Progressive Comparison for Ranking Estimation

Ryusuke Takahama,[1,2] **Toshihiro Kamishima,**[3] **Hisashi Kashima**[1]
[1]Kyoto University
[2]JST, ERATO, Kawarabayashi Large Graph Project
[3]National Institute of Advanced Industrial Science and Technology (AIST)
takahama@ml.ist.i.kyoto-u.ac.jp, mail@kamishima.net, kashima@i.kyoto-u.ac.jp

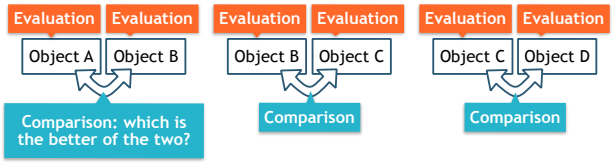
Abstract

Object ranking is a problem that involves ordering given objects by aggregating pairwise comparison data collected from one or more evaluators; however, the cost of object evaluations is high in some applications. In this paper, we propose an efficient data collection method called progressive comparison, whose objective is to collect many pairwise comparison data while reducing the number of evaluations. We also propose active learning methods to determine which object should be evaluated next in the progressive comparison; we propose two measures of expected model changes, one considering the changes in the evaluation score distributions and the other considering the changes in the winning probabilities. The results of experiments using a synthetic dataset and two real datasets demonstrate that the progressive comparison method achieves high estimation accuracy with a smaller number of evaluations than the standard pairwise comparison method, and that the active learning methods further reduce the number of evaluations as compared with a random sampling method.

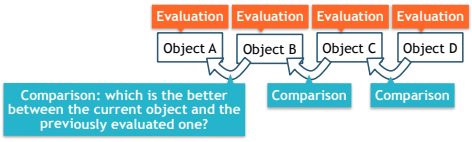
1 Introduction

Object ranking is a problem to order given objects by aggregating pairwise comparison data; It has various applications in optimizing Web services such as EC sites and search engines based on behaviors of customers or users. In sports or games played between multiple teams, match results during a particular period are aggregated into a team ranking list.

The pairwise comparison data is frequently discordant; when two players play chess, it naturally happens that one of the players wins in the first match and the other wins in the second match. In order to estimate a reasonable ranking list from such datasets, several existing approaches deal with such ambiguity using probabilistic models [Keener, 1993; Bradley and Terry, 1952; Glickman, 1999; Radlinski and Joachims, 2007; Elo, 1978; Herbrich *et al.*, 2006].



(a) Pairwise comparison



(b) Progressive comparison

Figure 1: (a) the standard pairwise comparison method and (b) the proposed progressive comparison method. In the standard pairwise comparison, an evaluator first evaluates two objects, and then submits the preferred one; therefore obtaining three pairwise comparison results requires six object evaluations. In the progressive comparison, the evaluator is asked if the current evaluated object is preferred to the previously evaluated object; therefore the evaluator needs only four object evaluations to obtain three pairwise comparison results, which means that the progressive comparison method can collect more comparison results with fewer evaluations than the standard pairwise comparison.

Estimation of a reasonably accurate ranking often requires many comparison results, and it is often costly to collect such datasets. Usually, an evaluator first evaluates two objects, and reports the preferred one to create one comparison result; to obtain K pairwise comparison results, we need $2K$ evaluations (Fig. 1(a)). In this paper, we especially consider the cases where the cost of an evaluation is much higher than that of a comparison; for example, ordering text documents by their quality, and ranking restaurants. An evaluation, i.e., reading a document or eating in a restaurant, is much more costly than comparing two of them. Moreover, increase in the number of evaluations also affects their quality and stability since the evaluator becomes tired or on a full stomach.

In this paper, we propose the *progressive comparison* method (Fig. 1(b)). Each time the evaluator makes an ob-

ject evaluation, the evaluator is requested to answer whether or not the current evaluated object is preferred to the previous one. In contrast with the standard pairwise comparison, the progressive comparison requires only $K + 1$ evaluations to obtain K pairwise comparison results.

We also propose *active learning* methods for the progressive comparison to further reduce data collection efforts. These methods select the object to be evaluated next based on the amount of information gained from the object evaluation and comparison. We propose two measures of such information gain: the *change in distributions* and the *change in winning probabilities*.

We present experiments conducted using a synthetic dataset and two real datasets in order to verify the cost effectiveness of the proposed methods. We show progressive comparison achieves higher accuracy with a smaller number of evaluations than the standard pairwise comparison, and our active learning methods more effectively estimate rankings than the method that randomly selects the next object.

In summary, this paper makes the following three contributions: (i) the progressive comparison method which is an efficient data collection method for pairwise comparison data with a small number of evaluations, (ii) active learning methods for the progressive comparison using two measures of expected model changes, and (iii) the property of datasets that the active learning methods applied in the progressive comparison method operate effectively.

2 Ranking Estimation Problem with High Evaluation Costs

We consider N objects $o_1, \ldots, o_N$, and denote by (o_i, o_j) a pairwise comparison result indicating that object o_i is preferred to object o_j. We also denote by C the multiset consisting of the comparison results. In many cases, the elements of C are discordant; in other words, C contains both (o_i, o_j) and (o_j, o_i). Our goal is to estimate an accurate ranking from C while keeping the collection cost of C as low as possible.

In this paper, we particularly focus on the cost in terms of the number of *evaluations*. An *evaluation* involves giving an (internal) score θ_i to an object o_i, for example, the beauty of an image, impressiveness of an article, or the deliciousness of food. A *comparison* involves determining the winner based on the two object scores to create one comparison result. Our important assumption is that the cost of an evaluation is substantially larger than that of a comparison; therefore, reducing the collection cost of C requires reducing the number of evaluations.

3 Progressive Comparison

3.1 Progressive Comparison

Linear ordering and *pairwise comparison* are two typical data formats in ranking estimation problems. In the linear ordering, each evaluator evaluates all objects $o_1, \ldots, o_N$, and submits a ranking list of the N objects. Since the evaluators have to remember the scores of all objects before creating the ranking list, this is difficult for a large N.

In the pairwise comparison, as shown in Figure 1(a), each evaluator evaluates two objects, and then reports the preferred one. For example, the evaluator receives objects o_i and o_j, and submits either (o_i, o_j) or (o_j, o_i). The evaluator needs only to remember the two objects to make a comparison. This method requires $2K$ evaluations to obtain K pairwise comparison results.

To reduce the evaluation costs of the standard pairwise comparison method, we propose the progressive comparison. As shown in Fig. 1(b), when an evaluator evaluates an object, she compares it with the preceding evaluated object and judges which one is preferred. This means that an evaluator first evaluates objects o_1 and o_2, and submits the comparison result for o_1 with o_2, and then evaluates object o_3 and submits the comparison result for o_2 with o_3. The evaluator repeatedly carries out the above operation. In order to compare K pairs of objects, this method requires only $K+1$ evaluations of objects. As compared with pairwise comparison, we can obtain more comparison results with fewer evaluations[1].

3.2 Ranking Estimation

We employ the Bradley-Terry model which is a popular probabilistic model for pairwise comparison [Bradley and Terry, 1952]. The Bradley-Terry model gives the probability an evaluator prefers object o_i to o_j as $P((o_i, o_j)) = \lambda_i \,/\, (\lambda_i + \lambda_j)$, where $\lambda_i, \lambda_j > 0$ are positive parameters that represent the object scores of o_i, o_j, respectively.

The glicko update equations are an approximate Bayesian estimation procedure for the Bradley-Terry model proposed by Glickman (1999). Following the original notations, we rescale λ_i as $\lambda_i = 10^{\theta_i/400}$ using the (scaled) object score θ_i. We assume that θ_i follows a Gaussian prior distribution $N(\mu_i, \sigma_i^2)$ where μ_i and σ_i^2 are the mean and variance parameters, respectively. In the progressive comparison scenario, when an object o_i is evaluated, θ_i is sampled from the Gaussian prior distribution $N(\mu_i, \sigma_i^2)$; then θ_j for the next object o_j is sampled from $N(\mu_j, \sigma_j^2)$, and a comparison result of o_j and o_i is determined with $P((o_i, o_j))$. Similarly, another object o_k is evaluated to obtain θ_k, which is then compared with θ_j to give another comparison result. Note that every time an object o_i is evaluated, θ_i is re-sampled from the corresponding Gaussian distribution.

The glicko update equations update parameters $\mu_i^{(o_i,o_j)}$ and ${\sigma_i^{(o_i,o_j)}}^2$ using comparison result (o_i, o_j) as follows:

$$\mu_i^{(o_i,o_j)} = \mu_i + \frac{q}{1/\sigma_i^2 + 1/\delta^2} g(\sigma_j^2)\{1 - E(\mu_i, \mu_j, \sigma_j^2)\},$$

$${\sigma_i^{(o_i,o_j)}}^2 = \left(\frac{1}{\sigma_i^2} + \frac{1}{\delta^2}\right)^{-1},$$

where $q = \log(10)/400$, $g(\sigma^2) = \left[\sqrt{1 + 3q^2\sigma^2/\pi^2}\right]^{-1}$, $E(\mu_i, \mu_j, \sigma_j^2) = \left[1 + 10^{-g(\sigma_j^2)(\mu_i - \mu_j)/400}\right]^{-1}$ and $\delta^2 =$

[1] One might wonder if the progressive comparison method is in fact a new idea; however, we are not aware of the method having been previously proposed in the literature.

$\left[q^2 g(\sigma_j^2)^2 E(\mu_i, \mu_j, \sigma_j^2)\{1 - E(\mu_i, \mu_j, \sigma_j^2)\}\right]^{-1}$. We repeatedly update each μ_i by using each comparison result in C, and finally obtain an object ranking list by sorting the conclusive $\{\mu_1, \ldots, \mu_N\}$ in descending order.

Note that since the results of pairwise comparisons obtained by the progressive comparison method are not i.i.d., the standard estimation methods are not applicable in the strict sense, which means that we approximate the likelihood or posterior by making the independence assumption. In the experimental section, we will see this approximation is effective in practice.

4 Active Learning for Progressive Comparison

4.1 Active Learning

To further reduce the number of evaluations and accelerate the estimation, we resort to active learning, which actively selects the unlabeled data to be labeled [Settles, 2012].

In active learning for the standard pairwise comparison, a learning algorithm actively selects two objects to be compared next. Conversely, an active learning algorithm in the progressive comparison selects one subsequent object. Let us assume that the preceding evaluated object is o_i. We first calculate u_{ij} for all j, where u_{ij} is the expected information gain by selecting o_j as the next object, and then o_{j^*} that maximizes u_{ij} is selected. In this study, we consider two types of information gain: *change in distributions* and *change in winning probabilities*.

4.2 Change in Distributions

When it turns out that an evaluator prefers object o_i to object o_j, the glicko update equations update the posterior distributions $N(\mu_i, \sigma_i^2)$ and $N(\mu_j, \sigma_j^2)$ to $N(\mu_i^{(o_i,o_j)}, {\sigma_i^{(o_i,o_j)}}^2)$ and $N(\mu_j^{(o_i,o_j)}, {\sigma_j^{(o_i,o_j)}}^2)$, respectively, where $\mu_i^{(o_i,o_j)}$, ${\sigma_i^{(o_i,o_j)}}^2$, $\mu_j^{(o_i,o_j)}$, and ${\sigma_j^{(o_i,o_j)}}^2$ are the parameters of the Gaussian distributions updated by assuming that an evaluator gives a comparison result (o_i, o_j). As shown in Fig. 2, a possible choice of our information gain is the amount of changes of the two Gaussian distributions, which is measured in the sum of the Kullback-Leibler (KL) divergences, that is,

$$\begin{aligned}\mathscr{F}((o_i, o_j)) =& D_{\mathrm{KL}}(N(\mu_i^{(o_i,o_j)}, {\sigma_i^{(o_i,o_j)}}^2) \parallel N(\mu_i, \sigma_i^2)) \\ &+ D_{\mathrm{KL}}(N(\mu_j^{(o_i,o_j)}, {\sigma_j^{(o_i,o_j)}}^2) \parallel N(\mu_j, \sigma_j^2)),\end{aligned}$$

where $D_{\mathrm{KL}}(P \parallel Q)$ is the KL divergence between two probability distributions P and Q.

Taking the expectation of $\mathscr{F}((o_i, o_j))$ over two cases (o_i, o_j) and (o_j, o_i), we define an expected information gain measure called *change in distributions* as

$$u_{ij} = P((o_i, o_j)) \times \mathscr{F}((o_i, o_j)) + P((o_j, o_i)) \times \mathscr{F}((o_j, o_i)),$$

where $P((o_i, o_j))$ is the probability that an evaluator prefers object o_i to o_j, which is readily calculated with

$$P((o_i, o_j)) = \int_{-\infty}^{\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right), \tag{1}$$

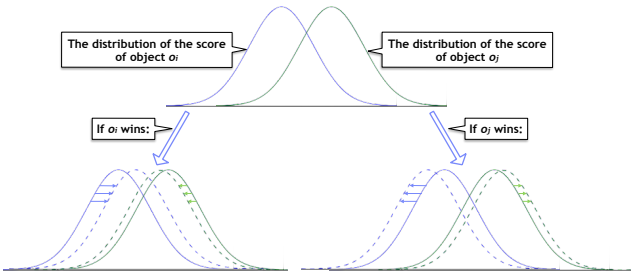


Figure 2: The information gain based on change in distributions. The information gain by comparing object o_i and object o_j is given as the expectation over two cases (o_i, o_j) and (o_j, o_i). For each case, the amount of change is measured by the sum of the KL-divergences between the Gaussian distributions of object scores.

where $\Phi(\cdot)$ is the cumulative Gaussian distribution [Mackay and Chaiy, 1982].

4.3 Change in Winning Probabilities

Since what we are interested in is the ranking list of objects which summarizes relative merits between any two objects, it seems a reasonable idea to consider the change in winning probabilities among objects as the information gain measure. Since a winner determination is a Bernoulli trial, such changes are measured in the KL divergences between Bernoulli distributions, that is,

$$\begin{aligned}&D_{\mathrm{KL}}\left(B(p_{xy}^{(o_i,o_j)}) \parallel B(p_{xy})\right) \\ &\quad = p_{xy}^{(o_i,o_j)} \log \frac{p_{xy}^{(o_i,o_j)}}{p_{xy}} + (1 - p_{xy}^{(o_i,o_j)}) \log \frac{(1 - p_{xy}^{(o_i,o_j)})}{(1 - p_{xy})},\end{aligned}$$

where $B(\cdot)$ denotes a Bernoulli distribution, and p_{xy} is the probability that an evaluator prefers object o_x to o_y, and $p_{xy}^{(o_i,o_j)}$ is the probability updated by the glicko update equations with a comparison results (o_i, o_j). Similar to Eq. (1), p_{xy} and $p_{xy}^{(o_i,o_j)}$ are readily calculated with

$$p_{xy} = \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right), p_{xy}^{(o_i,o_j)} = \Phi\left(\frac{\mu_x^{(o_i,o_j)} - \mu_y^{(o_i,o_j)}}{\sqrt{{\sigma_x^{(o_i,o_j)}}^2 + {\sigma_y^{(o_i,o_j)}}^2}}\right).$$

In contrast with measuring the change in two Gaussian distributions, changes in two Bernoulli distribution cause changes in all of the related winning probabilities, namely, they affect the winning probability one of whose objects is either o_i or o_j. Therefore, as shown in Figure 3, the total amount of changes in the winning probabilities is given as

$$\mathscr{G}((o_i, o_j)) = \sum_{\{(x,y)|x=i \text{ or } y=j\}} D_{\mathrm{KL}}\left(B(p_{xy}^{(o_i,o_j)}) || B(p_{xy})\right).$$

Now we define an expected information gain measure called *change in winning probabilities* defined as

$$u_{ij} = P((o_i, o_j)) \times \mathscr{G}((o_i, o_j)) + P((o_j, o_i)) \times \mathscr{G}((o_j, o_i))$$

where $P((o_i, o_j))$ is the probability that an evaluator prefers object o_i to o_j, which is calculated with Eq. (1).

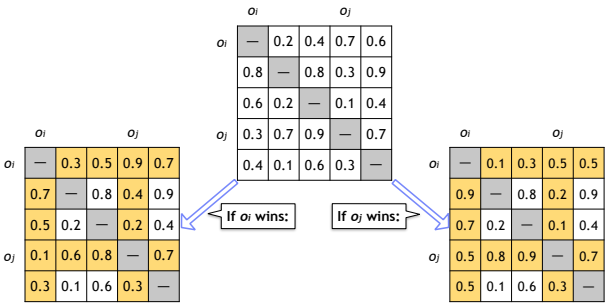


Figure 3: The information gain based on change in winning probabilities. The information gain by comparing object o_i and object o_j is given as the expectation over two cases (o_i, o_j) and (o_j, o_i). For each case, the amount of change is measured by the sum of KL-divergences between the Bernoulli distributions of winning probabilities.

5 Experiments

5.1 Experimental Settings

We conducted two experiments using a synthetic dataset and two real datasets to examine the effectiveness of our proposed approach. The first experimental result demonstrates that the progressive comparison requires fewer numbers of evaluations than the standard pairwise comparison method. The second result demonstrates that our proposed active learning approach is more efficient than the baseline method which randomly samples objects to be evaluated.

Throughout the experiments, the parameters are initialized $\mu_i = 1,500$ and $\sigma_i^2 = 147^2 = 21,609$ by following the settings of the glicko update equations [Glickman, 1999]. Every time the parameters are updated, we compare the current estimated ranking list with the true one. The estimated ranking list is obtained by sorting values $\mu_1, \ldots, \mu_N$ in descending order. The similarity between two ranking lists is measured using Spearman's rank correlation ρ [Spearman, 1906]. Let T and E be the rank vectors of the true ranking and the estimated ranking, respectively; ρ is given as $\rho = 1 - 6 \times d_{\mathrm{Spear}}(T, E) \,/\, (N^3 - N)$, where $d_{\mathrm{Spear}}(T, E)$ is the Spearman distance which is the squared Euclidean distance between T and E.

5.2 Datasets

The first dataset is a synthetic dataset comparing 100 objects. The true scores of object i is given as $\mu_i = i - 1$, and its true variance as $\sigma_i^2 = 100$. Each evaluation draws a score θ_i from the Gaussian distribution $N(\mu_i, \sigma_i^2)$, and the winner of o_i and o_j is determined by comparing θ_i and θ_j.

The image comparison dataset includes $25,500$ pairwise comparison results of 50 scenery images collected by using Lancers crowdsourcing marketplace[2]. Each crowdsourcing task presents ten pairs of images, and a crowdsourcing worker is asked to tell which picture of each pair is better for a postcard[3]. The true scores of the objects are defined as follows; we apply the glicko update equations to all of the pairs of the dataset in random order, and obtain final estimates. This procedure is repeated ten times, and the averaged final estimates are used as the ground truths of the ranking[4].

The Wikipedia article comparison dataset consists of $8,700$ pairwise comparison results of 30 Wikipedia articles. The dataset includes 15 featured articles and 15 non-featured articles selected randomly from "Wikipedia: Featured articles"[3]. Each crowdsourcing task presents five pairs of article links, and a crowdworker is asked to tell the more substantial article. Some exemplars are also presented in order to illustrate the evaluation criteria. Similar to the image comparison dataset, the average of ten sweeps of the whole dataset with different orders is used to create the ground truth ranking.

Figure 4 shows the histograms of the true object scores in the three datasets. Figure 4(a) corresponds to the synthetic dataset; the histogram is flat. On the other hand, the histogram for the image comparison dataset is unimodal (Fig. 4(b)). Figure 4(c) shows the histogram for the Wikipedia dataset; the objects with high scores and those with low scores are rather separated.

5.3 Results

Pairwise Comparison vs. Progressive Comparison

To compare the standard pairwise comparison and the proposed progressive comparison. we investigate the accuracy of the estimated ranking in terms of the numbers of comparisons and evaluations using the synthetic dataset.

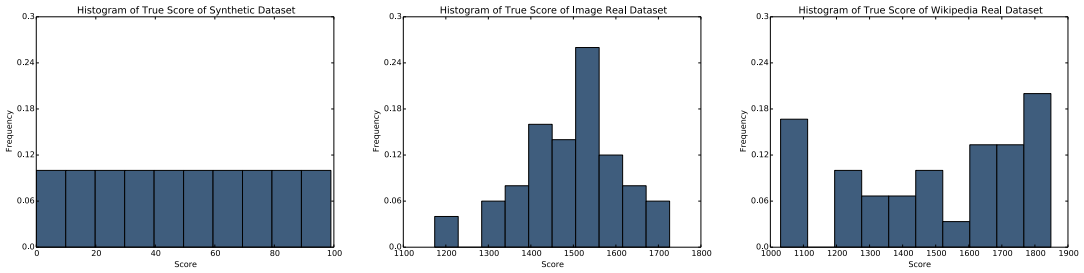
Figures 5(a) and 5(b) show the accuracy of the ranking versus the number of comparisons and evaluations, respectively. We denote the performance of the progressive comparison by 'Progressive' and the standard pairwise comparison by 'Pairwise.' We denote by 'Distribution' active learning using the change in distributions measure, 'WinProb' by the one using the change in winning probabilities measure, and by 'Random' the random sampling method.

Figure 5(a) shows the performance depending on the number of comparisons. The standard pairwise comparison method performs better than the progressive comparison in all of the random sampling method and two active sampling methods; this is because the progressive comparison has fewer options in selecting the next pair. In contrast, Figure 5(b) shows the progressive comparison method significantly outperforms the standard pairwise comparison in terms of the number of evaluations. This is because the standard pairwise comparison obtains $K/2$ comparison results by K evaluations, whereas the progressive comparison method can obtain $K - 1$ comparison results by K evaluations. As discussed earlier, the progressive comparison method collects comparison data in a non-i.i.d. manner, we need to approximate the likelihood or posterior by making the independence assumption. The result suggests this approximation is effective in practice. Due to the space limitation, we omit the results for the real datasets in Figure 5. We confirmed that they led to the same conclusion as the synthetic dataset.

[2]http://www.lancers.jp/

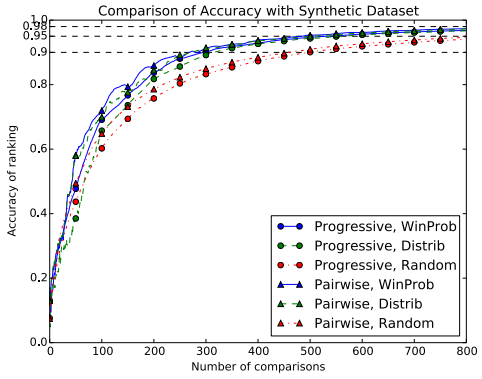
[3]The dataset is available at http://goo.gl/6MS9MK.

[4]The estimates were stable regardless of the presentation order.

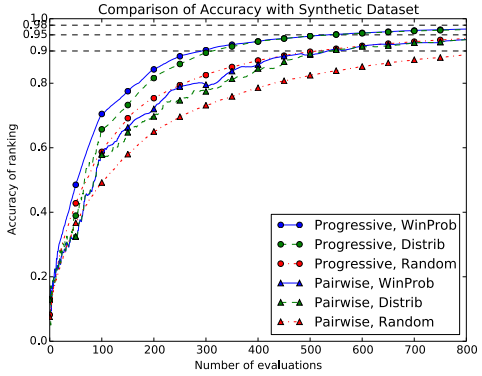


(a) Synthetic dataset (b) Image dataset (c) Wikipedia article comparison dataset

Figure 4: Histograms of the true object scores for three datasets.



(a) Estimation accuracy vs. number of comparisons



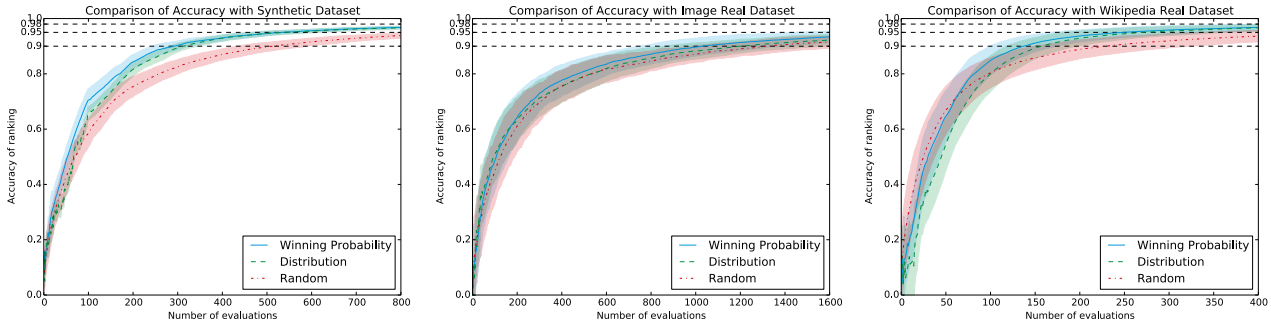
(b) Estimation accuracy vs. number of evaluations

Figure 5: Accuracy of ranking lists by the standard pairwise comparison and the progressive comparison for the synthetic dataset.

Efficiency of Active Learning Methods

We compare the proposed active learning methods using the two information gain measures: the change in distributions measure and the change in winning probabilities measure, with a baseline approach using uniform sampling. Figure 6 shows the comparison results of the accuracy of these methods. The accuracy of the estimated ranking lists is measured using Spearman's rank correlation between a true ranking list and an estimated ranking list. We denote the active learning method using the change in distributions measure by 'Distribution', and the one using the change in winning probabilities measure by 'WinProb', and the random sampling method by 'Random.' A solid line, a broken line, and a one-point broken line show the accuracy averages, and the transparently painted areas around these lines represent standard deviations. Figure 6(a) shows the result for the synthetic dataset. Our active learning methods obtain accurate ranking lists faster than the random method, and the method using the change in winning probabilities performs better than the one using the change in distributions. In terms of standard deviation, the area of the baseline method and those of active learning methods are clearly separated. However, for the image comparison dataset, the standard deviation areas of the three methods extensively overlap in Figure 6(b), and therefore we find no clear advantage of the proposed methods over the baseline method. The inconsistency of the two results is explained by the histograms of the true scores. Figure 4(a) shows the true scores of the synthetic dataset distribute uniformly, which means the true ranking is rather clear, and the proposed sampling methods are quite effective. On the other hand, Figure 4(b) shows the true scores of the image comparison dataset rather concentrate around the average, which means the dataset has many pairs whose winners are ambiguous, and the proposed sampling methods have no advantage in such datasets. Figure 6(c) shows the result for the Wikipedia dataset; the proposed methods outperform the baseline method especially after 200 evaluations, and there is no overlap between standard deviation areas. The method using the change in winning probabilities outperforms the one using the change in distributions when the number of evaluation is relatively small. In Fig. 4(c), the high-scored objects and the low-scored objects are separated in this dataset, that is, this dataset includes many pairs whose winners are readily determined as well as the synthetic dataset.

We also investigate the average number of evaluations required for Spearman's rank correlation to reach 0.9. The number of evaluations is the largest for the random sampling method and the smallest for the active sampling method using change in winning probabilities. The Wilcoxon signed-rank test [Wilcoxon, 1945] with 0.01 significance level shows statistically significant differences for all pairs of the methods except for the difference between the active sampling using



(a) Synthetic dataset (b) Image dataset (c) Wikipedia dataset

Figure 6: Comparison of active learning methods using two types of information gain measures, the "change in distributions" measure and the "change in winning probabilities" measure with a baseline approach using uniform sampling.

the change in distributions and the random sampling method in the image comparison dataset. The method using change in winning probabilities is superior to the random sampling method for all datasets with the statistical significance. In summary, the method using change in winning probabilities is the most promising to reduce the number of evaluations among the three methods.

6 Related Work

There has been much existing work in object ranking by aggregating the pairwise comparison results; to name a few, Keener proposed a method based on the eigenvector of the matrix that contains the comparison results [Keener, 1993]. Bradley and Terry proposed the Bradley-Terry model, which is a stochastic model of pairwise comparison, and its maximum likelihood estimation method [Bradley and Terry, 1952]. In addition, Glickman proposed an approximated Bayesian estimation method of the model [Glickman, 1999]. Based on Glickman's update equations, Radlinski and Joachims proposed an algorithm that estimates the Web page ranking ordered by relevance for a search query [Radlinski and Joachims, 2007]. The Elo rating system, which is a rating system for measuring the ability of chess players, was proposed by Elo [Elo, 1978]. Herbrich *et al.* proposed TrueSkill™ [Herbrich *et al.*, 2006], which is another rating system for estimating the ability of online game players, based on the Elo rating system.

To estimate a ranking list from the results of pairwise comparisons collected from crowdsourcing workers, the Crowd-BT model, which is an extension of the Bradley-Terry model to consider the ability of evaluators, was presented by Chen and Bennett [Chen and Bennett, 2013]. Matsui *et al.* proposed a method to aggregate linear orders collected by using crowdsourcing [Matsui *et al.*, 2014]. Wu *et al.* proposed a method that considers the reliability of labels annotated by multiple evaluators [Wu *et al.*, 2011].

There has been existing work on active learning from pairwise comparison data; for example, Pfeiffer *et al.* proposed adaptive information aggregation method for ranking estimation [Pfeiffer *et al.*, 2012]. The criterion that they employed for active sampling is essentially same as one we showed in Section 4.2. Brinker proposed active learning criteria [Brinker, 2004] for both the constraint classification [Har-Peled *et al.*, 2002] and the pairwise decomposition [Fürnkranz and Hüllermeier, 2003] which are approaches to reduce ranking problems to binary classification problems. Ailon discussed theoretical guarantees of active learning for ranking from pairwise preferences [Ailon, 2011].

7 Conclusion

We addressed the ranking estimation problem where the cost of object evaluation is large. We proposed a new pairwise comparison data collection method called progressive comparison, which reduces the number of evaluations to almost a half of that of the ordinary pairwise comparison method. We also proposed the active learning methods to determine which object should be evaluated next in the progressive comparison. We proposed two measures of expected model changes, one considering the changes in the evaluation score distributions and the other the changes in the winning probabilities. The experiments demonstrated that the progressive comparison achieved higher estimation accuracy with a smaller number of evaluations than the standard pairwise comparison, and the active learning methods further reduced the number of evaluations as compared with the random sampling method, especially for datasets with relatively clear true ranking.

In this study, we collected our crowdsourced datasets by using the standard pairwise comparison, and simulated the progressive comparison using them; however, it is still not clear if such experimental arrangement is always valid because evaluations in progressive comparison are possibly biased compared with the standard pairwise comparison. Further investigations regarding human cognition depending on sampling methods would be required.

References

[Ailon, 2011] Nir Ailon. Active Learning Ranking from Pairwise Preferences with Almost Optimal Query Complexity. In *Advances in Neural Information Processing Systems*, pages 810–818, 2011.

[Bradley and Terry, 1952] Ralph A. Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. In *Biometrika*, volume 39, pages 324–345. Biometrika Trust, 1952.

[Brinker, 2004] Klaus Brinker. Active Learning of Label Ranking Functions. In *Proceedings of the 21st International Conference on Machine Learning*, page 17. ACM, 2004.

[Chen and Bennett, 2013] Xi Chen and Paul N. Bennett. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 193–202, 2013.

[Elo, 1978] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*, volume 3. London: Batsford, 1978.

[Fürnkranz and Hüllermeier, 2003] Johannes Fürnkranz and Eyke Hüllermeier. Pairwise Preference Learning and Ranking. In *Proceedings of the 14th European Conference on Machine Learning*, pages 145–156. 2003.

[Glickman, 1999] Mark E. Glickman. *Parameter Estimation in Large Dynamic Paired Comparison Experiments*, volume 48, pages 377–394. Royal Statistical Society, 1999.

[Har-Peled *et al.*, 2002] Sariel Har-Peled, Dan Roth, and Dav Zimak. Constraint Classification: A New Approach to Multiclass Classification and Ranking. In *In Advances in Neural Information Processing Systems*, 2002.

[Herbrich *et al.*, 2006] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill^{TM}: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, pages 569–576. 2006.

[Keener, 1993] James P. Keener. The Perron-Frobenius Theorem and the Ranking of Football Teams. In *SIAM Review*, volume 35, pages 80–93. Society for Industrial and Applied Mathematics, 1993.

[Mackay and Chaiy, 1982] David B. Mackay and Seoil Chaiy. Parameter Estimation for the Thurstone Case III Model. In *Psychometrika*, volume 47. The Psychometric Society, 1982.

[Matsui *et al.*, 2014] Toshiko Matsui, Yukino Baba, Toshihiro Kamishima, and Hisashi Kashima. Crowdordering. In *Advances Knowledge Discovery and Data Mining*, pages 336–347. Springer International Publishing, 2014.

[Pfeiffer *et al.*, 2012] Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand. Adaptive Polling for Information Aggregation. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 122–128, 2012.

[Radlinski and Joachims, 2007] Filip Radlinski and Thorsten Joachims. Active Exploration for Learning Rankings from Clickthrough Data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 570–579. ACM, 2007.

[Settles, 2012] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

[Spearman, 1906] Charles Spearman. *'Footrule' for Measuring Correlation*, volume 2, pages 89–108. British Journal of Psychology, 1906.

[Wilcoxon, 1945] Frank Wilcoxon. Individual Comparisons by Ranking Methods. In *Biometrics Bulletin*, volume 1, pages 80–83. International Biometric Society, 1945.

[Wu *et al.*, 2011] Ou Wu, Weiming Hu, and Jun Gao. Learning to Rank under Multiple Annotators. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, volume 22, page 1571, 2011.