Assessment of Prediction Techniques: The Impact of Human Uncertainty

Kevin Jasberg and Sergej Sizov[✉]

Web Science Group, University of Duesseldorf,
Universitaetsstr. 1, 40225 Duesseldorf, Germany
{kevin.jasberg,sergej.sizov}@uni-duesseldorf.de
https://jasbergk.wixsite.com/research

Abstract. Many data mining approaches aim at modelling and predicting human behaviour. An important quantity of interest is the quality of model-based predictions, e.g. for comparative analysis and finding a competition winner with best prediction performance. In real life, human beings meet their decisions with considerable uncertainty. Its assessment and resulting implications for the statistically evident evaluation of predictive models are in the main focus of this contribution. We identify relevant sources of uncertainty as well as the limited ability of its accurate measurement, propose an uncertainty-aware methodology for more evident evaluations of data mining approaches, and discuss its implications for existing quality assessment strategies. Specifically, our approach switches from common point-paradigm to more appropriate distribution-paradigm. The proposed methodology is exemplified in the context of recommender systems and their established metrics of prediction quality. The discussion is substantiated by comprehensive experiments with real users and large-scale simulations.

Keywords: Human uncertainty · User noise · RMSE · Magic barrier · Distribution-paradigm

Track: Empirical evaluation, Exploratory

1 Introduction

A broad range of algorithms and approaches in data mining aims at modelling and predicting aspects of human behaviour. These efforts are motivated by many practically relevant applications, including various recommender systems, content personalisation, targeted advertising, along with many others. The comparative assessment of methods usually involves implicit or explicit knowledge about user behaviour, either by observing user interactions or by asking users explicitly. In many situations, particular individuals may meet their decisions with considerable uncertainty. In other words, they would not exactly reproduce their decisions when asked twice or multiple times. Consequently, observed decisions must be seen as single draws from individual "feeling"-distributions,

resulting from complex cognition processes, and influenced by multiple factors (e.g. mood, media literacy, etc.). Moreover, and even more important, our knowledge about such distributions may be very limited due to natural restrictions of human behaviour, i.e. it is practically not possible to require the necessary amount of repeated trials for precise location of the underlying distribution parameters. The presence of human uncertainty and our incomplete knowledge about its properties naturally raise the question of assessment validity and reliability. If some approach R_1 shows better results than approach R_2 in the sense of a certain quality metric (prediction accuracy, user satisfaction, etc.) given reference data, can we consider this as a statistically evident proof that approach R_1 *is indeed* better? In the common sense of statistical hypothesis testing, the confident conclusion can be made if the opposite case has a very low probability (type I error) to happen. Under appropriate accounting for human uncertainty, such evidence is often hard to reach.

As a motivating example, we consider the task of rating prediction (common to recommender systems research), along with the Root Mean Square Error (RMSE) [9] as a widely used metric for prediction quality. In a systematic experiment with real users (described in more detail in the forthcoming sections), individuals rated certain media items (movie trailers) multiple times. Only 27% of users have shown constant rating behaviour; 73% of them have given at least two different ratings to the same item; 49% of users have given three or more different responses. Based on the observations made so far, we constructed individual uncertainty models for every user and thus, the considered quality metric (in our case, RMSE) became a random quantity, characterized by a certain probability density function (PDF).
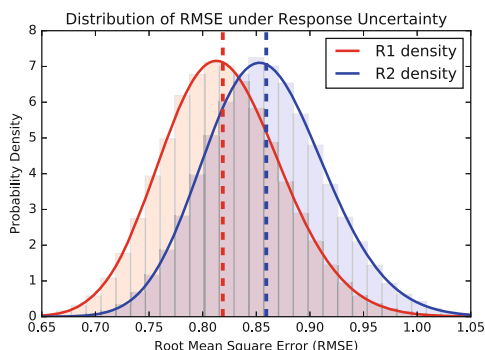

Fig. 1. RMSE as random quantity (Color figure online)

Figure 1 shows corresponding results for two sample recommenders: best possible prediction R_1 (the mean of observed user responses) (red chart) vs. random predictions around the mean R_2 (blue chart); here R_1 is supposed to be the better system by design. As can be seen, there is a large overlap between both PDFs

inducing a probability of $P(\text{"}R_2 \text{ better than } R_1\text{"}) \approx 0.33$. Insofar, the comparison of point-wise calculated quality metrics in a particular experiment is not necessarily evident for a statistically sound proof of method advantages. Without any loss of generality, the observations made so far can be considered as an indicative motivation for a more careful analysis of the following **research questions:**

Q1: How well is human uncertainty measurable and what are the implications of its incomplete assessment onto possible model comparisons?

Q2: How well can distinguishability be reached under the human uncertainty assumption, specifically

    (a) What is a natural metric for the distinguishability of two different models?
    (b) What kind of statistical evidence indicates that a model can still be improved?
    (c) What makes a difference between two models statistically significant?

2   Related Work

In the context of this paper, we exemplify our approach by scenarios from the field of recommender systems as summarised in [14] and focus specifically on comparative evaluation metrics. Recommender systems were initially based on demographic, content-based and collaborative filtering. An overview of these techniques is given in [4]. As collaborative filtering recently turned out to be one of the most successful techniques, they rapidly got into the centre of further research. A roadmap to collaborative filtering as well as a profound discussion on its predictive performance is provided by [17].

Due to the importance of evaluating those recommender systems in terms of their model-based prediction quality, different metrics have been introduced, such as the root mean squared error (RMSE), mean absolute error (MAE), mean average precision (MAP) and normalized discounted cumulative gain (NDCG) (see [1]). Further possible quality-related dimensions of interest in recommender assessment (user satisfaction, precision/recall, etc.) are summarised in [9].

All mentioned quantities have in common the need for human input, either by asking the users explicitly or by observing their interactions. In both the cases, human responses may show a considerable degree of uncertainty, resulting from complex cognition processes and multiple influential factors. Consequently, the main results shown in this contribution can be easily adopted for general cases without substantial loss of validity.

The idea of uncertainty is not only related to predictive data mining but also to measuring sciences such as physics or biology. In this area, a science called metrology has been developed, which is about accurate and precise measurement. Recently, a paradigm shift was initiated on the basis of a so far incomplete theory of error (see [7]), so that variables are currently modelled by probability density functions, and quantities of interest are obtained by means of a convolution of these densities. This model is described in [12]. A feasible framework for computing these convolutions via Monte-Carlo-simulation is given by [13].

We employ this model as a basis for our modelling of uncertainty for addressing similar issues in the field of computer science.

The complexity of human perception and cognition can be addressed by means of latent distributions (see [18]), resulting in varying observations. This idea is widely used in cognitive science and in statistical models for ordinal data. For example, so-called CUB models for ordinal data [10] assume the Gaussian as a latent response model underlying the observations. We adopt the idea of modelling user uncertainty by means of individual Gaussians following the argumentation in [10] for constructing our own response models.

The human impact on the prediction quality was noticed in 2009 when [2] stated, that users are inconsistent in giving feedback and therefore establish an unknown amount of noise that challenges the validity of collaborative filtering. In consequence, [15] has shown that quality metrics cannot exceed certain barriers, grounded in the collective uncertainty of observed user decisions. In order to collect information about human uncertainty, we follow [3] by using repeated rating scenarios for same users and items within conducted experiments designed in accordance with experimental psychology [6,11]. On the basis of the information gathered by using this approach, the authors of [3] were able to develop a preprocessing in order to de-noise the underlying data set of ratings and therefore yield better prediction accuracy. In contrast, we distinguish between non-significant deviations (natural human noise) and significant ones (model induced noise). In this paper, we use the same measuring instrument to collect uncertainty information as in [3] but in this contribution, we also focus on the influence of this uncertainty on the accuracy of recommender systems under the view of metrology. We also take the idea of a pre-processing to reduce the impact of human uncertainty on RMSE under this different perspective.

3   Modelling Human Uncertainty

For evaluating the quality of model-based predictions exemplified by recommender system accuracy, we compare internally computed predictors against real user ratings. Let $\mathcal{I} = \{1, \ldots, I\}$ be the index set of I items and $\mathcal{U} = \{1, \ldots, U\}$ the index set of U users. When several users have rated several items, we obtain $n \leq U \cdot I$ pairs (π_ν, r_ν) of predictors π_ν and ratings r_ν that can be matched against each other where $\nu \in \mathcal{U} \times \mathcal{I}$ is a multi-index. These quantities allow computing single scores of accuracy metrics (e.g. RMSE) which corresponds to the commonly used point-paradigm. By using the metrologic distribution-paradigm, we explicitly account for human uncertainty and its resulting rating uncertainty.

We consider all the given ratings to be a family of random variables $R_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$ which are assumed to be normally distributed as also done in [10]. From this point of view, a given rating r_ν can be seen as the output of a random experiment that is somehow related to human cognition. Hereunder, human uncertainty is strongly related to statistical randomness and the standard deviation σ_ν becomes a natural measure of human uncertainty. In this case, the RMSE becomes a random variable itself, since it is a composition of continuous

maps of random variables. The distribution emerges as a convolution of n density functions under the given mathematical model

$$\text{RMSE} = \sqrt{\sum_{\nu \in \mathcal{U} \times \mathcal{I}} \frac{(\pi_\nu - R_\nu)^2}{n}}. \tag{1}$$

As an example, we consider all n rating distributions to be i.i.d. with $R_\nu \sim \mathcal{N}(\pi_\nu, 1)$ that is, the predictors of our recommender systems perfectly match with the mean of our rating distributions. With these distributions, we want to derive the RMSE's density gradually by specifying the densities for every step of computation that has to be done for calculating the entire RMSE. First, we consider the initial step $S_\nu^1 := \pi_\nu - R_\nu$ which is a random variable distributed by $\mathcal{N}(0, 1)$. Then as sum of n standard normal distributed random variables, the second step $S_\nu^2 := \sum_\nu (S_\nu^1)^2$ yields a $\chi^2(n)$-distribution with n degrees of freedom. Hence, a scaling by $1/n$ will lead to a gamma distribution $S_\nu^3 := \frac{1}{n} \cdot S_\nu^2 \sim \Gamma(\frac{n}{2}, \frac{2}{n})$ and finally for the last step, $S_\nu^4 := \sqrt{Z_\nu^2} \sim \text{Nakagami}(\frac{n}{2}, 1)$ yields the Nakagami-distribution since it is the square root of a gamma-distributed random variable. Under all these conditions, we yield the RMSE not to be a single point but rather to be a Nakagami-distributed random variable with density function

$$f(x) = \frac{2m^m}{\Gamma(m)} x^{2m-1} \exp\left(-mx^2\right) \quad \text{where} \quad m = n/2. \tag{2}$$

whose expectation

$$\mathbb{E}(\text{RMSE}) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \sqrt{\frac{2}{n}} \tag{3}$$

is the average RMSE score according to the point paradigm when repeating the rating scenario infinitely. The advantage of this approach is, that it additionally provides a non-vanishing variance

$$\mathbb{V}(\text{RMSE}) = 1 - \frac{2}{n} \cdot \left(\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\right)^2 \tag{4}$$

as a measure of the uncertainty that is related to the RMSE. The fact that a different RMSE score is achieved each time the rating scenario is repeated, corresponds to drawing a random number from a given RMSE distribution within the distribution-paradigm. Considering a data set of uncertain ratings, two different recommender systems would obtain different RMSEs on this dataset, denoted X_1 and X_2. Let $f_{X_1}(x)$ and $f_{X_2}(x)$ the probability density functions of X_1 and X_2. If those densities overlap, then there is also a non-vanishing possibility of error when building a ranking order by evaluating single scores only (point-paradigm). Let x_1 and x_2 denote two realisations of the RMSEs X_1 and X_2 and let $x_1 < x_2$ be the ranking order by using the point-paradigm, then the probability P_ε of error for this decision is given by $P_\varepsilon := P(X_1 > X_2)$ with

$$P(X_1 > X_2) := \int_{-\infty}^{\infty} f_{X_2}(x)\big(1 - F_{X_1}(x)\big)\, \mathrm{d}x \leq 0.5 \tag{5}$$

where $F_{X_1}(x) := \int_{-\infty}^{x} f_{X_1}(t)\,\mathrm{d}t$ denotes the cumulative distribution function of f_{X_1}. Later, it will be shown that a ranking built by using the point-paradigm is associated with considerable errors caused by human uncertainty. However, this can virtually be subtracted out in a pre-processing step.

From the view of the distribution-paradigm, each time a given rating is compared with a model-based prediction, we must examine whether the observed deviations are significant or just in nature of contingency, i.e. the influence of human uncertainty. In doing so, we divide the set of all deviations into two subsets. One subset contains all the deviations around the predictor π_ν that can be considered as human uncertainty and the other subset contains all deviations whose extent cannot be explained by this uncertainty and thus seems to be induced by the prediction model. In this case, it seems viable to calculate the RMSE by taking into account only those deviations that are related to the algorithm rather than to human uncertainty. Similarly to the classic RMSE, we refer to this more natural metric as the significant RMSE (**sRMSE**). Following this approach, we have to use statistical hypothesis testing to decide whether a realisation r_ν of the rating distribution R_ν is equal to a model-based prediction π_ν or not. In mathematical notation, we have to test

$$H_0 \colon r_\nu = \pi_\nu \quad \textbf{vs.} \quad H_1 \colon r_\nu \neq \pi_\nu \tag{6}$$

for every multi-index ν at a given significance level α. For known density functions f_{R_ν} of the rating distributions R_ν the critical region can be constructed as the complement of $I_\alpha = [\pi_\nu - a; \pi_\nu + a]$ where a is chosen such that

$$\int_{\pi_\nu - a}^{\pi_\nu + a} f_{R_\nu}(x)\,\mathrm{d}x = 1 - \alpha. \tag{7}$$

We now yield the probability density function of the sRMSE by a convolution of the pseudo-restrictions $f_{R_\nu}|_{I_{95}^{\complement}}(x) := \mathbb{I}_{I_{95}^{\complement}}(x) \cdot f_{R_\nu}(x)$ where $\mathbb{I}$ is the indicator function. Due to this definition, the sRMSE grants assessment of different recommender systems with much lower probabilities of error. This can be explained by not taking into account the stabilising centre of all the rating-distributions and as the RMSE amplifies the remaining extreme values by its quadratic term (see Eq. 1), the distributions rapidly differ under increase of false predictions. Having in mind this mathematical model of human uncertainty in terms of the novel metrologic distribution-paradigm, we elaborate on our research questions by examination of real life scenarios.

4   User Study and Simulations

In practice, the application of the previously described model is technically challenging. Let the rating distributions $R_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$ be not necessarily equal for every ν. As it has been shown in [5], the sum of squared deviations receives the density of a non-central χ^2-distribution. At this point, it is quite hard to find a closed form for the RMSE density. It turns out that efficient dealing with

the RMSE's distribution can only be maintained by using statistical simulations when general cases are taken into account. In this paper we use **Monte-Carlo-Simulations** (MC) as described in [13]: For every input variable $R_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$ we take a sample $\mathcal{S}(R_\nu) := \{r_\nu^1, \ldots, r_\nu^\tau\}$ of τ pseudo-random numbers (trials) that are drawn from this specific distribution. Due to the randomness, further computations may fluctuate slightly, but his effect diminishes for a high number of trials. In our analyses, we reached stable results by setting $\tau = 10^6$. With these samples we compute $\mathcal{S}(\mathrm{RMSE})$ by

$$\mathcal{S}(\mathrm{RMSE}) = \left\{ y_j = \sqrt{\sum_\nu \frac{(\pi_\nu - r_\nu^j)^2}{n}} : j = 1, \ldots, \tau \right\}. \tag{8}$$

Post hoc illustration of this sample by a normalised relative histogram with b bins lead to an approximation of the RMSE's density. Our analyses often focus on the error probability P_ε as described in Eq. 5. In the following numerical simulations this probability is efficiently computed by

$$P_\varepsilon = P(\mathrm{RMSE1} > \mathrm{RMSE2}) = |A|/\tau \tag{9}$$

where A is the set of all $(r_i, s_j) \in \mathcal{S}(\mathrm{RMSE1}) \times \mathcal{S}(\mathrm{RMSE2})$ holding the condition $r_i > s_i$ for $i = 1, \ldots, \tau$. For modelling human uncertainty, we assume a set of known rating distributions, based on perceptions about real user behaviour from comprehensive user experiments.

User Experiments

Our experiment is set up with Unipark's[1] survey engine whilst our participants were committed from the crowdsourcing platform Clickworker[2]. During the experiment, participants watched theatrical trailers of popular movies and television shows and provided ratings on a 5-star scale multiple times in random order. The submitted ratings have been recorded for five out of ten fixed trailers so that the remaining trailers act as distractors triggering the misinformation effect, i.e. memory is becoming less accurate because of interference from post-event information. Altogether, we received a rating tensor $R_{u,i,t}$ with $\dim(R) = (67, 5, 5)$, having $N = 1\,675$ data points in total, where the coordinates (u, i, t) encode the rating that has been given to item i by user u in the t-th trial. From this dataset we derive a unique rating distribution for every user-item-pair by considering tensor-slices in trial-dimension $R_{u,i} := R_{u,i,\bullet} = \{R_{u,i,t} | t = 1, \ldots, 5\}$ which can be easily depicted in a relative histogram and modelled by a certain rating distribution. In our experiment, only few tensor slices contain constant ratings and hence lead to a vanishing variance. Performing an item-wise analysis, the fraction of tensor slices with non-zero variance ranges from 50 to 90% that is, only every second participant is able to reproduce its own decisions for the best case.

---

[1] http://www.unipark.com/de/.
[2] https://www.clickworker.de/.

For the worst case, only one out of ten participants is able to precisely reproduce a rating. All tensor slices containing a non-vanishing variance are checked for normality by a KS-test at $\alpha = 0.05$. The null hypothesis was never rejected, allowing to keep the Gaussian distribution as a possible model (rationally, it exhibits maximum entropy among all distributions with finite mean/variance and support on $\mathbb{R}$).

Research Question Q1: Measurability of Human Uncertainty

Description: Based on our user study, we assume $R_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$. Since this study only surveyed a sample rather than an entire population, point estimates for the distribution parameters would be inappropriate. Instead, confidence intervals have to be specified. Following [8], the confidence interval for the parameter μ_ν can be received by

$$\mu_\nu \in \left[\bar{x}_\nu - t_{(1-\frac{\alpha}{2};n-1)} \frac{s_\nu}{\sqrt{n}} \; ; \; \bar{x}_\nu + t_{(1-\frac{\alpha}{2};n-1)} \frac{s_\nu}{\sqrt{n}} \right] \quad (10)$$

where $\bar{x}$ and s are the point estimates for the mean and bessel-corrected standard deviation and $t_{(p;k)}$ represents the p-quantile of the t-distribution with k degrees of freedom. Following [16], the confidence interval of σ_ν is given by

$$\sigma \in \left[s\sqrt{(n-1)/\chi^2_{(1-\frac{\alpha}{2};n-1)}} \; ; \; s\sqrt{(n-1)/\chi^2_{(\frac{\alpha}{2};n-1)}} \right] \quad (11)$$

where $\chi^2_{(p;k)}$ is the p-quantile of the χ^2-distribution with k degrees of freedom. This means that we can not simply determine a single rating distribution for each data set. Instead, a variety of rating distributions needs to be computed for each user-item-pair where the associated parameters are drawn from the corresponding confidence interval. Even for large-scale computations, the resulting RMSE does not possess a stable density function. However, we can consider borderline cases which reveal the maximum span in which we can expect results for the density function of the RMSE. On this basis we run three simulations:
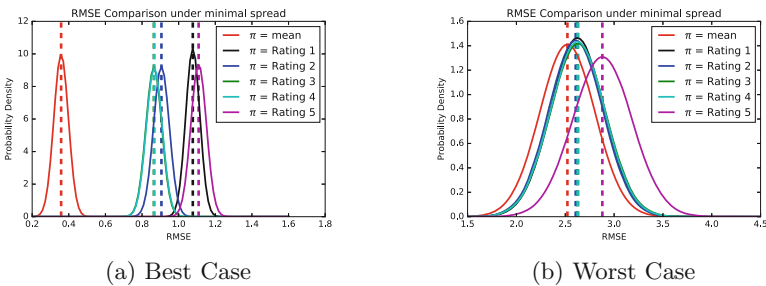

(a) Best Case   (b) Worst Case

Fig. 2. Borderline cases of RSME for different recommender systems

Simulation 1: In Simulation 1 we compute these borderline cases by assigning the parameters μ_ν and σ_ν as the lower limits of the corresponding confidence interval and the upper limits respectively. In doing so, we first build six recommender systems by defining their predictors via where k denotes the k-th recommender systems. Then, for every recommender systems we compute a sample $\mathcal{S}(\mathrm{RMSE}(R\,k))$ for all borderline cases as described in Eq. 8 and generate the ML-density functions. In this simulation we use $\tau = 10^6$ MC-trials for steadiness of histograms as well as $b = 55$ bins for accurate display of densities. Figure 2 shows the impact of the uncertainty of the re-rating-proceeding. Whilst we can recognise a good resolution for three groups of RMSEs in the minimum case, this is virtually no longer possible for the maximum case. The true distributions of the individual RMSEs can vary between these two thresholds but remain unknown to us on the basis of the information collected. In short, with only five re-ratings it is not possible to get high-quality uncertainty information, but it must be said that this phenomenon is not grounded within the point-paradigm itself. In practice, we have to distinguish between two different types of uncertainty: On the one hand, there is the human uncertainty (leading from scores to distributions) which is in the main focus of this contribution. But on the other hand, there is also a kind of measurement error which we call the method uncertainty. The variability for the RMSE distributions in Fig. 2 is completely explained by the impact of this method uncertainty.

Simulation 2: The method uncertainty can be reduced by decreasing the width of the confidence intervals that scale with $1/n^q$ for some $q \in \mathbb{R}$. To this end, it is necessary to increase the number of re-ratings. Accordingly, the borderline cases of the RMSE converge to a stationary state for large n. In this Simulation we estimate the number of re-ratings in order to get stable results, so we can speak of the *true* RMSE. As a measure of this convergence, we calculate the intersection area of the minimum and maximum density for each recommender system. As can be seen from Fig. 3(a), we need about 1000–2000 re-ratings, so that both distributions converge to a steady state by more than 90%. This means that users in a real rating scenario would have to re-evaluate the same item at least 1000 times in order to locate the RMSE-distribution accurately.

Simulation 3: If it is not feasible to calculate the stationary state with the re-rating-proceeding, then it might be sufficient to only gather samples as large as to exclude the high error probabilities of the maximum case. This is simulated by fixing the point estimates $\bar{x}$ and s and artificially increasing the sample size n to calculate the boundary points of our confidence intervals in Eqs. 10 and 11. With those, we determine the error probabilities for a point-paradigm ranking of recommender system 1 to all the other recommender systems for each of the borderline cases. Figure 3(b) depicts the error probabilities $P_\varepsilon = P(\mathrm{RMSE}(R1) > \mathrm{RMSE}(R3))$ for the minimum and the maximum case. All the other cases of $P_\varepsilon = P(\mathrm{RMSE}(R1) > \mathrm{RMSE}(R\,k))$ lead to equivalent results for $k \neq 1$. As we can see, we would need about 500 re-ratings to regard the RMSE approximation to be satisfactory, if we accept a maximum of $P_\varepsilon \approx 0.10$.
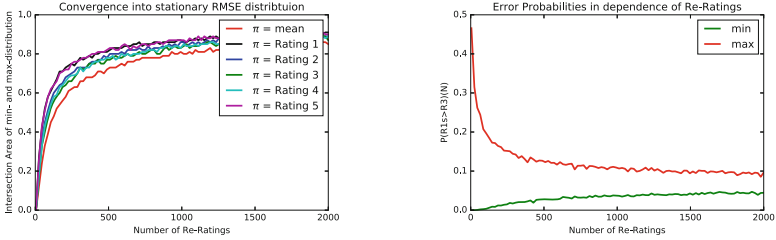
(a) Convergence of minimum and maximum RMSE

(b) Example of convergence of error probabilities

Fig. 3. Convergence into the stationary state

Research Question Q2b: Statistical Evidence for Improvements

Here, we examine the conditions under which a single recommender system can not be distinguished from a theoretically optimal recommender system by means of the RMSE. The idea of this investigation is to create a copy of a given recommender system and to distort this copy by artificial uniform-noise. This is done by resampling its predictors $\pi_1 \in [(1-p)\pi_0 \,;\, (1+p)\pi_0]$ assuming a uniform distribution. In this case, a noise fraction of p means that those new predictors deviate from the originals by $100p\%$. The RMSE thereby receives a shift on the x-axis so that it's possible to build a ranking along with its associated error probability. We can apply these as a function of the noise component. Noise is, in this context, a specific quantity for inducing differences in recommender system quality in a controlled manner.
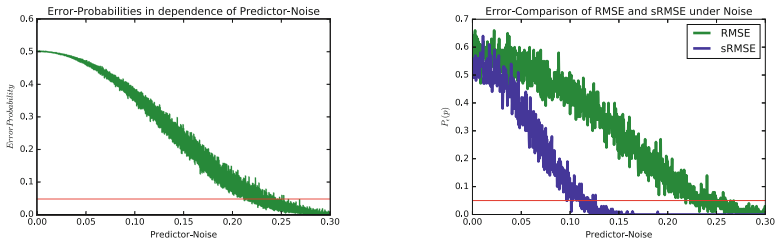
(a) RMSE

(b) RMSE / sRMSE in comparison

Fig. 4. Error probabilities as a function of artificial predictor noise

Simulation 4: The expected value of a random variable is the value which is obtained on average in the case of infinite repetitions of the random experiment and thus has the smallest sum of squared deviations. Theoretically, this property makes the arithmetic mean $\bar{x}_{u,i}$ of the data series $R_{u,i}$ the optimal predictor. Hence, we define the optimal recommender system by setting $\pi_{u,i} := \bar{x}_{u,i}$, so statements can be generated which are correct for very large investigations on the average. To this optimum, we additionally create a copy which we distort by

artificial uniform-noise as described and specify that two recommender systems can be distinguished significantly if the error probability is less than 5%. In this simulation we again use $\tau = 10^6$ MC-trials for each of the 10^6 data points (p, P_ε), having 10^{12} trials in total. Figure 4(a) shows the curve of the error probability where the width of this graph is an artefact of the uniform-noise. We can see that the error probability drops below the 5% mark in a range of 21% to 24%, i.e. only then distinctions to the optimum can be reliably detected. This proves the existence of a certain barrier of prediction quality so that any superior recommender system can not be differentiated from the best possible recommender system anymore.

Research Question Q2c: Significant Differences of two Models

In real life, assessments compare several recommender systems among each other. This is taken into account in the following simulations.

Simulation 5: We generate two copies of an optimal recommender, with different proportions of added noise in such a way that the relative noise difference of both copies remains constant. Then, we compute the resulting RMSEs for both copies together with an error probability for the point-paradigm ranking. By increasing the noise for both copies whilst keeping their relative difference constant, we generate an offset (deviation from the optimum or prediction quality) and can thus plot the error probabilities against this offset for different noise ratios. This simulation was performed with 10^{12} data points. Figure 5 depicts the family of curves mapping the noise offset to the corresponding error probabilities. The offset represents background noise and is a measure of the deviation from the best possible recommender system, i.e. the larger the offset, the worse the prediction quality of the recommender system. The colours encode the relative difference Δ of two recommender systems among each other. For the green curve (representing 10% noise of difference), an x-value of 0.15 means that RS1 has a noise of 15% whereas RS2 has a noise of 25%. The corresponding y-value indicates the error probability for ranking both of these recommender systems using the point paradigm. It is apparent from this Figure, that two systems can not be brought into a ranking order without considerable error probability if their relative difference is below 15%, regardless of their basic prediction quality. Figure 5 also reveals that only for noise differences of more than 20%, two different systems can be distinguished starting from a certain quality. As a result, we recognise the following: The better a system becomes, the more improvement does a revision need in order to be detected with statistical evidence.

Simulation 6: In order to make our results more tangible and comparable to current competitions (e.g. the Netflix Prize), we define the RMSE difference as the relative difference in the expectation values of both distributions for this difference uses to be the best estimation for an infinitely repeated rating scenario. We rerun the last simulation, but now determine the RMSE distances by using adaptive noise: We only add so much noise until we reach the desired
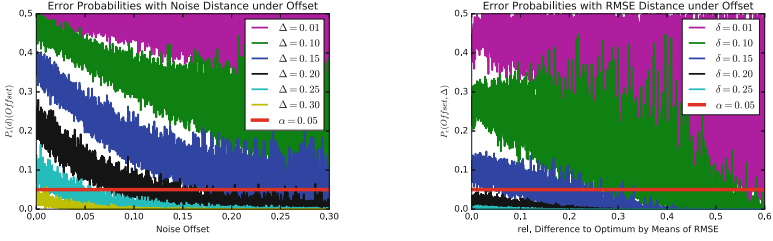
(a) artificial noise as a measure for recommender difference

(b) relative RMSE-differences as a measure for recommender difference

Fig. 5. Error Probabilities for two suboptimal recommender systems

RMSE difference. Then we compare the error probabilities by means of those RMSE distances. For the RMSE distances, a similar result is obtained. Two systems differing by 10% in terms of RMSE must deviate more than 40% from the optimum to be distinguished significantly. In reverse, if the closeness of two systems to the theoretical optimum (i.e. the offset) remains unknown - which is probably always the case in real life assessment - then both systems would only be distinguishable with statistical evidence, if they differ at least 20% in terms of the RMSE (since only the 20%-curve is below the 5%-mark for any offset).

Human Accuracy Metrics

At this point, we investigate the resolution properties of two recommender systems by means of the sRMSE. This is performed by a hypothesis test as described in Sect. 3 and considering only significant deviations from the rejection range to compute an RMSE. As a result, the sRMSE could theoretically distinguish between two recommender systems even with fewer deviations.

Simulation 7: In practice, the hypothesis test is performed by constructing a symmetric interval around the predictor π_ν within the rating distribution of R_ν until the density's area over this interval sums up to 0.95. All values in this interval do not represent any significant deviations and are not taken into account in the sRMSE. We hence generate pseudo-random numbers according to R_ν until we have $\tau = 10^6$ values in the rejection range and use these to compute the sRMSE distribution. For these density functions, we now repeat the procedure from simulation 4. The results are depicted in Fig. 4(b). Here we see error curves under noise in the form of a comparison of RMSE and sRMSE. It can be seen that the sRMSE grants substantially faster distinguishability from an optimum with statistical evidence than the traditional RMSE. Using this metric, a recommender system can already be distinguished from a theoretical optimum with 10% of noise whereas the RMSE would probably need more than 20%. A repetition of simulation 5 and 6 leads to equivalent results. This proves the better distinguishing features of the sRMSE as predicted by theory.

5   Discussion

The lessons learned so far can be summarised as follows:

1. Due to the blur of the RMSE, an ordering relation is sometimes very difficult to define; we can only give probabilities for the existence of a particular order relation: The probability $P_\varepsilon := P(R1 > R2|\mathbb{E}(R1) < \mathbb{E}(R2))$ for making an error when following the point-paradigm has proven to be an intuitive and very good metric. It correlates positively with the overlap of two RMSE distributions and is hence a good measure for the distinguishability of two recommender systems and also serves as a p-value for hypothesis testing.
2. A recommender system is only to be significantly distinguished from an optimum if it differs by more than 21 to 24% in terms of noise. Below this limit, it cannot be distinguished with evidence.
3. The distinguishability of two systems is not dependent solely on its (noise) difference, but also on their basic quality, that is, from their distance to a theoretical optimum. The worse two recommender systems predict, the less they have to differ in order to be distinguished evidently and vice versa.
4. Methods for collecting uncertainty information are yet to imprecise; the parameters of the rating distributions have such wide confidence intervals, that specifying RMSE densities is not reliable. We need between 500 and 1000 re-ratings to exclude the worst case and about 2000 re-ratings for stable results. The method of re-rating-proceeding must, therefore, be improved.

The most notable results are 2 and 3, since they show a natural limit for the resolution of evaluation metrics (which is also always present in the point paradigm but can not be made accessible). Result (2) implies the existence of an equivalence class of optimal recommenders because all recommender systems below a certain RMSE value are no longer to be distinguished from the optimum. Result (3) generalises this fact and raises the fundamental question of assessment evidence. On the basis of our results, the suggested solution of using the sRMSE has proven to be quite fruitful for evaluating prediction quality. In the our simulations, the sRMSE outperformed the traditional RMSE by far, i.e. the resolution capability for two recommender systems was doubled.

6   Conclusion and Future Work

It has been shown that accounting for natural human uncertainty is essential for objective and statistically evident interpretation of ratings and their predictions. In this contribution, we considered recommender systems and their assessment by means of the RMSE, as a characteristic evaluation scenario. It can be assumed that similar influences might be observed for other metrics accounting for uncertain inputs, such as ratings and browsing behaviour. For example, the results presented here could be reproduced in an equivalent form for the metrics average absolute deviation and mean signed deviation. Similar influences might be found not only in recommender systems but also anywhere in predictive data mining

where human behaviour is to be analysed. We were, therefore, able to provide initial indications that human uncertainty may have a striking influence on the predictive data mining and thus on all the areas that build upon it. On this basis, further research may lead into various directions: For **theoretical research**, the overall goal is to develop a complete mathematical model of human uncertainty providing large connectivity for practical applications. For **practical research**, it would be quite profitable to assimilate technical approaches and sensitising them for human uncertainty. This could be done by developing Bayesian prediction models with informative priors based on advanced experiments.

References

1. ACM: Workshop on Recommendation Utitlity Evaluation: Beyond RMSE, vol. 9, Dublin, Ireland, September 2012
2. Amatriain, X., Pujol, J.M., Oliver, N.: I like It.. I like it not: evaluating user ratings noise in recommender systems. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 247–258. Springer, Heidelberg (2009). doi:10.1007/978-3-642-02247-0_24
3. Amatriain, X., Pujol, J.: Rate it again: increasing recommendation accuracy by user re-rating. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 173–180. ACM (2009)
4. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl.-Based Syst. **46**, 109–132 (2013)
5. Chan, F.K.: Miss distance-generalized variance non-central chi distribution. In: AAS/AIAA Space Flight Mechanics Meeting, pp. 11–175 (2011)
6. Döring, N., Bortz, J.: Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer-Lehrbuch, 5th edn. Springer, Heidelberg (2016). doi:10.1007/978-3-642-41089-5
7. Grabe, M.: Grundriss der Generalisierten Gauß'schen Fehlerrechnung. Springer, Heidelberg (2011). doi:10.1007/978-3-642-17822-1
8. Henze, N.: Stochastik für Einsteiger. Springer Spektrum, Heidelberg (2013)
9. Herlocker, J.L.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
10. Iannario, M.: Modelling uncertainty and overdispersion in ordinal data. Commun. Stat. - Theory Methods **43**, 771–786 (2014)
11. Intraub, H.: Presentation rate and the representation of briefly glimpsed pictures in memory. J. Exp. Psychol. **6**, 1–11 (1990)
12. JCGM: Guide to the expression of uncertainty in measurement. Technical report, BIPM (2008)
13. JCGM: Supplement 1 to the GUM - propagation of distributions using a Monte Carlo method. Technical report, BIPM (2008)
14. Ricci, F., Rokach, L., Shapira, B. (eds.): Recommender Systems Handbook. Springer, Boston (2015). doi:10.1007/978-1-4899-7637-6
15. Said, A., Jain, B.J., Narr, S., Plumbaum, T.: Users and noise: the magic barrier of recommender systems. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 237–248. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31454-4_20
16. Sauerbier, T.: Statistik für Wirtschaftswissenschaftler. Oldenbourg (2003)

17. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell., 01 (2009)
18. Vandekerckhove, J.: A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. J. Math. Psychol. **60**, 58–71 (2014)