^{ML})\right)$$

$$h_{\text{est}}(w) = \frac{1}{\tau + M}\left(\tau\alpha(w) + \sum_{i=1}^{M} \xi_i\, \delta(w - w_i^{ML})\right)$$

Predictions: a committee machine!



Treating the learned $h_{est}(w)$ as fixed parameters, we can now make predictions for the active user with D_a:

$$P(y \mid x, D_a, h_{\text{est}}) = \frac{1}{Z}\left(\tau P(D_a \mid \alpha) P(y \mid x, D_a) + \sum_{i=1}^{M} \xi_i P(D_a \mid w_i^{ML}) P(y \mid x, w_i^{ML})\right)$$

How well does w_i^{ML} explain the active user's data

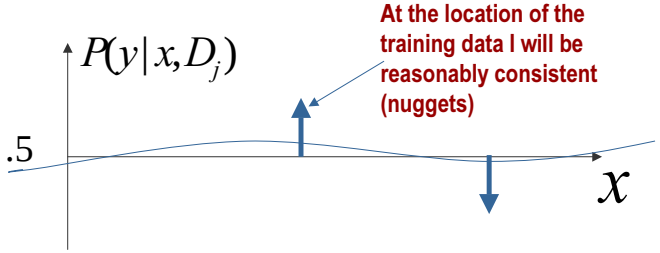
Predictive distribution given w_i^{ML}

Connections to Content-based and Collaborative filters

- If the precision parameter goes to infinity, the learned prior remains the same as the base distribution. This case is equivalent to *pure content-based filtering*
- If the precision parameter goes to zero, the learned prior assigns zero chances on new values of parameters. Thus predictions for a new user (i.e. the active user) completely depend on previous users' ML models.
- If the precision parameter goes to zero and features are deficient or not available, then *pure collaborative filtering* is derived.

Nugget Effect

- So far does our approach does not give valid predictions if features are useless – in contrast to collaborative filtering ->
- *Nugget effect*



Nugget Effect: pure collaborative filtering

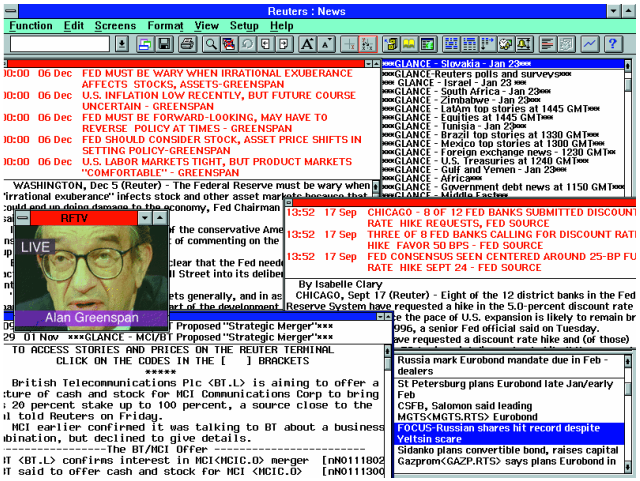
- In case of pure collaborative filtering, the logarithm of the weighting term takes on an intuitive form

$$\log P(D_a \mid D_j) = const + c_1 \#^{+} - c_2 \#^{-}$$

c_1, c_2 : Positive constants

$\#^{+}, \#^{-}$: Number of agreements, resp. disagreements

Text Retrieval on Reuters-21578 Data



Workshop: Ensemble Methods 2004

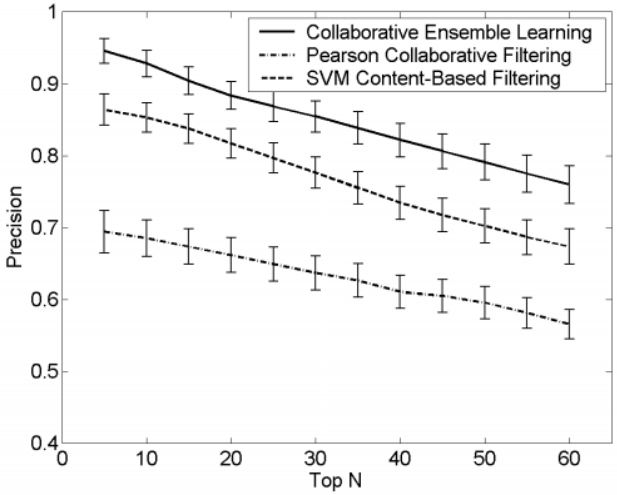
Text Retrieval on Reuters-21578 Data

- We used 36 categories covering totally 10,034 news articles from the Reuters text data set (1,152 articles belong to more than one category)
- For the experiments, we assume that each user is interested in exactly one of these categories
- We generate example data for L=360 (artificial) users by choosing at random a set of 30 (positive and negative) example items.
- The goal is to predict the probability that a user likes an article
- Rank all unseen articles and select top N ranked articles;
- Plot: how many of the top N articles are truly positive

- (features: word counts, classifier: probabilistic SVM with linear kernel)

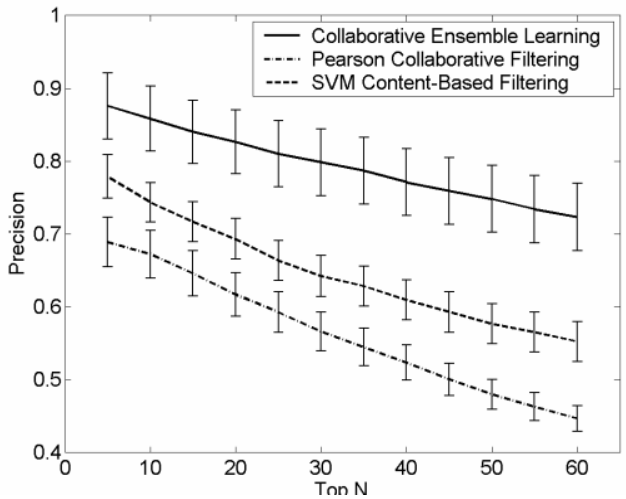
Workshop: Ensemble Methods 2004

Active user: given 20 known



Workshop: Ensemble Methods 2004

Active user: given 5 known



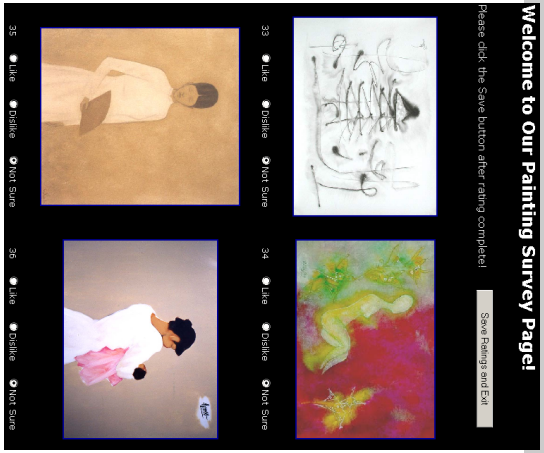


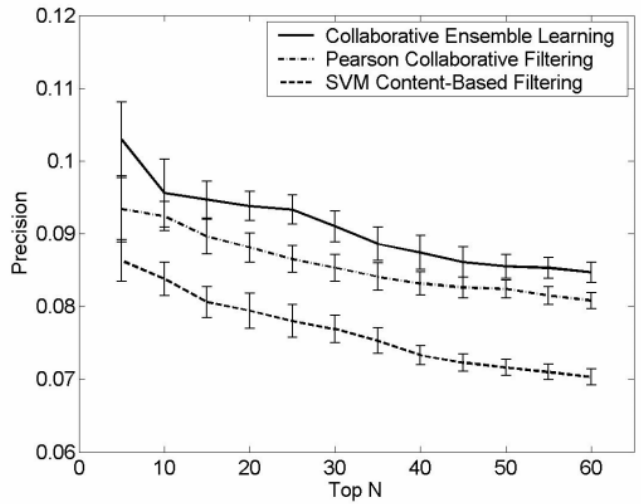
Image Retrieval

Image Retrieval

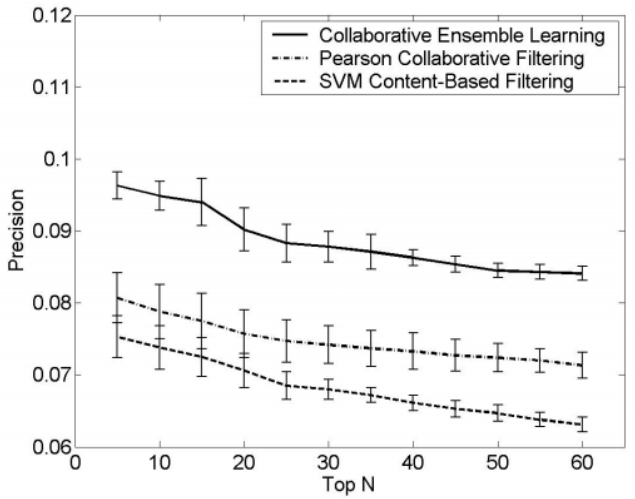
- Task: Predict image rating (642 images)
- 190 users with 89 rated images on average
- Image features:
  - 256 correlogram features (colour/texture)
  - 10 features based on wavelet texture
  - 9 features on colour moments
  - giving a 275-dimensional feature vector for each image.
- Weak indicators of high-level information about an image
- The model predicts N highest ranked pictures; among those, how many were rated positively (in comparison to unrated or negatively rated)

- Classifier: probabilistic SVM with Gaussian kernel
- `http://honolulu.dbs.informatik.uni-muenchen.de:8080/paintings/index.jsp`

Active user: given 20 known



Active user: given 5 known



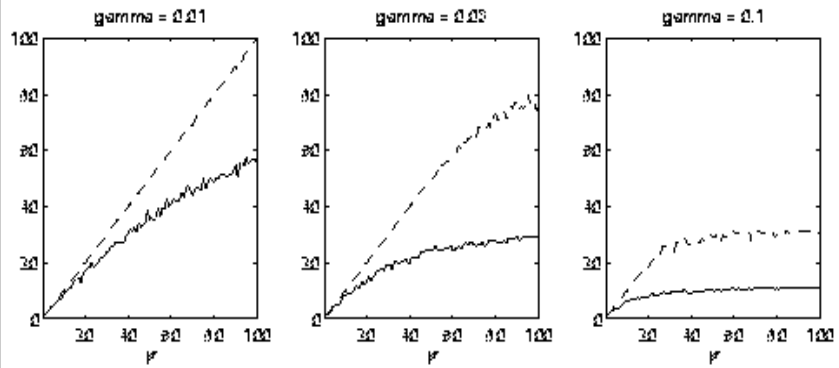
Summary and Outlook

- The work combines content filtering and collaborative filtering under a hierarchical Bayesian framework. A principled hybrid filter is proposed.
- We used nonparametric Bayesian modeling in form of a Dirichlet process to solve a fundamental problem in HB
- Users -> agents, plants, hospitals

- References:
  - Kai Yu, Anton Schwaighofer, Volker Tresp, W.-Y. Ma, and H. J. Zhang. Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes. In *Proceedings of 19th International Conference on Uncertainty in Artificial Intelligence (UAI'03))*, 2003, http://www.tresp.org/papers/uai03.pdf
  - Kai Yu, Volker Tresp, Shipeng Yu: A nonparametric hierarchical Bayesian framework for information filtering. Submitted.



Workshop: Ensemble Methods 2004



- The effective degrees of freedom (continuous) as a function of the number of training data
- Q should be larger than the effective degrees of freedom
- Even better: Q is larger than the effective rank of (any) K
- (gamma is scale parameter)