Volker Schlecht (Germany)

On empirical comparisons of prediction methods for ratings

Abstract

In recent years a multitude of collaborative and hybrid methods for predicting individual ratings has been proposed. However, an overall empirical comparison of the most promising procedures with respect to the same data basis and other performance-measures than the widely-used *AAD* is still missing. We show, that the restriction to the sole use of the goodness-of-fit measure *AAD* might lead to misleading results with respect to business objectives. Furthermore, the ability of these procedures to provide good predictions under realistic assumtions based on the real implications of the user's rating process has never been discussed – let alone investigated. Users can only rate objects if they know them and they also tend to focus more on items with respect to which they (correctly or incorrectly) assume that these objects might appeal to them. Hence, the average of the provided ratings is most likely different from the average of ratings, which they would have provided if they had known and rated each and every item. Thus, one might conclude that the usual equal structure of test and training data is unrealistic. This paper investigates the aptitude of the most important collaborative and hybrid procedures to produce good predictions based on realistic test and training data. For the empirical investigation the MovieLens data set, which consists of whole-number ratings with respect to movies, is used.

Keywords: e-commerce, data mining, prediction, analytics, customer relationship management, econometrics, MovieLens data, SVD, two-mode clustering, MMMF, hierarchical GP-procedure, Hierarchical Bayes, Hierarchical Linear Regression Model, collaborative filtering, bias, AAD, precision, Breese-score, utility, profitability.
JEL Classification: C01, C11, C5, M1.

Introduction

Recommender systems provide their users with individual recommendations with respect to items. Typical online-stores like Amazon.com tend to offer their visitors a multitude of products, which we will refer to with the more general term items. While a visitor of a more traditional brick-and-mortar store can just take a quick look at the items displayed in the relevant section of the shop, this procedure is less satisfying for online-shoppers, since the number of items in the relevant section of the online-store is usually much bigger. Thus, it is both necessary and customary for online-businesses to help their customers to find items they might wish to buy. One approach to solve this problem is to exploit the navigational patterns of electronic visitors to recommend pages, which might lead to the ability to present the relevant (sets of) items to the user at an earlier stage of his or her search (e.g., Gaul, Schmidt-Thieme, 2000, 2001, 2002). Another method is to ask the customers to rate items and to use these ratings to infer which items they might prefer. A huge number of heuristics and also some econometric models exist, which are able to estimate or predict the rating of a particular user for any item which he or she has not rated so far – at least if the user has provided enough ratings and all items have been rated by enough other users. While this approach requires a bit of an effort on the part of the user, it might also lead to serendipitous recommendations for items which the user might otherwise never have discovered – at least if the underlying quantitative method works well. Thus, more helpful personalized recommendations might constitute a competitive advantage which in turn might enhance customer satisfaction and customer retention. Therefore, it is no surprise, that a vast multitude of such rating-estimation procedures has been developed during the recent years. Traditionally, those procedures are divided into three different classes which are referred to as content-based, collaborative and hybrid methods. For every particular user content-based methods only use his or her ratings and all relevant attributes of all items in order to estimate his or her ratings with respect to all items, which he or she has not yet rated. In contrast to that collaborative methods utilize every available rating to estimate one user's ratings but no attribute information. Hybrid procedures always use both all given ratings and the attribute information to deduce the estimated ratings of every single user. With respect to some of the elder methods hybrid procedures tend to outperform both collaborative and content-based approaches (see, e.g., Balabanoviv, Shoham, 1997; Soboroff, Nicholas, 1999; Melville et. al., 2002). Usually content-based procedures lead to the worst results. Moreover, content-based and hybrid approaches require the relevant attributes to be both measurable and available, which has made the collaborative procedures the most popular research subject. During the recent years a number of very promising collaborative methods have been proposed. However, most of the latest methods have been never compared to each other. Also their performace has only been evaluated with respect to the accuracy

of the prediction. Economically more informative measures like the Breese score were rarely reported. Given the recent advances of the collaborative procedures, some of those new collaborative methods might even outperform the hybrid ones. Therefore, the performace of a number of new and highly promosing procedures will be compared empirically with respect to different performance measures in section 1.

The previously published empirical investigations divide the given data set into test and training set by random number procedures. Hence, the distribution of test and training data ought to be quite similar for large test and training data sets. This approach is followed in section 2. However, it should be pointed out that the general approach is highly problematic, since it does ignore the structure of real data. In practice, users acquire more experience and knowledge with respect to items, which they, at least at some point in the past, considered to be probably of interest. E.g., under general circumstances most people only watch movies, which they assumed to be appealing to their tastes. Since the number of existing movies is huge, most users have just time and patience enough to watch a small fraction of those movies. As numerous sources are available that provide more or less detailed information about the movie content, it might be a reasonable assumtion, that the user's guesses are right most of the time or at least if the monetary or opportunity costs involved are high. Thus, the ratings provided by the users should be on average considerably higher than the ratings those users would supply if they were asked to watch and rate randomly chosen films. As a consequence, the predictions for any movie (or even movies which might be more likely to be disliked) are generated from ratings with respect to a set of movies which were mostly liked. Nevertheless, it might be argued that in consequence another effect distorts the real data structure: Given their high expectation the audience might react with excessively low ratings to movies which disappoint(ed) them on a special occasion (e.g., on Saturday night or if they had to pay to watch the movie). However, the danger that a participant is overreacting is less severe, if there are no monetary or opportunity costs involved or if those costs are considered either negligible or simply part of a certain life style (like going out on Saturday night). Since movie recommenders only inquire about the perceived qualitiy of the movie irregardless whether any monetary or opportunity costs were involved or when it was watched, it seems safe to assume, that it is comparatively rarely the case that both costs were involved and the movie was rated directly after the screening of the movie. Since low ratings are much less common in the known data sets (Eachmovie, MovieLens) than high or mediocre ratings, it ought be safe to hypothesize that data distortions due to overreactions are not nearly as important as the above mentioned effects. Hence, only the distortions which are due to the fact that predictions for any item have to be based on ratings concerning mostly favored items, are empirically investigated in section 3. Nevertheless, the capabilities of the different procedures to cope with the ill-tempered ratings could be an interesting subject for future research. Finally, in the last section the results and (economic) implications of these empirical studies are discussed and issues for future research are deduced. In the last section conclusions are drawn from the empirical investigations reported in sections 2 and 3 and practical guidelines with respect to recommendation systems are derived.

The empirical investigation is based on the MovieLens data. This is a data set, which consists of whole-numbered ratings from 1 to 5, which are supplied by the users of an online movie-recommender and which represent their personal degree of enjoyment or dissatisfaction with the respective movie. (The higher the rating, the bigger the user's preference for the movie.) This data set is perhaps the most frequently used one in the data mining and online marketing literature and can be visualized by a huge data matrix, whose rows represent users and whose columns belong to specific items (movies). Most of the elements in this matrix are missing. The aim is to predict the missing values based on the available elements.

1. Performance evaluation

The most frequently used performance indicator for the evaluation of predictions of whole-number (movie) ratings is still the average absolute deviation (AAD), which is similar to the mean absolute deviation (MAD), but measures the average absolute deviation of the actual rating from the respective prediction not from the mean of the ratings. Besides evaluation measures like the average absolute deviation (AAD), which simply measure how well the estimators, whose calculation was based on the training data, fit the actual test data, a number of procedures are used, which are inspired by the goal of online-recommendation. The latter depends on the kind of items which are to be recommended. If the items are movies, users are more interested in getting correct recommendations for items they like very much, since the number of available movies is huge. Thus, only the highest predictions have to be correct. In contrast, it is immaterial, whether all predictions actually fit the

test data well. Also, most people do not wish to invest so much time that they could watch every movie they like. Therefore, it is no big disadvantage, if only a small fraction of the items, that they actually like, is recommended. In that case the precision (PREC) is the appropriate performance measure. With respect to other items the opposite is true. If the items are scientific papers and the users were researchers, it might be more important, that every interesting item can be discovered based on the predicted ratings. In this case it would be less important, if lots of items are recommended, which turn out to be of no interest at all to the respective user. If the researchers are extremely diligent and time is not an issue at all, it might be completely immaterial, how many uninteresting items are recommended – as long as all interesting ones are recommended. Thus, with respect to scientific papers and researchers, the recall (REC) might be considered more important. On the other hand, if the researchers have to meet a deadline, it might become equally important that as many of the interesting items are recommended and as few recommendations for uninteresting ones are given. Under those circumstances precision and recall are equally of the essence, which is why the F-measure should be applied, since it combines precision and recall. Since the precision can often be improved by decreasing the recall and vice versa, precision and recall should in all cases be recorded together. In this work a subsection of the MovieLens data is used. Thus precision is more important. Nevertheless, the recall has to be recorded as well. In addition, the Breese-score is given. The Breese-score (Breese et al., 1998) is based on the assumption, that for every user a list of recommended items is generated. Only items he has not rated so far are put on this list. The higher the predicted rating with respect to an item is, the higher it is positioned on the list. A recommendation is most helpful at the top of the list and nearly useless at the bottom. Therefore, the usefulness of such a list can be quantified as

$$R_{B,i} = \sum_{j_{i,list}=1}^{J_{i,list}} \frac{max(Y_{ij_{i,list}} - d_i, 0)}{2^{(j_{i,list}-1)/(\alpha_c - 1)}}.$$

Here, Y_{ij} is the rating of user i concerning item j. The index $j_{i,list}$ ranks all given test set ratings of person i according to the corresponding predicted ratings. (The higher the predicted rating is, the lower the index will be.) d_i is supposed to be a neutral rating and is therefore chosen to be the average (training set) rating, which the user supplied. α_c is referred to as half-life value and can be interpreted as the rank of the item on the list that has a chance to be seen by the user. Usually, α_c is set to 5. Let $R_{B,i}^{max}$ be the maximal achievable utility, which is the value $R_{B,i}$ might take, if all unknown items were recommended in order of their actual ratings. Then the Breese-score measures the usefulness of the recommendation lists which can be derived by the predicted ratings.

$$R_B = 100 \frac{\sum_{i=1}^{I} R_{B,i}}{\sum_{i=1}^{I} R_{B,i}^{max}}.$$

Higher values for precision and Breese-score indicate higher customer satisfaction and customer retention. Thus, from an economic point of view, precision and Breese-score are much more important than the widely-reported measures of overall accuracy.

Even though the task at hand is actually to predict the ordinal data, rank correlation types of measures like, for example, the Kendall tau rank correlation coefficient or Somers' D are hardly ever used in the literature. While these measures might indeed be more appropriate metrics for the accuracy of whole-numbered predictions, they are less adequate, when the task is to measure the usefulness of predictions, since they tend to ignore the degree of difference between actual and predicted ratings if prediction and actual rating are different. However, given that the actual rating is 5, it makes a tremendous difference, if the predicted rating is 4 or 1. In the first case, the prediction has still captured the right tendency while in the latter case the prediction was completely misleading.

2. Traditional empirical comparison

In this section a number of recent and promising prediction procedures are compared with each other. They are traditionally divided into collaborative and hybrid procedures.

Among the collaborative methods are the singular-value-decomposition-based approach (SVD) of Sarwar et al. (2000), the Maximum Margin Matrix Factorization (MMMF) by Rennie, Srebro (2005), various procedures for two-mode clustering which are referred to as $\hat{S}_Y^1$ (which is the traditional approach, based on the ADCLUS-model by Shepard, Arabie (1979) and further developed by Gaul, Schader (1996)), $\hat{S}_Y^2$ (Banerjee et al., 2004; George, Merugu, 2005) and $\hat{S}_Y^3$ (Cheng, Church, 2000), a special procedure referred to as ordinal two-mode clustering (OTMC) for two-mode clustering of

ordinal data (Schlecht, 2007) and an HB-procedure based on Gaussian processes (HGP) introduced by Yu et al. (2006). The considered hybrid approaches are a hybrid method by Pazzani (HP) (Pazzani, 1999) and two Bayesian procedures, the Hierarchical Bayesian Linear Regression Model (HBLR) by Rossi, McCulloch and Allenby (1996) and HHGP, the Hierarchical Bayesian Linear Regression Model based on Gaussian processes (Schlecht, 2007a), which is a hybrid version of the HGP-model.

A brief introduction of the collaborative prediction procedures is provided in Appendix A. Appendix B contains a short explanation of the hybrid methods.

Table 1. AAD of all procedures with respect to different test data set fractions of the whole data set

Procedure	Test data set of the whole data set								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
SVD	0.712	0.730	0.743	0.750	0.760	0.768	0.776	0.783	0.798
$\hat{S}_Y^1$-TMC	0.723	0.725	0.736	0.739	0.743	0.755	0.768	0.795	0.899
$\hat{S}_Y^2$-TMC	0.717	0.721	0.728	0.727	0.730	0.734	0.738	0.754	0.789
$\hat{S}_Y^3$-TMC	0.716	0.809	0.826	0.848	0.869	0.906	0.944	0.993	1.025
MMMF	0.683	0.667	0.673	0.676	0.680	0.695	0.713	0.747	0.827
OTMC	0.664	0.670	0.672	0.692	0.705	0.717	0.738	0.771	0.843
HGP	0.648	0.664	0.677	0.689	0.714	0.752	0.778	0.807	0.834
HHGP	0.704	0.715	0.720	0.723	0.724	0.734	0.734	0.767	0.795
HP	0.713	0.721	0.721	0.724	0.729	0.727	0.735	0.747	0.785
HBLR	0.741	0.744	0.746	0.746	0.749	0.744	0.755	0.760	0.779

Table 2. Best performing procedures with respect to the AAD-measure and varying test set fractions

Type of procedure	Small test set fraction	Large test set fraction
Collaborative	HGP	$\hat{S}_Y^2$-TMC
Hybrid	HHGP	HP
Collaborative or hybrid	HGP	HP

For the empirical comparison a fraction of the MovieLens data set which contains 1067 users and 418 items is used. Users and items were chosen in such a way that comparatively few values are missing from the resulting data matrix. Still, approximately 78% of the data are missing. All items were rated with integers from 1 to 5, with 5 (1) indicating a very high level of (dis)satisfaction.

Since the applied procedures are very competitive, no procedure exists, that outperforms all other ones with respect to all sizes of the test data fraction and all performance measures. However, the HGP procedure produces the smallest AAD of all collaborative procedures if small test data fractions are used. With respect to small test data fractions and AAD HHGP is the best hybrid procedure for small test data fractions according to the AAD-measure. If high test data fractions are used, the $\hat{S}_Y^2$-TMC method leads to the smallest AAD-values among all collaborative procedures and the hybrid approach of Pazzani (1999) produces the smallest AAD of all procedures altogether. Remarkably, the AAD-values of the $\hat{S}_Y^2$-TMC procedure are almost always very similar to the AAD of the best performing procedure.

A comparison of the procedures based on the Breese-score leads to different conclusions: With respect to the utility of the recommendation lists generated according to the predicted ratings (measured by the Breese-score) HBLR leads to the best results for small test data fractions.

Although MMMF leads to smaller AAD-values up to a test set fraction of 70% than $\hat{S}_Y^2$-TMC, MMMF also performs considerably worse than $\hat{S}_Y^2$-TMC with respect to the utility (which is estimated by the Breese-score) for test set fractions that are smaller than 40%. Thus, we have empirically verified that better overall accuracy of the predicted ratings does not necessarily imply equivalent or improved (estimated) utility of the recommendations generated according to those predicted ratings. Furthermore, even though HHGP is able to achieve comparatively small *AAD*-values for small test set fractions, HHGP also leads to remarkably low Breese-scores. Therefore, the methodology commonly used in the vast majority of empirical investigations so far, has to change. It is highly recommendable, not just to

evaluate the empirical performance of different procedures in terms of accuracy, but to take into account, for which economic (or social) purpose those recommendation procedure is designed and to define a quantitative measure of success (like, e.g., the Breese-score) accordingly.

Table 3. Breese-score of all procedures with respect to different test data set fractions of the whole data set

Procedure	Test data fraction of the whole data set								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
SVD	86.05	78.17	73.66	71.10	69.37	68.32	67.61	66.78	66.51
$\hat{S}_Y^1$-TMC	85.87	74.37	71.76	67.37	61.01	64.22	57.98	54.58	48.56
$\hat{S}_Y^2$-TMC	86.84	77.62	72.27	68.25	66.86	63.51	63.17	61.67	65.34
$\hat{S}_Y^3$-TMC	86.85	58.12	56.04	55.97	51.19	49.98	48.01	46.58	46.27
MMMF	82.89	76.50	72.11	69.55	67.30	65.84	63.90	61.87	56.14
OTMC	87.05	77.91	71.45	68.51	65.55	63.31	61.98	60.65	58.30
HGP	84.41	77.34	74.28	71.37	70.12	68.44	67.36	65.51	64.09
HHGP	80.87	71.37	65.86	63.19	61.36	57.75	57.31	54.03	51.83
HP	85.28	77.19	73.06	70.10	68.59	68.41	68.03	67.17	66.88
HBLR	89.91	81.73	73.37	66.32	62.86	60.27	56.55	53.57	50.95

Table 4. Best performing procedures with respect to the Breese-score and varying test set fractions

Type of procedure	Small test set fraction	Large test set fraction
Collaborative	SVD/OTMC	SVD
Hybrid	HBLR	HP
Collaborative or hybrid	HBLR	HP

With respect to the precision it is a bit harder to identify the best performing procedure since the recall has also to be taken into account (see section 1). Here, HGP, SVD and $\hat{S}_Y^2$-TMC are to be recommended for small test set fractions and HP ought to be used if the test set fraction is large.

Interestingly, the procedures which are recommendable with respect to the Breese-score are not the same methods that achieve the best precision-values. The Breese-score measures how well each user's real ranking of the items can be predicted based on the predicted ratings, while the precision just measures the fraction of correctly identified highest ratings. If a user generally provides quite low ratings, the ranking of his or her ratings might be correctly discovered even though no high ratings are predicted and vice versa. Since the detection of true ranking patterns is much more useful in the context of practical recommendation, the Breese-value is the most important established performance measure.

From an economic point of view it seems desirable, that each new rating can be used as quickly as possible to generate new predictions and derive from these new recommendations. Thus, one needs procedures which can be calculated (preferably even updated) quickly even when both the number of users and items is huge.

Table 5. Precision (top) and recall (bottom) for all procedures with respect to different test data fractions of the whole data set

Procedure	Test data fraction of the whole data set								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
SVD	0.696	0.699	0.708	0.687	0.684	0.680	0.678	0.651	0.570
	0.171	0.156	0.137	0.131	0.116	0.089	0.076	0.067	0.056
$\hat{S}_Y^1$-TMC	0.433	0.445	0.414	0.365	0.463	0.522	0.509	0.493	0.429
	0.098	0.097	0.093	0.072	0.078	0.109	0.111	0.127	0.148
$\hat{S}_Y^2$-TMC	0.626	0.621	0.611	0.594	0.596	0.590	0.563	0.545	0.495
	0.187	0.204	0.193	0.192	0.201	0.200	0.218	0.219	0.231

Table 5 (cont.). Precision (top) and recall (bottom) for all procedures with respect to different test data fractions of the whole data set

Procedure	Test data fraction of the whole data set								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
$\hat{S}_Y^3$-TMC	0.603	0.462	0.421	0.422	0.401	0.411	0.403	0.376	0.366
	0.179	0.132	0.146	0.139	0.133	0.094	0.089	0.078	0.067
MMMF	0.610	0.618	0.621	0.617	0.610	0.603	0.571	0.521	0.406
	0.277	0.273	0.256	0.254	0.238	0.211	0.195	0.161	0.167
OTMC	0.607	0.610	0.594	0.562	0.552	0.525	0.513	0.462	0.408
	0.269	0.273	0.284	0.262	0.270	0.272	0.276	0.292	0.287
HGP	0.687	0.697	0.720	0.721	0.708	0.658	0.567	0.463	0.383
	0.227	0.214	0.178	0.143	0.076	0.030	0.025	0.032	0.049
HHGP	0.615	0.583	0.580	0.563	0.571	0.546	0.543	0.474	0.452
	0.115	0.129	0.118	0.123	0.117	0.141	0.140	0.167	0.178
HP	0.645	0.596	0.621	0.631	0.608	0.601	0.585	0.571	0.509
	0.201	0.178	0.182	0.186	0.181	0.180	0.171	0.199	0.202
HBLR	0.682	0.674	0.652	0.628	0.656	0.659	0.650	0.614	0.538
	0.079	0.083	0.061	0.065	0.062	0.054	0.049	0.061	0.055

Table 6. Best performing procedures with respect to the precision and varying test set fractions

Type of procedure	Small test set fraction	Large test set fraction
Collaborative	HGP, SVD, $\hat{S}_Y^2$-TMC	SVD, $\hat{S}_Y^2$-TMC
Hybrid	HBLR	HP
Collaborative or hybrid	HGP, SVD, $\hat{S}_Y^2$-TMC	SVD, $\hat{S}_Y^2$-TMC, HP

Table 7. CPU-time with respect to all procedures

Procedure	SVD	$\hat{S}_Y^1$-TMC	$\hat{S}_Y^2$-TMC	$\hat{S}_Y^3$-TMC	MMMF	OTMC	HGP	HHGP	HP	HBLR
CPU-time [s]	215.7	51.4	53.2	67.4	189.4	70.6	48.2	47.1	2705.7	8031.9

Additionally, it should be taken into consideration that HGP can lead to less favorable results if the number of items is considerably larger (see Schlecht, 2007). Thus, for many practical applications HGP might be a less desirable choice than these results suggest.

Generally, even if compared to the optimal procedure, the performance of the $\hat{S}_Y^2$-TMC method is almost always only slightly worse. Also, $\hat{S}_Y^2$-TMC is one of the fastest procedures for which highly efficient and quick update algorithms already exist (George, Merugu, 2006). Thus, $\hat{S}_Y^2$-TMC might be a recommendable in many cases, even though the hybrid methods outperform the collaborative procedures with respect to AAD and Breese-score.

3. Empirical comparison considering the biased data structure

In order to obtain test and training data sets which account for the systematic difference between test and training data sets in practical applications, only a (varying) part of the test data set is selected from the set of ratings that do not belong to the test set yet by random number procedures. The remaining part of the test data set is selected from the set of ratings that do not belong to the test set so far and are smaller than 4 and the percentage of those ratings in the test set is referred to as bias-degree and measured by percentage. Table 8 displays the test and training set averages for all five used bias-degrees. In order to be in a position to produce test sets with very high bias-degree, a quite small test set size was chosen (30% of the selected data part).

Table 8. Average rating of test and traing data set with respect to varying bias-degrees

	Bias-degree				
	20%	40%	60%	80%	90%
Training data set	3,52	3,60	3,68	3,77	3,83
Test data set	3,32	3,15	2,96	2,74	2,61

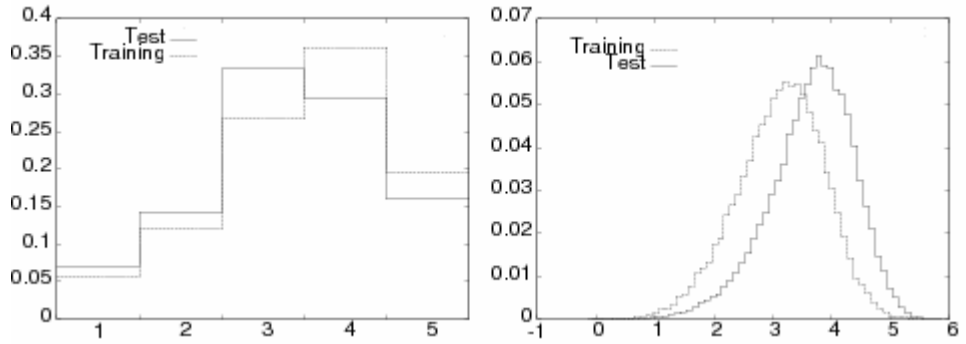

Fig. 1. Histogram of the biased test and training set data, 20% bias-degree (left). Histogram of the biased test and training set estimators, 20% bias-degree (right)

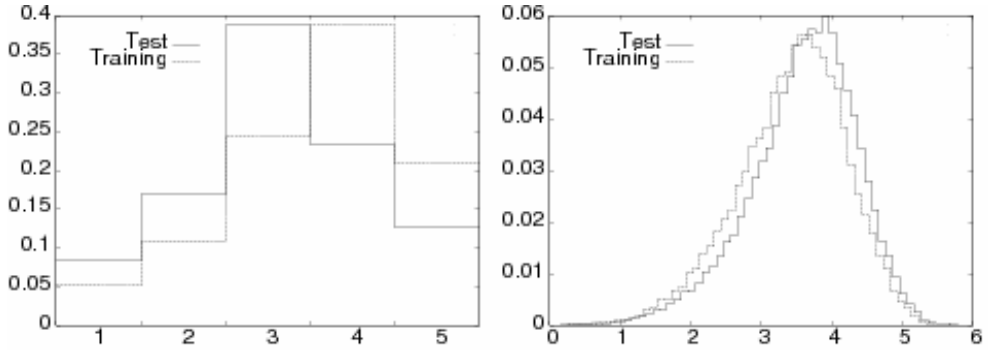

Fig. 2. Histogram of the biased test and training set data, 40% bias-degree (left). Histogram of the biased test and training set estimators, 40% bias-degree (right)

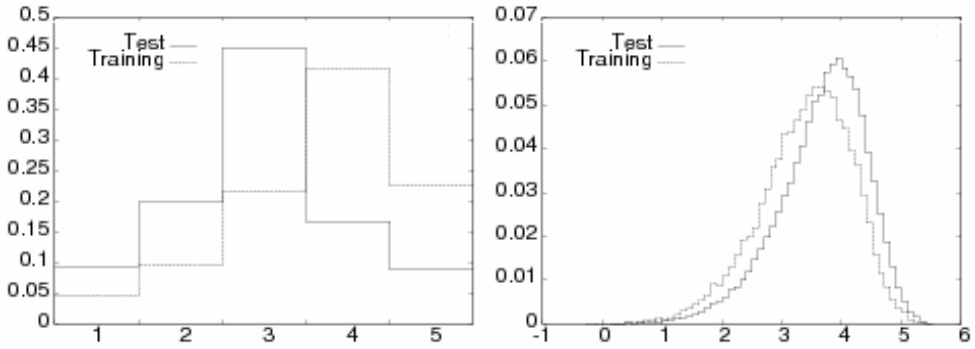

Fig. 3. Histogram of the biased test and training set data, 60% bias-degree (left). Histogram of the biased test and training set estimators, 60% bias-degree (right)

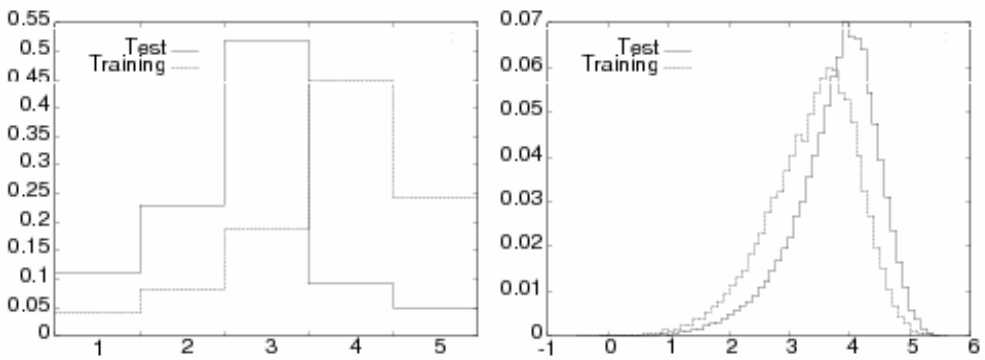

Fig. 4. Histogram of the biased test and training set data, 80% bias-degree (left). Histogram of the biased test and training set estimators, 80% bias-degree (right)

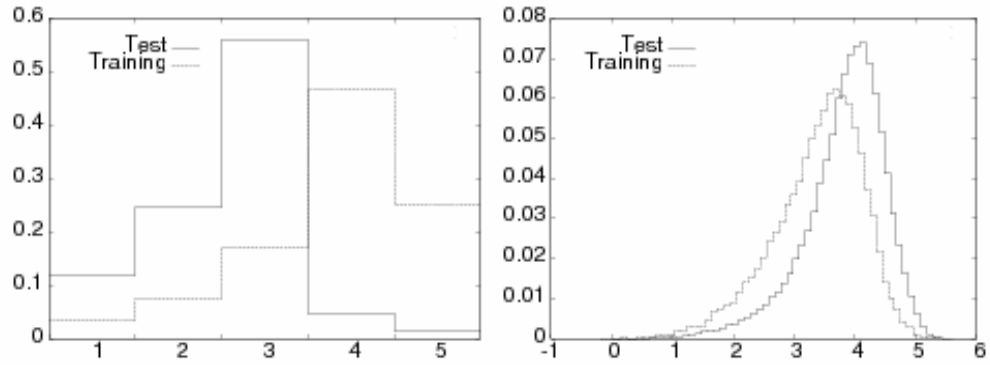

Fig. 5. Histogram of the biased test and training set data, 90% bias-degree (left). Histogram of the biased test and training set estimators, 90% bias-degree (right)

The histograms depicted in Figures 1 to 5 illustrate the systematic effect of the biases on the (average) relationship between actual and predicted (by the $\hat{S}_Y^2$-TMC procedure) ratings. Even though the difference between the distribution of the test and training set is increasing strongly, the distributions of the $\hat{S}_Y^2$-estimators of the test and training sets remain very similar. One can infer directly from the shape of the histograms, that the $\hat{S}_Y^2$-predictions more and more overestimate the actual test set ratings as the bias-degree is increased. This systematic error causes the increase of the *AAD*-values. Since systematic effects can be estimated and thus corrected one could always improve the *AAD* by simply correcting the (biased) $\hat{S}_Y^2$-predictors according to the (estimated) underlying systematic effect.

Again, always the most successful parameters were chosen for all procedures. The results are given in Tables 9 and 10. In Table 9 the Breese-score is reported in parentheses and the *AAD*-value without parentheses. Table 10 contains the precision and the recall, which is given in parentheses. Here, all used performance measures lead to the same or at least very similar conclusions. With respect to small bias-degrees HGP is the best procedure of all under consideration and HP is the best hybrid procedure. If the bias-degree is high, $\hat{S}_Y^2$-TMC is the best choice for minimizing *AAD*, while HGP leads to larger Breese-scores. With respect to high degrees of biasedness, HHGP (HP) is the most successful hybrid procedure for minimizing the *AAD*-measure (maximizing the Breese-score). However, it is remarkable, that so many collaborative procedures outperform the hybrid procedures even though the hybrid procedures additionally exploit content information about the items. Thus, collaborative procedures do not only require no content information at all, are generally considerably quicker, but also tend to be much more reliable if the data structure of the available ratings differs from the general structure, which could be expected if the users knew and rated each and every item.

Table 9. *AAD* and Breese-score (in parentheses) for several procedures with respect to varying bias-degrees

21.5 cmprocedure	Bias-degree				
	30%	40%	60%	80%	90%
SVD	0.751 (71.86)	0.788 (68.89)	0.848 (66.28)	0.948 (62.93)	1.021 (60.44)
$\hat{S}_Y^1$-TMC	0.737 (68.90)	0.779 (67.09)	0.814 (64.93)	0.912 (62.28)	0.947 (65.74)
$\hat{S}_Y^2$-TMC	0.732 (69.69)	0.760 (68.03)	0.807 (65.34)	0.884 (63.93)	0.940 (65.49)
$\hat{S}_Y^3$-TMC	0.834 (59.07)	0.866 (57.32)	0.902 (51.22)	0.942 (49.97)	0.987 (49.66)
MMMF	0.690 (68.56)	0.735 (67.00)	0.806 (65.22)	0.930 (63.71)	1.010 (64.96)
OTMC	0.702 (70.05)	0.736 (67.36)	0.803 (65.42)	0.889 (64.34)	0.954 (65.86)
HGP	0.689 (72.79)	0.734 (71.31)	0.802 (61.41)	0.920 (67.28)	1.008 (68.20)
HHGP	0.734 (66.27)	0.752 (63.72)	0.874 (61.33)	0.892 (61.30)	0.894 (61.77)
HP	0.729 (71.77)	0.760 (69.46)	0.805 (69.07)	0.884 (68.74)	0.936 (68.77)
HBLR	0.748 (66.76)	0.779 (64.13)	0.833 (60.79)	0.925 (57.89)	0.980 (57.15)

Table 10. Precision and recall (in parentheses) for several procedures with respect to varying bias-degrees

21.5 cmprocedure	Bias-degree				
	30%	40%	60%	80%	90%
SVD	0.704 (0.157)	0.641 (68.89)	0.552 (0.119)	0.427 (0.223)	0.261 (0.245)
$\hat{S}_Y^1$ -TMC	0.617 (0.163)	0.450 (0.057)	0.485 (0.182)	0.355 (0.140)	0.268 (0.190)
$\hat{S}_Y^2$ -TMC	0.566 (0.243)	0.478 (0.274)	0.443 (0.271)	0.359 (0.309)	0.209 (0.310)
$\hat{S}_Y^3$ -TMC	0.419 (0.295)	0.358 (0.303)	0.303 (0.323)	0.269 (0.318)	0.194 (0.299)
MMMF	0.587 (0.264)	0.555 (0.295)	0.483 (0.300)	0.353 (0.316)	0.219 (0.342)
OTMC	0.571 (0.266)	0.533 (0.273)	0.483 (0.296)	0.365 (0.304)	0.220 (0.298)
HGP	0.691 (0.206)	0.651 (0.238)	0.567 (0.226)	0.492 (0.268)	0.354 (0.282)
HHGP	0.546 (0.157)	0.475 (0.198)	0.390 (0.218)	0.293 (0.263)	0.325 (0.234)
HP	0.615 (0.210)	0.541 (0.219)	0.465 (0.218)	0.382 (0.263)	0.255 (0.278)
HBLR	0.627 (0.093)	0.587 (0.112)	0.828 (0.123)	0.408 (0.163)	0.274 (0.177)

Table 11. Best performing procedures with respect to the *AAD* and Breese-score (in parentheses) for varying test set fractions

Type of procedure	Small bias-degree	Large bias-degree
Collaborative	HGP (HGP)	$\hat{S}_Y^2$ -TMC (HGP)
Hybrid	HP (HP)	HHGP (HP)
Any	HGP (HGP)	$\hat{S}_Y^2$ -TMC (HP)

Discussion and outlook

The results of this empirical investigation illustrate the increased importance of collaborative procedures.

In section 2, it has been shown empirically, that higher accuracy (lower AAD) does not necessarily imply a higher estimated utility (higher Breese-score). Since the Breese-score is much more important economically, a comparison of different predictive procedures in the context of recommender systems and online-businesses should be rather based on the Breese-score than the AAD.

With respect to precision and recall all matters are slightly more complicated. If the threshold is chosen high, only the ability of the system to identify the highest possible ratings is explored. Given, that anyway far too many items exist, ideally only items for which the respective user supplied the highest possible rating should be considered a success. Otherwise, only the system's ability to identify items which receive above average ratings is explored. Thereby it would not be guaranteed that the user might be satisfied with the resulting recommendations if the precision is high. Therefore, the threshold should be chosen high. However, if a prediction/recommendation is no success by the terms of the definition of precision and recall, it would be useful to know, how bad the mistake actually is. E.g., a mistake might not harm the recommender system if the item which was predicted to be rated with 5 points in reality only received 4 points but it will certainly be detrimental if it received 1 point. Such considerations are accounted for by the Breese-score, not by the precision. In addition, every comparison of the precision has to take the respective levels of the recall into consideration (and vice versa). Thus, the Breese-score not only yields more accurate information with respect to the utility of the ratings but is much easier to interpret.

All known measures have in common, that they never account for how well-known the recommended item is. Even though recommendations for new and less well-known items are more important to the user, since he might know other items anyway, science has rarely tackled the task of developing procedures, that are able to predict new items. Conveniently, the scientists decided, that the originality of the recommendation need not be accounted for in an evaluation of different procedures. Lately, several procedures have been developed and analyzed, which are able to estimate ratings of so far unrated items (Schlecht, 2007; Schlecht, 2008). The best procedures are able to predict ratings for new items with Breese-scores and a level of acurracy, that is almost comparable to the performance of well-known procedures for well-known items. However, if measures were to evaluate the utility of a recommendation, they would definitely need to account for the originality of the recommendations. The development of more appropriate measures to evaluate different predictive procedures for recommendation is necessary and overdue.

Even in the context of recommendation, the Breese-score also has another shortcoming. Since online-businesses venture to make profits and not to maximize their customer's expected utility, it might be advisible to define a performance measure which accounts for the profitability of the resulting recommendations. Furthermore, it should be empirically investigated, how large the actual bias-degree is and if and how strong this bias-degree differs from user to user. If the degree of biasedness varies strongly, it would be interesting to know, whether or not sets of users can be identified for which this degree is stronger or weaker. Depending on the (expected) bias-degree of each user, a different procedure might be recommendable.

In the context of online recommendation it is vital, that reliable recommendations can be made if only few items are rated by the users. Especially at the beginning, when the user is not accustomed to investing his or her time to provide recommendations, the user wonders if the quality of the resulting recommendations is high enough to justify his or her time and efforts. If the user is disappointed at this early stage he or she is highly likely to quit using the recommendation system forever. Moreover, if the recommender system is operated by an online business, he might feel deceived by this business, if the recommended items prove to be unsatisfying to his or her needs and expectations. Thus, the user's loyalty towards this store might decline and in the worst case the user might even stop to use this store. Thus, through bad initial recommendations a recommender system could even become counterproductive and not improve customer satisfaction and retention but impair it.

The selected part of the MovieLens data set contains approximately 98000 ratings of 1067 different users. If the highest test set fraction of 90% is used, only 9.2 ratings are used for the average user. Like in real applications, with respect to some of the users considerably more or less ratings might be available (and actually no user exists for which the training set contains exactly 9.2 ratings). This example illustrates, that high test set fraction simulate the situation when many/most users recently joined the recommender system and have therefore mostly only few ratings. Since (as we have already advocated) good predictions with respect to new persons are vital, the performace of the procedures for low test set fractions is economically most important. As shown in section 3, the HP-procedure yields the highest Breese-scores with respect to high test set fractions and randomly selected test and training sets. Unfortunately, it is comparatively slow (see section 3) and requires content information. Thus the SVD-based approach might be preferable, even though this method yields Breese-values for large test set fractions that are a little bit smaller. However, the SVD-based procedure is not the fastest available procedure (see section 2). Additionally it has been argued that the SVD-based procedure can be updated less efficiently than the TMC-methods (George, Merugu, 2005). Nevertheless, the best performing TMC-approach, the $\hat{Y}_Y^2$-TMC method, results in considerably smaller Breese-scores for large test set fractions.

Both the HBLR-procedure and the HP-method are hybrid procedures. Thus, if test and training data sets are selected randomly, the additional use of information about the item-content leads (as expected) to the best results.

It has been argued (section 1) that in the context of recommender system applications the difference between the distribution of the given ratings and the distribution of the ratings which would result if all the unrated items were known and rated would be considerable. Mainly, the unrated items were supposed to be much more likely to receive unfavorable ratings than the rated ones. Consequently, the simulated large bias-degree (of test and traing set) could to be the most realistic prediction situation. Under those circumstances the (Bayesian) HGP-procedure yields the best Breese-score of all collaborative procedures but is slightly outperformed by the more time-consuming HP-procedure. However, so far it is unknown, if and how strongly test and training set are biased. Thus, it is an important advantage, that HGP is not only very fast in comparison to HP but also yields the largest Breese-score if the bias-degree is low. Nevertheless, it is important to remember that HGP might be less suitable if the number of items is large (Schlecht, 2007a). Thus, the HP-method (and among the collaborative procedures the $\hat{Y}_Y^2$-TMC method) might be a better choice since it performs well with respect to the Bresse-score regardless of the number of items involved and yields good results for all test data fractions and all bias-degrees. Since the HP-procedure is one of the slowest procedures and also requires content-information it might be preferable in many cases to apply the SVD-based approach or the $\hat{Y}_Y^2$-TMC method.

References

1. Balabanovic, M., Shoham, Y. (1997), Content-Based, Collaborative Recommendation, *Communications of the ACM,* 40, 66-72.
2. Banerjee, A., Dhillon, I.S., Ghosh, J., Merugu, S., Modha, D.S. (2004), A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation, *Proceedings of the 10-th ACM SIGKKD International Conference on Knowledge Discovery and Data Mining*, 509-514.
3. Breese, J.S., Heckerman, D., Kadie, C. (1998), Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 43-52.
4. Cheng, Y., Church, G.M. (2000), Biclustering of Gene Expression Data, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 93-103.
5. Gaul, W., Schader, M. (1996), A New Algorithm for Two-Mode Clustering, In: Bock, H.H., Polasek, W. (Eds.): *Data Analysis and Information Systems, Springer*, 15-23.
6. Gaul, W., Schmidt-Thieme, L. (2000), Mining Web Navigation Path Fragments, *Proceedings of the Workshop Web Mining for E-Commerce, The Sixth ACM SIGKDD International Conference on Data Mining*, 105-110.
7. Gaul, W., Schmidt-Thieme, L. (2001), Mining Generalized Association Rules for Sequential and Path Data, *Proceedings of the 2001 International Conference on Data Mining*, 593-596.
8. Gaul, W., Schmidt-Thieme, L. (2002), Recommender Systems Based on User Navigational Behavior in the Internet, *Behaviormetrika*, 29, 1-22.
9. George, T., Merugu, S. (2005), A Scalable Collaborative Filtering Framework Based on Co-Clustering, *Proceedings of the 5th International IEEE Conference on Data Mining (ICDM)*, 625-628.
10. Melville, P., Mooney, R.J., Nagarajan, R. (2002), Content-Boosted Collaborative Filtering for Improved Recommendations, *Proceedings of the 18th National Conference on Artificial Intelligence*, 187-192.
11. Rennie, J.D.M., Srebro, N. (2005), Fast Maximum Margin Matrix Factorisation for Collaborative Prediction, *Proceedings of the 22nd International Conference on Machine Learning*, 713-719.
12. Rossi, P.E., Mcculloch, R.E., Allenby, G.M. (1996), The Value of Purchase History Data in Target Marketing, *Marketing Science*, 15, 321-340.
13. Sarwar, B., Karypis, G., Konstan, J., Riedl, J. (2000), Application of Dimensionality Reduction in Recommender System – A Case Study, *Proceedings of the WEBKDD*, 82-90.
14. Schlecht, V. (2007), Ordinal Data Two-Mode Clustering, *ETU Working Paper*.
15. Schlecht, V. (2007a), Zur Vorhersage ranggeordneter Bewertungen auf unvollständiger Datengrundlage im Marketing, Dissertation, university of Karlsruhe.
16. Schlecht, V. (2008), How to Predict Preferences for New Items, *Investment Management and Financial Innovations*, 5, 4, 7-24.
17. Shepard, R.N., Arabie, P. (1979), Additive Clustering Representation of Similarities as Combinations of Discrete Overlapping Properties, *Psychological Review*, 86, 87-123.
18. Soboroff, I.M., Nicholas, C.K. (1999), Combining Content and Collaboration in Text Filtering, *Proceedings of the IJCAI'99 Workshop on Machine Learning for Information Filtering*, 86-91.
19. Yu, S., Yu, K., Tresp, V., Kriegel, H.-P. (2006), Collaborative Ordinal Regression, to appear in: *Proceedings of the 23rd International Conference on Machine Learning*.

Appendix A. Brief explanation of the collaborative prediction procedures

One of the earliest collaborative procedures applied to the MovieLens data is based on the singular value decomposition (SVD). The rating data are essentially a matrix, in which every row represents a user and every column refers to an item. The SVD-procedure by Sarwar et al. (2000) tries to distill the preference (dis)similarities of the users by a singular value decomposition from this matrix. Obviously this procedure does not account for the ordinal structure of the rating data but proved to be quite effective in practice.

Very similar to the SVD-method are the two-mode clustering (TMC) approaches. Here, users are (re)grouped into user-clusters and items are (re)grouped into item-clusters until the underlying preference pattern is simplified and the complexity of the preference information given by the matrix of ratings is thereby reduced. While the traditional TMC-procedures, $\hat{S}_Y^1$-TMC (Gaul, Schader, 1996), $\hat{S}_Y^2$-TMC (Banerjee et al., 2004; George and Merugu, 2005) and $\hat{S}_Y^3$-TMC (Cheng and Church (2000)), neglect the ordinal data structure, a new TMC-procedure called ordinal two-mode clustering (OTMC) exists (Schlecht, 2007), which takes the ordinality of the data into account.

Another approach, which tries to reduce the complexity of the matrix of ratings is referred to as Maximum Margin Matrix Factorization (MMMF) and uses an objective function, which includes a penalty term, which is suitable for rank-ordered data (Rennie and Srebro, 2005). However, there are other parts of the objective function, which might be inconsistent with the ordinal data structure.

Finally, there is the Hierarchical Gaussian Process (HGP) method (Yu et al., 2006), which is a hierarchical Bayesian procedure based on Gaussian processes. This procedure is also inappropriate for rank-ordered data.

Appendix B. Brief explanation of the hybrid prediction procedures

The hybrid method by Pazzani (1999) first aims at determining a set of attributes. Based on these attributes individual user profiles are derived, which reflect the importance of each of the selected attributes to the specific user. This simple user profile is used to predict the ratings of the missing data (or those from the test data set) by a simple heuristic procedure.

By contrast the hierarchical Bayesian linear regression model (HBLR) by Rossi, McCulloch and Allenby (1996) is a regression approach, which uses the ratings as dependent variable and the (relevant) attributes as independent variables. The most important feature of this regression approach is that it accounts for the individuality of each customer as well as for general tendencies among the users.

Similarly the Hybrid Hierarchical Gaussian Process (HHGP) procedure is nothing but a hierarchical Bayesian lines regression model. The only distinctive feature of the HBLR procedure is a different estimation algorithm.

All hybrid approaches do not take the ordinality of the data into account.