Final Essay: Solving the integration problem of mutual scoring

Zhuoyang Song

School of Computer Science and Engineering, Beihang University, Beijing, China
16061170@buaa.edu.cn

1 Introduction

In many cases, people need to evaluate the performance of each other. In this session, everyone score the performance of themselves and others which finally generate large amount of two-dimensional data. It is a basic method to compute the final score via simply averaging the origin data. However the result of this method is often in a low accuracy since the evaluation criteria of each person has great difference. In other words, different criteria cause that each sub-score have different weight while averaging. Thus it is necessary to transform the origin data to ensure the same contribution of each sub-score.

This essay introduces a statistical method to minimize the influence of noise mentioned above in the origin data and finally propose a specific function computing the score. Besides, in the experiment part, this essay uses a simple data set to implement the algorithms proposed in section 3. The results of experiment were analyzed in section 4.

2 Problem Analysis

In this section, we will have a detailed analysis about the integration problem of mutual scoring.

2.1 Noises

Criterion Difference When a person is to others rate, it is common that he or she can't know the scoring criteria of other people timely which causes mathematical features(e.g., mean and variance) different from each other. These differences make integrating score by computing average inaccurate. There are three different criteria in Table1

When integrating scores given under these three criteria, 'excellent' in criterion 3 might have greater contribution, 'poor' in criterion 1 might have positive contribution, and 'excellent' in criterion 1 might not have contribution. Therefore, it's essential to normalize the origin scores before integrating or analyzing.

Table 1. Criterion examples

No.	Excellent	Average	Poor
Criterion 1	95	90	85
Criterion 2	85	80	75
Criterion 3	95	80	65

Personal Bias Influence In many circumstances, people ought to do self-evaluation. It is unavoidable to be more or less biased since people are more likely to overestimate themselves. Besides people may also underestimate themselves when there is a lack of self-confidence. Weakening the influence of personal bias is of great value.

2.2 Problem Abstraction

In this essay, we abstract the integration problem of mutual scoring as Equation 1.

$$F(S_{m\times n}) = s_{1\times n} \tag{1}$$

$S_{m\times n}(n \leq m)$ is an m by n matrix containing origin scoring data given by m person to n person. $s_{1\times n}$ is an n-dimensional vector representing final score of n person. Function F processes the input matrix $S_{m\times n}$ using the algorithm in section 3.

What's more, we assume that everyone's score given is Gaussian in this paper. That is, at any particular time, true score can be approximated as the mean of scores given in similar criteria.

3 Algorithm

To weaken the noises mentioned in section 2, we propose two main statistical methods. This section will have a detailed introduction about these methods.

3.1 Criteria Normalization

In order to educe the impact of different evaluation criteria we use z-score normalization method as Equation2.

$$s^*_{i,j} = \frac{s_{i,j} - \mu_i}{\sigma_i} \tag{2}$$

$s_{i,j}$ is the element in the data matrix at location (i, j) representing the score of person j given by person i. μ_i is the mean of scores given by person i and σ_i is the mean of variance given by person i.

By doing z-score optimization, the evaluation criteria of each person can be transformed into same mean and variance which becomes more comparable. It's more accurate to integrate scores via averaging the data z-score-normalized.

3.2 Personal Bias Evaluation

There is a simple method to evaluate person bias via comparing the score that a person give to himself and to others. When there is a huge gap between self-evaluation and evaluation of others, the self-evaluation can be considered untrustworthy. However, this method is often lack of accuracy. It's a more accurate evaluate method is by comparing the score given by himself and by others. For instance, when a person really has a excellent performance, it's acceptable to have higher self-evaluation which is probably considered untrustworthy.

This essay uses confidence interval(Equation 3) to determine whether self-evaluation is trustworthy.

$$Pr(c_1 \leq \mu \leq c_2) = 1 - \alpha \tag{3}$$

At this stage, all data has been normalized with the method mentioned before. We firstly assume that everyone's score given is in a Gaussian distribution. Using the score given by other people, we can fit a Gaussian distribution and compute the confidence interval in a precise credibility which is set as 90 percent in this essay. This interval will be used as a criterion for judging the person bias.

4 Experiment

This section will have a detailed introduction about data preprocessing and experiment result. This essay uses a real data set to implement the algorithms proposed in section 3.

4.1 Preparing

The real data consist of large amount of missing which increases the difficulty of implementation. Before starting analyzing, it's essential to preprocess the origin data. For different kinds of missing data, we propose three kinds of treatment options:

1. Discard all missing data
2. Discard partial missing data
3. Keep all data

If the missing data is retained in the data set, it's necessary to do the data completion. Since the data completion may introduce new errors into the computation, more missing data means lower accuracy. In this essay, we choose the first and the second options as the pretreatment strategy.

4.2 Result

The result of experiment is in Figure 4.2 using two different pretreatment strategies. In order to show the effect more clearly, the change of ranking is shown in Table 2 and Table 3.

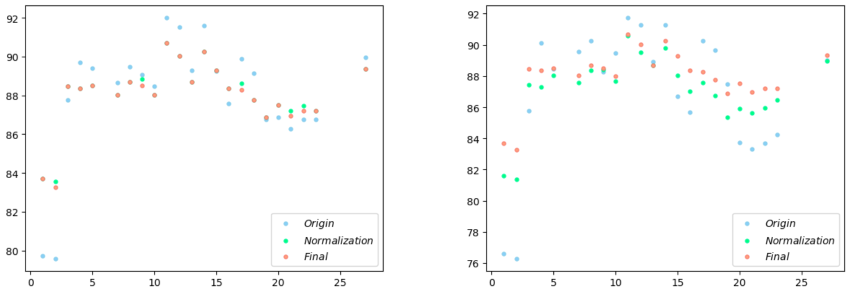

Fig. 1. Left: Discard all missing data; Right: Discard partial missing data

In the figures, x-axis represents student number and y-axis represents the score of each student. The blue points represent the origin score. The orange points represent the score calculated by only normalizing. The green points represent the score calculated using all methods in section 3.

Table 2. Ranking(discarding all missing data)

No.	Origin	Normalization	Final
1	22	22	22
2	23	23	23
3	15	11	10
4	6	12	11
5	8	10	8
7	13	14	14
8	7	7	6
9	12	6	9
10	14	15	15
11	1	1	1
12	3	3	3
13	9	7	6
14	2	2	2
15	10	5	5
16	16	13	12
17	5	9	13
18	11	16	16
19	18	21	21
20	17	17	17
21	21	20	20
22	18	18	19
23	18	19	18
27	4	4	4

Table 3. Ranking(discarding partial missing data)

No.	Origin	Normalization	Final
1	22	22	22
2	23	23	23
3	16	13	10
4	6	14	11
5	12	8	8
7	8	11	14
8	4	7	6
9	13	6	9
10	9	10	15
11	1	1	1
12	2	3	3
13	11	5	6
14	2	2	2
15	15	9	5
16	17	15	12
17	4	11	13
18	7	16	16
19	14	21	21
20	19	19	17
21	21	20	20
22	20	18	19
23	18	17	18
27	10	4	4

After doing normalization, some students' rankings have changed. At the stage of personal bias evaluation, student 9, 17, 21, 22 are recognized selfish giving untrustworthy excessive self-evaluation. And after tuning this, we finally have a trustworthy scores.

5 Conclusion

In this essay, we propose two statistical methods to minimize the influence of noise mentioned above in the origin data and finally propose a specific function computing the score. At the experiment part, we use a real mutual scoring data as example. After implementing the statistical methods and processing the origin data, we finally obtain more accurate score of each student.

What's more, the source code is uploaded to GitHub.