

Computing Marks from Multiple Assessors Using Adaptive Averaging

Aleksandar Ignjatovic, Chung Tong Lee, Cat Kutay,
Hui Guo, Paul Compton

School of Computer Science and Engineering,
University of New South Wales,
Sydney, Australia

Email: {ignjat, ctle, ckutay, huig, compton}@cse.unsw.edu.au

Abstract

Consider a situation in which a group of assessors mark a collection of submissions; each assessor marks more than one submission and each submission is marked by more than one assessor. Typical scenarios include reviewing conference submissions and peer marking in a class. The problem is how to optimally assign a final mark to each submission. The mark assignment must be robust in the following sense. A small group of assessors might collude and give marks which significantly deviate from the marks given by other assessors. Another small group of assessors might give arbitrary marks, uncorrelated with the others' assessments. Some assessors might be excessively generous while some might be extremely stringent. In each of these cases, the impact of the marks by assessors from such groups has to be appropriately discounted. Based on the work in [2], we propose a method which produces marks meeting the above requirements. The final mark assigned to each submission is a weighted average of marks by individual assessors; the weight given to each assessor's mark is inversely related to the total variance of all his marks from the final marks. Clearly, such definition is circular, and the existence of a final mark assignment having such a property is proved using the Brouwer Fixed Point Theorem for continuous maps on convex compact sets [1]. We provide a fast converging iterative algorithm for computing such a fixed point and give results of empirical tests of the robustness and adequacy of the marks calculated by our algorithm.

1. Introduction

Assume that M assessors $\{a_i: 1 \leq i \leq M\}$ are marking N submissions $\{s_j: 1 \leq j \leq N\}$. Each submission is marked by at least two (preferably more) assessors, and each assessor marks at least two (preferably more) submissions. A typical example is the common reviewing process of submissions for a large conference. The assessors might also be the students submitting their assignments, i.e., our method applies to peer marking as well.

As it is often the case in practice, assessors might not have uniform criteria; some might consistently be tougher, some are more generous with their marks; some might mark erratically, allowing a large random component in their marks. Further, in case of peer marking, there might be collusions of smaller groups, giving members of the colluding group higher marks than warranted, and to everyone else low marks.

The aim of this paper is to show how adaptive averages can be used to design a marking procedure which is robust with respect to:

- discrepancies in the strictness of marking criteria of individual assessors,
- influence of collusion of smaller groups (in case of peer marking), and
- presence of assessors with somewhat arbitrary (random) marking practice.

The method should also:

- detect which assessors give anomalous marks, and indicate the nature of the anomaly;
- be reasonably efficient and allow significant number of both assessors and submissions.

In our procedure, final marks are obtained as a fixed point of a weighted average operator, with weights assigned to marks given by each assessor reflecting the variance of the marks of that assessor from the finally assigned marks. Our procedure satisfies all of the above criteria, as our simulations and empirical testing on actual data show.

2. Basic Notations

For each assessor a_i , let D_i be the domain of the set of indices of all submissions that a_i has marked. For each submission s_j such that $j \in D_i$, let $m(i, j)$ be the mark given by a_i to s_j . Let also G_j be the set of indices of all assessors that have marked s_j . All marks are non-negative real numbers in a bounded range, i.e., there exists $R > 0$ such that $0 \leq m(i, j) \leq R$ for all i, j where $1 \leq i \leq M, 1 \leq j \leq N$.

For the purpose of analysis, let us consider an unspecified marking method \mathcal{M} . Denoting by μ_j the mark assigned by \mathcal{M} to the submission s_j , we define two metrics for variance of an assessor a_i from the assigned marks:

$$\begin{aligned} v(i) &= \frac{1}{\|D_i\|} \sum_{j \in D_i} |m(i,j) - \mu_j|^p \\ sv(i) &= \frac{1}{\|D_i\|} \sum_{j \in D_i} \text{sgn}(m(i,j) - \mu_j) \cdot |m(i,j) - \mu_j|^p \end{aligned} \quad (1)$$

where $\|D_i\|$ denotes the number of submissions marked by a_i ;
 $p \geq 1$ is a real parameter; and

$$\text{sgn}(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases} .$$

Thus, the variance metric $v(i)$ takes into account only the absolute values of the differences between μ_j (assigned by \mathcal{M}) and $m(i,j)$ (given by a_i), while the variance metric $sv(i)$ retains the sign of these differences. If $v(i)$ is large and absolute value of $sv(i)$ is small, the assessor a_i has a high degree of arbitrariness, because his marks are often and equally likely excessively high and excessively low. If both $v(i)$ and $sv(i)$ are of high positive values, a_i tends to be excessively generous compared to the marking method \mathcal{M} , while a large value of $v(i)$ and a large negative value of $sv(i)$ indicates that a_i is a harsh assessor. The higher the value of the parameter p , the more contributing to the sum the large differences become.

Let $\vec{\mu} = \langle \mu_i : 1 \leq i \leq M \rangle$ and let $q \geq 1$ be another real parameter. Consider a submission s_j ; we define a weight function $w(i, j, \vec{\mu})$ for all $i \in G_j$ and then a weighted average $f_j(\vec{\mu})$ of marks $m(i, j)$ given to s_j by a_i .

$$\begin{aligned} w(i, j, \vec{\mu}) &= \frac{\left(1 - \frac{v(i)}{\sum_{k \in G_j} v(k)}\right)^q}{\sum_{l \in G_j} \left(1 - \frac{v(l)}{\sum_{k \in G_j} v(k)}\right)^q} \\ f_j(\vec{\mu}) &= \sum_{i \in G_j} w(i, j, \vec{\mu}) \cdot m(i, j) \end{aligned} \quad (2)$$

If we interpret μ_j as the final mark assigned to s_j , then for a fixed j , $f_j(\vec{\mu})$ is a weighted average of marks $\{m(i, j) : i \in G_j\}$.

Since $0 \leq v(i) \leq \sum_{k \in G_j} v(k)$, we have $0 \leq 1 - \frac{v(i)}{\sum_{k \in G_j} v(k)} \leq 1$. Because all these values can be

close to one, we need to use a “spreading function” to emphasize the variances of values of $1 - \frac{v(i)}{\sum_{k \in G_j} v(k)}$ and taking the term to a power of q serves the purpose. In our experiments, taking q with order of magnitude equal to the average size of the sets G_j worked very well. Finally, to obtain the correct weighted average, we must normalize weights so that for each s_j , $\sum_{i \in G_j} w(i, j, \vec{\mu}) = 1$. This explains the form of the weight formulas (2). Note that for two different submissions s_{j_1} and s_{j_2} , the corresponding weights $w(i, j_1, \vec{\mu})$ and $w(i, j_2, \vec{\mu})$ for the same assessor a_i may be different, because the sets of G_{j_1} and G_{j_2} might not be the same.

The sum $f_j(\vec{\mu})$ can be seen as an adaptive weighted average of all marks given to s_j , in the sense that the weight $w(i, j, \vec{\mu})$ assigned to the marks of an assessor a_i is inversely related to his share $v(i)$ in the total variance of all assessors who marked s_j . If the values μ_j satisfy $f_j(\vec{\mu}) = \mu_j$ for all $j: 1 \leq j \leq N$, we will have precisely the desired properties of the marking system. Namely that the impact on the final marks of marks assigned by assessors with larger variance in their marking is appropriately diminished. The weights reflect appropriately the reliability of the corresponding assessor. Consequently, if we define the operator $F(\vec{\mu}) = (f_j(\vec{\mu}))_{1 \leq j \leq N}$, $\vec{\mu}$ should be a fixed point of the operator F in the hypercube $[0, R]^N$, i.e., $F(\vec{\mu}) = \vec{\mu}$ with $0 \leq \mu_j \leq R$.

If all assessors of a submission s_j give the same mark, this mark is assigned to s_j as μ_j . Thus, we can assume that for every $1 \leq j \leq N$ there are at least two assessors a_{i_1} and a_{i_2} such that $m(i_1, j) \neq m(i_2, j)$. Note that for all such j the denominators in equations (2) are non zero, regardless of the value of μ_j ; consequently, all weights given by (2) are well-defined and for all $j \leq M$:

$$\min_{i \in G_j} \{m(i, j)\} \leq f_j(\vec{\mu}) = \sum_{i \in G_j} w(i, j, \vec{\mu}) \cdot m(i, j) \leq \max_{i \in G_j} \{m(i, j)\}. \quad (3)$$

The operator F maps a compact and convex subset of \mathbb{R}^N , namely the N -dimensional cube $[0, R]^N$, into itself, and this mapping is continuous. Thus, by the Brouwer Fixed Point Theorem F has a fixed point in $[0, R]^N$, which produces the assignment of marks which satisfy the appropriate weighted average equations.

3. Experiments

The system of equations $\{f_j(\vec{\mu}) = \mu_j: 1 \leq j \leq N\}$ is a set of polynomial equations, and it can be solved efficiently using standard iterative procedures, with the simple arithmetic mean of the given marks $m(i, j)$ as the starting point for iterations. In our tests, we use Wolfram’s

Mathematica software package to obtain a fixed point solution. The results of one of our tests are presented in Figure 1.

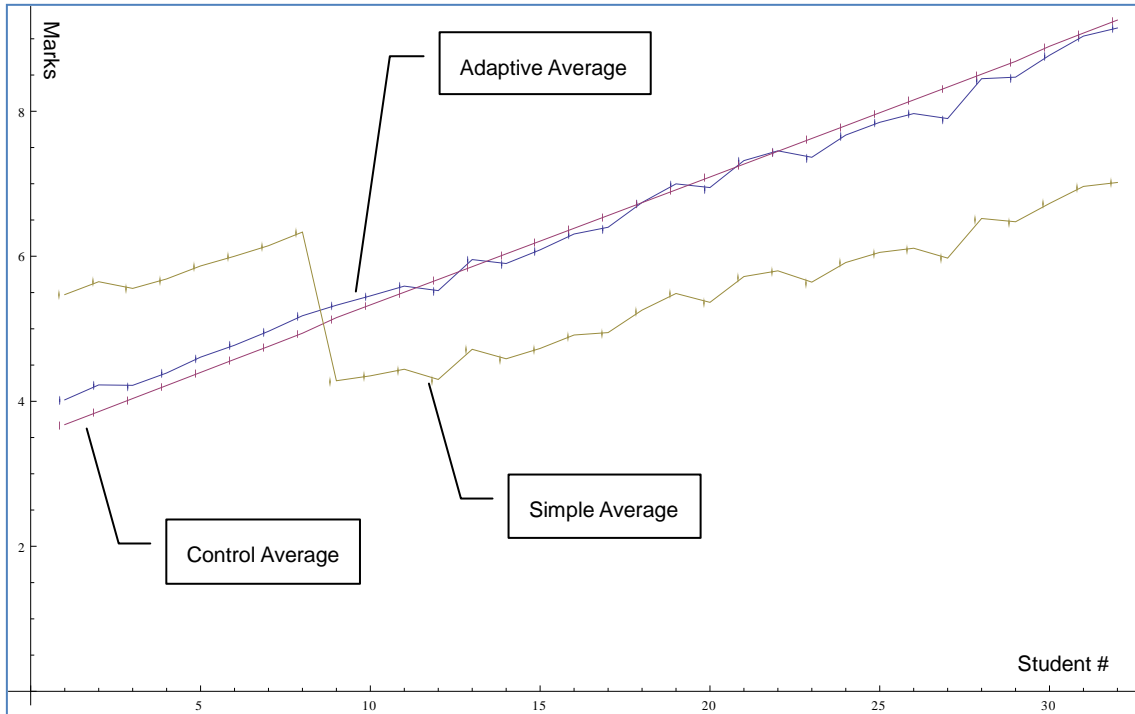


Figure 1 Simulation Result of Assessments of 32 Students

We simulated a situation of a class of 32 students. Each submission was marked by all students except the author. The range of marks is from 1 to 10. The students are arranged so that the higher the student number, the better the ability. The first eight students are weakest and they colluded by giving each other 10 points and everyone else 1 point. In addition to being the students with highest ability, the last three were also lazy assessors and assigned marks using a random number generator. Thus, we argue that the fair marks should be the average of marks by all assessors, excluding those colluders and lazy ones. However, in practice it is hard to ascertain who colluded and who was careless in marking in order to eliminate the marks given by such assessors. Note that our algorithm does not require making any decisions regarding the quality of assessors, but instead relies on essentially self-adapting averaging method.

We can see that, for example, the first eight colluding students would have managed to significantly increase their marks, while also significantly reducing the marks of the best students, if the final marks were determined as simple averages of all given marks. However, as can be seen from the Figure 1, the algorithm dramatically reduces the effect of unworthy

marks, both in terms of reducing the benefit for the colluding students as well as reducing the impact on marks of good submissions.

After the system has been solved and the marks μ_j 's have been obtained, one can evaluate $v(i)$ and $sv(i)$ for all assessors a_i . From the statistics of these values, one can find assessors with large variance as well as groups that might have colluded.

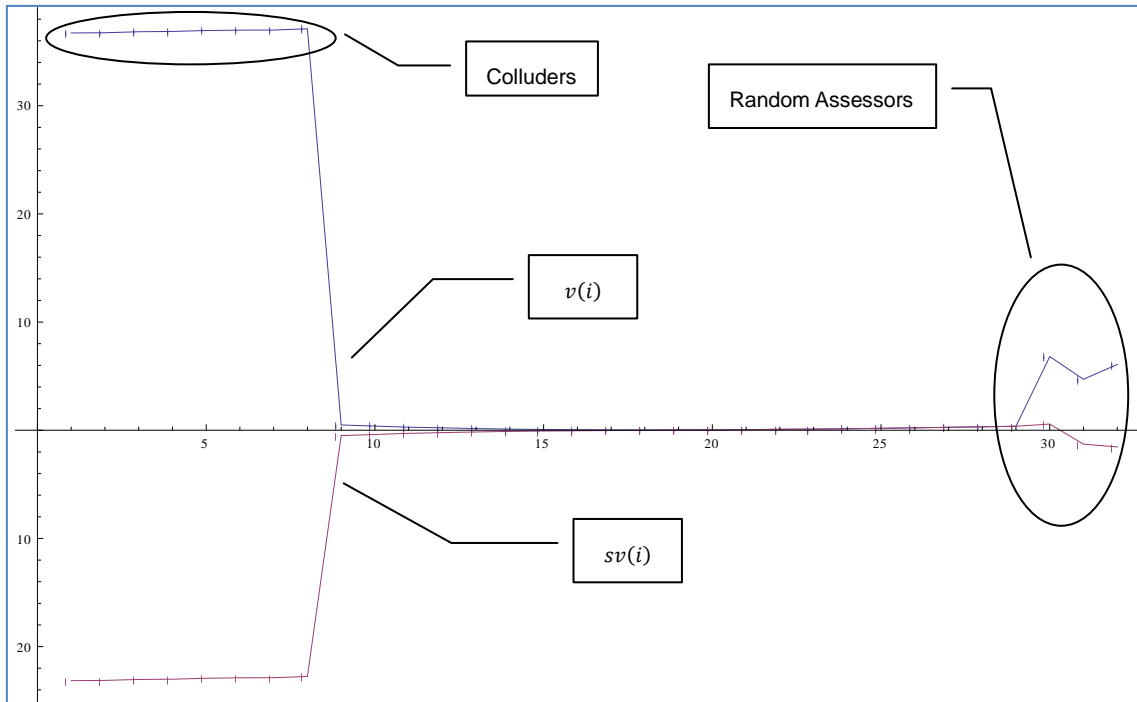


Figure 2 Variance Analysis

For peer marking, the weight equation can be modified as follows:

$$w(i, j, \vec{\mu}) = \frac{\left(1 - \frac{v(i)}{\sum_{k \in G_j} v(k)}\right)^q \cdot (\mu_i)^r}{\sum_{l \in G_j} \left(\left(1 - \frac{v(l)}{\sum_{k \in G_j} v(k)}\right)^q \cdot (\mu_l)^r\right)} \quad (4)$$

where r is a positive real number. This way we give more weight to assessors who will receive higher marks themselves, with the assumption that a higher mark reflects increased competence in the subject and thus more reliable marking.

4. Conclusions

The same procedure applies to various other scenarios that involve a data fusion process. For example, we might have a network of wireless temperature sensors that report their readings to a central processing unit for monitoring operation environment. Sensors might have variable accuracy over time, due to battery life, occasional loss of individual packets, intermittent exposure to direct sun etc. Our adaptive averaging method can be used to estimate temperature of the environment at a given moment from individual readings, while appropriately discounting data from sensors with large variance from the estimated values. Trust development can be regarded as a form of data fusion for experiences. Details about using adaptive averaging for trust and reputation evaluation can be found in [2].

The parameters p and q in equations (1) and (2) can be used to tune the method for different setups. For example, if each submission is marked by small number of assessors, taking larger p improves rejection of anomalous marks. Similarly, for a larger number of assessors, or if each assessor has marked small number of submission, taking a larger q improves the performance.

References

- [1] D. H. Griffel, *Applied Functional Analysis*, Dower Publications Inc., 1981.
- [2] A. Ignjatovic, N. Foo, C. T. Lee, An Analytic Approach to Reputation Ranking of Participants in Online Transactions, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.