Combining preference and absolute judgements in a crowd-sourced setting

Peng Ye PENGYE@UMIACS.UMD.EDU
Institute for Advanced Computer Studies University of Maryland, College Park, MD, USA

David Doermann DOERMANN@UMIACS.UMD.EDU
Institute for Advanced Computer Studies University of Maryland, College Park, MD, USA

Abstract

This paper addresses the problem of obtaining gold-standard labels of objects based on subjective judgements provided by humans. Assuming each object can be associated with an underlying score, the objective of this work is to predict the underlying score efficiently and accurately based on preference and absolute judgements via experiments in a crowd-sourced setting. Unlike previous information aggregation methods which consider preference and absolute judgements independently or convert one to another in an ad-hoc way, the proposed method combines the two types of judgements directly via an unified probabilistic model. Additionally, we introduce a batch-mode active learning method which actively constructs a set of queries consisting of preference and absolute judgement tests which maximize the expected information gain at each iteration of the experiment. Experimental results show the effectiveness of the proposed method.

1. Introduction

Estimating gold-standard labels (or strengths, scores, etc) of objects based on subjective judgements provided by humans is a critical step in psychological experiments with applications in many research fields. Depending on the specific applications, there are two primary objectives of these experiments. The first is to estimate the genuine score for the object and the second is to obtain a consensus ranking over a set of objects. For example, with the goal of improving user's experience with a service (web browsing, phone calls, video chatting, online shopping, etc), in Quality of Experience (QoE) studies, the degree of users' subjective satisfaction has to be estimated. On the other hand, a consensus ranking over objects is required in studies, such as information retrieval, collaborative filtering, social choice and online gaming. In this paper, we focus on the task of inferring gold-standard scores of objects, however the proposed method can also be used to rank the objects.

Various types of judgements can be used to estimate the underlying score. The most widely used two types of judgements are the absolute judgement and the preference judgement. An absolute judgement involves assigning a categorical label to a single object. For example, with Mean Opinion Score (MOS) test for QoE assessment, subjects are asked to rate objects using an ordinal scale: "Bad", "Poor", "Fair", "Good" and "Excellent". Despite the popularity of the MOS test, there are many known problems (Carterette et al., 2008; Chen et al., 2009). First, most previous work on QoE (Ribeiro et al., 2011) treat the MOS scale as an interval scale instead of ordinal scale and assume that the cognitive distances between the consecutive MOS scales are the same. However, assumptions such as: "Fair"-"Poor"="Good"-"Fair", are usually not true in practice. Second, absolute rating procedures are somewhat obscure to experiment subjects in that the subjects can be easily confused about which scale they should give in each test and the resulting absolute judgements can be very noisy.

Preference judgements can be in the form of full or partial rankings, relative item comparisons, or a combination. In this paper, we consider preference judgements obtained by comparing pairs of objects, since different forms of the preference can be converted to a set of pairwise preferences. In the simplest experimental unit, two objects *A* and *B* are presented to a single subject, who must "prefer" one of them (Thurstone, 1927; David, 1988; Carterette et al., 2008; Chen et al., 2009). Compared to the ordinal scale rating test, making a decision in a paired comparison test is much simpler and leads to less confusion for the subject. However, when n objects need to be compared, the total number of pairs is $\binom{n}{2}$ and when n is large, the cost for obtaining a full set of pairwise comparisons can be very high.

To overcome these limitations of absolute and preference judgements, we propose a systematical way to combine the two information sources. In this paper, we will answer the following two questions:

1. Given a collection of absolute judgements and preference judgements, how can we combine this two information sources to obtain more accurate estimation of the underlying score?

2. In a crowd-sourced setting, subjective judgements are obtained at a certain cost. In order to minimize the cost, how to determine which samples shall we ask subjects to judge and what types of judgements we ask them to make.

2. Related Work

In the QoE community, most previous work considers absolute judgements and preference judgements independently. (Ribeiro et al., 2011) performed MOS test for QoE assessment using crowdsourcing. They developed a two-way random effects model to model uncertainty in subjective tests and proposed a post-screening method and rewarding mechanism to facilitate the process. (Chen et al., 2009) proposed a crowdsourceable QoE assessment framework for multimedia content, in which interval-scale scores are derived from a full set of paired comparisons, but the problem of how to efficiently down-sample pairs was not discussed.

Another line of related work has focused primarily on the ranking problem (Pfeiffer et al., 2012; Volkovs & Zemel, 2012; Chen et al., 2013), where preference aggregation models are developed for combining pairwise comparisons to get a consensus ranking. (Pfeiffer et al., 2012) introduced an active learning method based on the Thurstone-Mosteller Case V model (Thurstone, 1927; Mosteller, 1951) for pairwise ranking aggregation. At each iteration of an experiment, this method adaptively chooses one pair of objects and asks a subject to compare this pair of objects. The paper shows that the active learning method can reduce the total cost of paired comparisons relative to a random sampling method. However, when an observation on a new pair is obtained, the model has to be retrained, which can be prohibitively expensive to use in a crowd-sourced setting. To overcome this problem, (Chen et al., 2013) proposed an active learning model based on the Bradley-Terry Model (Bradley & Terry, 1952) which adopts an efficient online Bayesian updating scheme and does not require retraining of the whole model when new observations are obtained. However, this process can still be inefficient in a crowdcourced setting, where multiple subjects may work in parallel and workers may expect to work on multiple tests instead of making only one judgement in each working session. Therefore, it is desirable to develop a batch-mode active learning method for information aggregation in a crowd-sourced setting.

(Gleich & Lim, 2011) introduced several ad-hoc methods for building a preference matrix from rating based on: arithmetic mean of score differences, geometric mean of score ratios, binary comparison, strict binary comparison and logarithmic odds ratio. However, this process may introduce some information loss. Our method combines absolute judgements and preference judgements directly via a unified probabilistic model.

3. Combining Absolute Judgements and Preference Judgements

Throughout this paper, we denote the set of objects as $A_1, A_2, .., A_n$ and let $s = (s_1, s_2, ..., s_n)$ represent the underlying scores of n objects. We model a subject's perceived score of object A_i as a random variable: $r_i = s_i + \varepsilon_i$, where the noise term is a Gaussian random variable $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We refer the test for obtaining absolute judgements as rating test and the test for obtaining preference judgements as preference test.

In the following of this section, we first derive the likelihood functions of the underlying score given absolute judgements and preference judgements independently. We then present the a hybrid system which estimates the underlying score by Maximum A Posteriori Estimation (MAP).

3.1. Absolute Judgements

Assume the perceived categorical observation is m_i and $m_i \in \mathcal{M}$, where $\mathcal{M}$ is a finite set of K ordered categories. Without loss of generality, these categories are denoted as consecutive integers: $\mathcal{M} = \{1, 2, ..., K\}$. We further introduce a set of cutoff values $-\infty \equiv \gamma_0 < \gamma_1 < \gamma_2 < ... < \gamma_{K-1} < \gamma_K \equiv \infty$. When r_i falls between the cutoffs γ_{c-1} and γ_c, the observed categorical label is c, i.e. $m_i = c$ and we have

$$\begin{aligned} &Pr(m_i|s_i) = Pr(\gamma_{m_i-1} < s_i + \varepsilon_i \leq \gamma_{m_i}) \\ &= \Phi(\tfrac{\gamma_{m_i} - s_i}{\sigma}) - \Phi(\tfrac{\gamma_{m_i-1} - s_i}{\sigma}) \end{aligned} \quad (1)$$

where $\Phi(\cdot)$ represents Cumulative Density Function (CDF) of standard Gaussian distribution.

In the rating test, repeated observations are made for each object. We define the rating observation matrix M as follows:

$$M = \begin{pmatrix} M_{1,1} & M_{2,1} & \cdots & M_{n,1} \\ M_{1,2} & M_{2,2} & \cdots & M_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1,K} & M_{2,K} & \cdots & M_{n,K} \end{pmatrix} \quad (2)$$

where $M_{i,j}$ is the number of times the object A_i is observed as in the j-th category. Given the underlying score s, we

assume the categorical observations of each object are conditionally independent. We then have the probability of observing M as follows:

$$\begin{aligned} Pr(M|s) &= \prod_{i=1}^{n} Pr(M_{i,1}, M_{i,2}, ..., M_{i,K}|s_i) \\ &= \prod_{i=1}^{n} \binom{M_{i,1}+...+M_{i,K}}{M_{i,1},...,M_{i,K}} \prod_{k=1}^{K} Pr(m_i = k|s_i)^{M_{i,k}} \\ &= c_1 \prod_{i=1}^{n} \prod_{k=1}^{K} (\Phi(\tfrac{\gamma_k - s_i}{\sigma}) - \Phi(\tfrac{\gamma_{k-1} - s_i}{\sigma}))^{M_{i,k}} \end{aligned} \tag{3}$$

where c_1 is a constant.

3.2. Preference Judgements

In a pairwise comparison experiment, if the perceived score $r_i > r_j$, we say that A_i is preferred to A_j, which is denoted as $A_i \succ A_j$. The probability of $A_i \succ A_j$ is given by:

$$\begin{aligned} Pr(A_i \succ A_j) &= Pr(s_i + \varepsilon_i > s_j + \varepsilon_j) \\ &= \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma}) \end{aligned} \tag{4}$$

Eq. 4 is known as the Thurstone-Mosteller Case V model (Thurstone, 1927; Mosteller, 1951). Preferences obtained from a set of paired comparison experiments can be characterized by a preference matrix. We define the preference matrix P as follows:

$$P = \begin{pmatrix} 0 & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & 0 & \cdots & P_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & 0 \end{pmatrix} \tag{5}$$

where $P_{i,j}$ is the number of times $A_i \succ A_j$ is observed. Then the probability of observing P is:

$$\begin{aligned} Pr(P|s) &= \prod_{i,j \in 1,...,n, i<j} Pr(P_{i,j}, P_{j,i}|s_i, s_j) \\ &= \prod_{i<j} \binom{P_{i,j}+P_{j,i}}{P_{i,j}} Pr(A_i \succ A_j)^{P_{i,j}} Pr(A_j \succ A_i)^{P_{j,i}} \\ &= \prod_{i<j} \binom{P_{i,j}+P_{j,i}}{P_{i,j}} \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma})^{P_{i,j}} \Phi(\tfrac{s_j - s_i}{\sqrt{2}\sigma})^{P_{j,i}} \\ &= c_2 \prod_{i \neq j} \Phi(\tfrac{s_i - s_j}{\sqrt{2}\sigma})^{P_{i,j}} \end{aligned} \tag{6}$$

where c_2 is a constant.

3.3. Posterior Probability of the Underlying Score

Given both absolute and preference judgements, the hybrid system finds the estimate of underlying score by maximizing the posterior probability

$$\hat{s} = argmax_s Pr(s|P, M)$$

Computing $Pr(s|P, M)$ is not a trivial task. The likelihood functions in Eq. 3 and Eq. 6 are conditioned on several unknown model parameters including: noise variance σ and cut-off parameters $\gamma_1, ..., \gamma_{K-1}$. Since the likelihood functions are scale-invariant, i.e. $Pr(M|s, \gamma, \sigma) = Pr(M|ks, k\gamma, k\sigma)$ and $Pr(P|s, \sigma) = Pr(P|ks, k\sigma)$ for a constant $k \neq 0$, without loss of generality, we may fix $\sigma = 1/\sqrt{2}$. With σ fixed, the likelihood functions are still translation-invariant, i.e. $Pr(M|s, \gamma) = Pr(M|s + k, \gamma + k)$ and $Pr(P|s) = Pr(P|s + k)$ for a constant k. To make the objective identifiable, we further assume $\gamma_1 = 0$. $K - 2$ model parameters $\gamma_2, ..., \gamma_{K-1}$ remain unknown. We denote the set of unknown model parameters $\gamma = \{\gamma_2, ..., \gamma_{K-1}\}$.

In a full Bayesian treatment, computing $Pr(s|P, M)$ requires integrating the model parameters over all possible values, which in practice can be implemented using Monte Carlo methods. However, these computations might be prohibitively expensive. Alternatively, we approximate $Pr(s|P, M)$ by $Pr(s|P, M, \hat{\gamma})$ where $\hat{\gamma}$ refers to the optimal setting of γ. Specifically, $\hat{\gamma} = argmax_\gamma Pr(M, P|\gamma)$, which is the Maximum Likelihood Estimate of γ. To obtain an analytical form of the gradients of $Pr(M, P|\gamma)$ w.r.t γ and a Gaussian form approximation to the posterior probability $Pr(s|M, P, \hat{\gamma})$, we apply Laplace approximation (Chu & Ghahramani, 2005). To illustrate the approximation procedure, let's define:

$$\mathcal{F}_\gamma(s) = -logPr(M|s, \gamma) - logPr(P|s) - logPr(s) \tag{7}$$

where we assume a Gaussian prior on s: $Pr(s) \sim N(\mu, \Omega)$. The Hessian matrix of $\mathcal{F}_\gamma(s)$ is given by:

$$R_\gamma(s) = \frac{\partial^2 \mathcal{F}_\gamma(s)}{\partial s \partial s^T} \tag{8}$$

Denoting the minimizer of $\mathcal{F}_\gamma(s)$ as $\hat{s}_\gamma$ and $\hat{R}_\gamma = R_\gamma(\hat{s}_\gamma)$, applying a Laplace Approximation, we have

$$\mathcal{F}_\gamma(s) \approx \mathcal{F}_\gamma(\hat{s}_\gamma) + \frac{1}{2}(s - \hat{s}_\gamma)^T \hat{R}_\gamma (s - \hat{s}_\gamma) \tag{9}$$

Using the above approximation, $Pr(M, P|\gamma)$ can be computed analytically as follows:

$$\begin{aligned} Pr(M, P|\gamma) &= \int Pr(s) Pr(M|s, \gamma) Pr(P|s) ds \\ &= \int exp(-\mathcal{F}_\gamma(s)) ds \\ &\approx exp(-\mathcal{F}_\gamma(\hat{s}_\gamma))(2\pi)^{\frac{n}{2}} |\hat{R}_\gamma|^{-1/2} \end{aligned} \tag{10}$$

Using Eq. 10, the gradients of the $log(Pr(M, P|\gamma))$ w.r.t γ can be computed analytically. Gradient-based optimiza-

tion methods can be used to find MLE of γ. The posterior probability of s can then be approximated by:

$$\begin{aligned} Pr(s|P,M) &\approx Pr(s|P,M,\hat{\gamma}) \\ &\propto Pr(M|s,\hat{\gamma})Pr(P|s)Pr(s) \\ &= exp(-\mathcal{F}_{\hat{\gamma}}(s)) \propto N(\hat{s}_{\hat{\gamma}}, \hat{R}_{\hat{\gamma}}^{-1}) \end{aligned} \tag{11}$$

The MAP estimate of s is $\hat{s}_{\hat{\gamma}}$. In order to ensure a global optimal solution of the MAP estimate, Eq. 7 has to be a convex function. It has been shown in (Chu & Ghahramani, 2005) that $-logPr(M|s,\gamma) - log(Pr(s))$ is convex. However, in order to make sure $-log(Pr(P|s))$ to have unique minimizer, Ford's condition (Ford, 1957) has to be satisfied. In practice, this can be achieved by adding a small constant τ to zero-valued elements in the preference matrix P. This is also known as smoothing. For the rating observation matrix M, we also add a small smoothing constant to zero-valued element in it.

4. Batch-mode Active Learning

In a crowd-sourced setting, subjective judgements are obtained at certain cost. It is therefore desirable to design cost-efficient experiments by applying an active learning strategy. Let $\mathcal{E}_i$ denote the experiment of making one absolute judgement on the object A_j and $\mathcal{E}_{ij}$ be the experiment of making a pairwise comparison on A_i and A_j.

4.1. Information Measure of Experiment

The purpose of experiments is to gain knowledge about the state of nature. We adopt the Bayesian Optimal Design framework introduced in (Lindley, 1956) and evaluate an experiment using the Expected Information Gain (EIG) provided by conducting this particular experiment. Suppose the state of nature (or parameters) to be estimated is θ. Before conducting the experiment $\mathcal{E}$, our knowledge on θ is characterized by the prior distribution of θ - $Pr(\theta)$. The EIG provided by an experiment $\mathcal{E}$ is denoted as $I(\mathcal{E}, Pr(\theta))$. The general formula of $I(\mathcal{E}, Pr(\theta))$ is given by (Lindley, 1956):

$$I(\mathcal{E}, Pr(\theta)) = E_\theta \left[\int log\{\frac{Pr(x|\theta)}{Pr(x)}\} Pr(x|\theta) dx \right] \tag{12}$$

Where $E_\theta(\cdot)$ is expectation taken w.r.t $Pr(\theta)$. For absolute judgement, suppose the outcome of $\mathcal{E}_i$ is $x_i \in \{1, 2, ..., K\}$. Let $\theta = s_i$ be the underlying score of A_i. Suppose $p_{ik} = P(x_i = k|\theta)$. It is easy to verify that $p(x_i = k) = E_\theta(p(x_i = k|\theta)) = E_\theta(p_{ik})$ and we have

$$\begin{aligned} I(\mathcal{E}_i, Pr(\theta)) &= E_\theta[\textstyle\sum_{k=1}^K p_{ik} log(\frac{p_{ik}}{p(x_i=k)})] \\ &= E_\theta[\textstyle\sum_{k=1}^K p_{ik} log(p_{ik})] - \sum_{k=1}^K E_\theta(p_{ik}) log E_\theta(p_{ik}) \end{aligned} \tag{13}$$

For preference judgements, suppose the outcome of $\mathcal{E}_{ij}$ is x_{ij} and $x_{ij} = 1$ if $A_i \succ A_j$; $x_{ij} = 0$ if $A_i \prec A_j$. Let $\theta = \{s_i, s_j\}$ be the underlying scores of A_i and A_j. Define $p_{ij} = p(x_{ij} = 1|\theta)$ and $q_{ij} = 1 - p_{ij}$. It is easy to verify that $p(x_{ij} = 1) = E_\theta(p(x_{ij} = 1|\theta)) = E_\theta(p_{ij})$ and $p(x_{ij} = 0) = E_\theta(q_{ij})$. The information gain provided by $\mathcal{E}_{ij}$ is:

$$\begin{aligned} I(\mathcal{E}_{ij}, Pr(\theta)) &= E_\theta[p_{ij} log(\tfrac{p_{ij}}{p(x_{ij}=1)}) + q_{ij} log(\tfrac{q_{ij}}{p(x_{ij}=0)})] \\ &= E_\theta[p_{ij} log(p_{ij}) + q_{ij} log(q_{ij})] \\ &- E_\theta(p_{ij}) log(E_\theta(p_{ij})) - E_\theta(q_{ij}) log(E_\theta(q_{ij})) \end{aligned} \tag{14}$$

Assuming that judgements on a pair objects are conditionally independent given the underlying score, it has been proved in (Glickman & Jensen) that the information gain obtained by a set of pair-wise comparisons is the sum of contributions from each pair. It is worth noting that the prior distribution $Pr(\theta)$ is actually conditioned on previous observations given by Eq. 11, but we omit the conditions here for ease of representation. In Eq. 11, we have introduced a Gaussian approximation to the posterior distribution. Therefore, we can use Gauss-Hermite quadrature (Press et al., 1992) to compute the expectation efficiently.

4.2. Batch Selection

In this section, we introduce a method for batch selection based on integer programming. Let Z_i and Z_{ij} denote the occurrence of experiments $\mathcal{E}_i$ and $\mathcal{E}_{ij}$ respectively. $Z_{ij} = 1$ ($Z_i = 1$) if $\mathcal{E}_{ij}$ ($\mathcal{E}_i$) is selected to be performed, otherwise $Z_{ij} = 0$ ($Z_i = 0$). [1] We formulate the batch selection problem as a binary integer programming problem for finding the optimal $\{Z_i, Z_{ij}\}$:

$$\begin{aligned} &Z^* = argmax \textstyle\sum_{i=1}^n Z_i I(\mathcal{E}_i) + \sum_{i<j} Z_{ij} I(\mathcal{E}_{ij}) \\ &s.t. \\ &(1) Z_i \in \{0,1\}, Z_{ij} \in \{0,1\}, i < j, i, j = 1, ..., n. \\ &(2)\ C_a \textstyle\sum_{i=1}^n Z_i + C_p \sum_{i=1}^{n-1} \sum_{j=i+1}^n Z_{ij} \leq C_{max} \\ &(3) Z_k + \textstyle\sum_{j:j<k} Z_{j,k} + \sum_{j:j>k} Z_{k,j} \leq 1, k = 1, ..., n \end{aligned} \tag{15}$$

where C_a and C_p denote the cost associated with each absolute judgement and each preference judgement respec-

[1] Z_{ij} and Z_{ji} represent the same pair. We may only consider Z_{ij} where $(i < j)$.

tively. C_{max} is the maximal cost assigned to each iteration of the experiment, $I(\cdot)$ represents the information gain. The first constraint $Z_{ij}(Z_i) \in \{0, 1\}$ ensures that a given experiment is performed at most once in each iteration, since duplicate experiments are redundant. The second constraint is to help control the cost of the experiment at each iteration, which is necessary to make sure the crowdsoucing experiment cost is within a defined budget. When $C_a = 1$ and $C_p = 1$, this constraint is also useful for specifying the batch size. The third constraint ensures that each object be observed only once in each iteration.

5. Experiments

5.1. Protocol

5.1.1. Evaluation Measure

Two evaluation measures were used, including the Spearman Rank Order Correlation Coefficient (SROCC) and Wilcoxon-Mann-Whitney statistics (ACC). SROCC is for evaluating how well the relationship between the predicted score and true score can be described using a monotonic function. ACC is primarily used for evaluating the ranking performance: $ACC = \sum_{i,j} I(y_i > y_j \wedge s_i > s_j) / \sum_{i,j} I(s_i > s_j)$ where y_i is the estimated score and s_i is the true score.

5.1.2. Implementation Details

We use Ipopt (Wchter & Biegler, 2006) for computing the MAP estimate of underlying score, i.e the minimizer of Eq. 7, which is a convex optimization problem. The prior distribution of the underlying score is specified by an uninformative prior defined $N(\mu, \Omega)$, where $\mu = \mathbf{0}, \Omega = 1000 \times \mathbf{I}$ and $\mathbf{I}$ is an identity matrix. Parameters in the second constraint in Eq. 15 are set to $C_a = 1, C_p = 1$. By default, the smoothing constant for both types of test is $\tau = 0.01$.

5.2. Simulations

In this section, we evaluate our method using synthetic data. We generated 20 random samples with underlying scores s uniformly distributed in the range $[-2.5, 2.5]$. The number of scales in the rating test is $K = 3$ and the cut-off parameters for generating the rating observations are $\gamma_0 = -\infty, \gamma_1 = -1, \gamma_2 = 1, \gamma_3 = \infty$.

We compare the hybrid active learning methods (**HY-ACT**) with three other methods: (1) **RATING**, which estimates underlying score based on rating test results and the estimated score of an object is the average of all rating observations of this object; (2) **PAIR**, where only preference tests are performed and MAP estimates are computed based on preference matrix and (3) **HYBRID**, where rating and preference tests are randomly selected and MAP is performed to find estimate of s. [2]

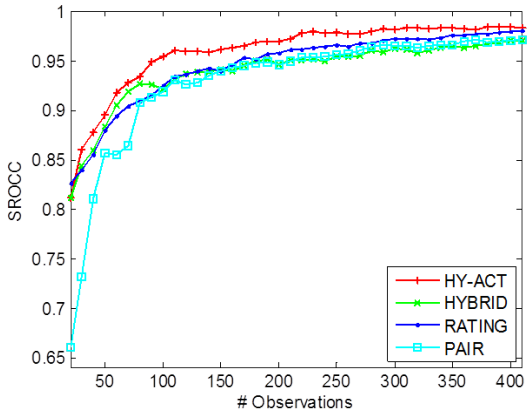

Figure 1. SROCC in Simulation.

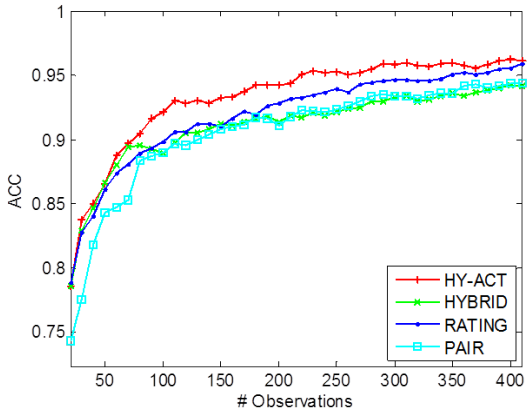

Figure 2. ACC in Simulation.

Each experiment is repeated 10 times and the average results are presented. In this experiment, we initialize all four methods by 20 rating tests. Fig. 1 and Fig. 2 show how SROCC and ACC of the four different methods increase as more tests are performed. Table 1 shows the number of tests required by each method for achieving a given level of SROCC. It can be seen that the hybrid active learning method outperforms the other three competing methods. While a hybrid system with random sampling does not outperform methods based only on rating or preference test.

Fig. 3 shows the number of different tests selected at each iteration of the experiment. In this particular case, our system prefers to select rating test, since when the noise level of the two types of test are the same, rating test is more informative than preference test.

[2]For RATING, PAIR and HYBRID, we randomly select C_{max} tests to perform at each iteration. These observations together with observations obtained from previous iterations are used to estimate s.

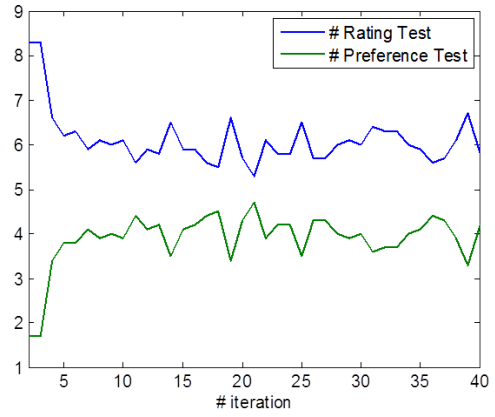

Figure 3. Number of selected tests at each iteration.

SROCC	RATING	PAIR	HYBRID	HY-ACT
0.97	267	330	388	126
0.96	203	243	323	104

Table 1. Number of samples required to achieve SROCC at given level.

5.3. Subjective Image Quality Assessment

In this section, we apply the proposed model to the task of subjective image quality assessment.

5.3.1. Dataset

In our experiment, we use a subset of 120 images from the Fast-Fading category in the LIVE IQA dataset (Sheikh et al., 2006). The 120 images include 20 non-distorted reference image and 100 distorted images derived from the 20 reference images. The groundtruth of image quality is obtained in a laboratory testing under controlled condition. We collected 1200 MOS observations (10 for each image) and a total of 21420 pairwise judgements (3 for each pair) from 182 distinct subjects on the Amazon Mechanical Turk platform. [3] In the MOS test, objects are labeled by five ordinal scales: "Bad", "Poor", "Fair", "Good" and "Excellent".

5.3.2. Results

When using all judgements obtained from MTurk for inference with a smoothing constant $\tau = 0.6$, we have $SROCC = 0.952$ and $ACC = 0.925$. This result serves as an upper bound for evaluating our active learning method.

For purpose of comparison, we implemented two heuristic methods (Gleich & Lim, 2011) in which the rating observations are converted to the preference matrices denoted as P_1 and the pairwise comparisons results in a preference matrix P_2, then the final preference matrix is given by $P = P_2 + P_1$. The MAP estimate is computed based on P. The first method (**H1**) computes object ranking based on the average of rating observations first, then the normalized rank difference is used to construct a preference matrix. The second method (**H2**) is based on the arithmetic mean of score differences.

[3] https://www.mturk.com/

	HYBRID	H1	H2
$N_r = 120, N_p = 50$	**0.842**	0.781	0.747
$N_r = 120, N_p = 120$	**0.852**	0.786	0.754
$N_r = 500, N_p = 500$	**0.928**	0.917	0.897
$N_r = 1000, N_p = 1000$	**0.942**	0.940	0.919

Table 2. SROCC given by the hybrid system and the heuristic method. N_r: number of rating test, N_p: number of preference test.

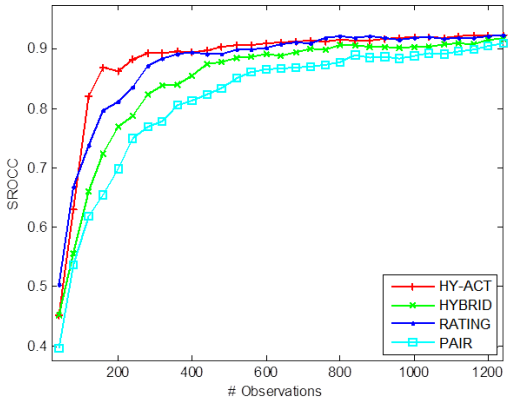

Figure 4. SROCC in Crowdsourcing experiment.

Table 2 shows SROCC averaged over 10 repeated experiments, where N_r rating observations and N_p preference observations are sampled from the real data. It can be seen that the hybrid method outperforms the heuristic method. In particular, when the number of observations is small, the difference is more significant.

To test the performance of the our active learning method, we simulate the active learning process by repeatedly sampling from the real judgements collected from MTurk. In this experiment C_{max} is set to 40. As is shown in Fig. 5 and Fig. 1, the active sampling based hybrid system (HY-ACT) outperforms the other three methods. However, the simple RATING based method is only slightly worse than HY-ACT, which implies that in this particular case, judgements obtained from MOS test are of relatively high quality.

6. Discussions and Conclusions

The proposed model assumes that the variances of observations noise for different objects in different types of test, given by different annotators are the same. This assump-

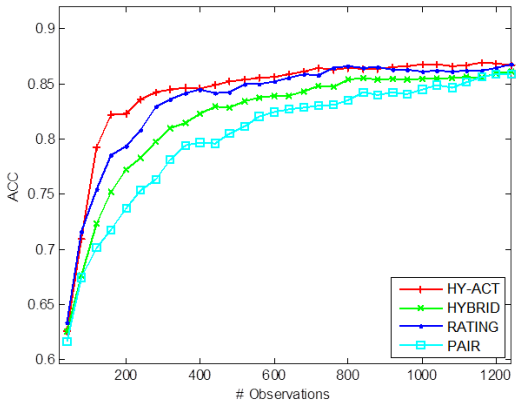

Figure 5. ACC in Crowdsourcing experiment.

tion does not necessarily hold in practice. Our future work will extend this model to take into consideration these factors. In our experiment, we have used an uninformative prior for the underlying score, however, when additional information about the underlying score is available, it can be easily incorporated into our model by constructing the prior distribution using prior information.

We have presented a hybrid system which combines absolute and preference judgements in an unified probabilistic model for estimating the underlying score of objects. Additionally, a batch-mode active learning method was proposed to efficiently construct queries of preference tests and rating tests which maximize the expected information gain. Experimental results show the effectiveness of the proposed method.

References

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Carterette, Ben, Bennett, Paul N., Chickering, David Maxwell, and Dumais, Susan T. Here or there: preference judgments for relevance. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pp. 16–27. Springer-Verlag, 2008.

Chen, Kuan-Ta, Wu, Chen-Chi, Chang, Yu-Chun, and Lei, Chin-Laung. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of ACM Multimedia*, 2009.

Chen, X., Bennett, P.N., Collins-Thompson, K., and Horvitz, E. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of WSDM*, 2013.

Chu, Wei and Ghahramani, Zoubin. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6: 1019–1041, Dec. 2005.

David, H.A. *The Method of Paired Comparisons*. Hodder Arnold, 1988.

Ford, L. R., Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

Gleich, David F. and Lim, Lek-heng. Rank aggregation via nuclear norm minimization. In *the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 60–68, New York, NY, USA, 2011. ACM.

Glickman, Mark E. and Jensen, Shane T. Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127:2005.

Lindley, D. V. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Mosteller, Frederick. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:3–9, 1951.

Pfeiffer, Thomas, Gao, Xi Alice, Mao, Andrew, Chen, Yiling, and Rand, David G. Adaptive polling and information aggregation. In *The 26th Conference on Artificial Intelligence*, 2012.

Press, William H., Flannery, Brian P., Teukolsky, Saul A., and Vetterling, William T. *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, October 1992.

Ribeiro, F., Florencio, D., Zhang, Cha, and Seltzer, M. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*, pp. 2416–2419, May 2011.

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

Thurstone, L.L. A law of comparative judgement. *Psychological Review*, 1927. 34:273-286.

Volkovs, Maksims N. and Zemel, Richard S. A flexible generative model for preference aggregation. In *International World Wide Web Conference*, 2012.

Wchter, Andreas and Biegler, Lorenz T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106:25–57, 2006.