https://doi.org/10.2352/EI.2024.36.11.HVEI-228

Aggregating Metric Values using Kumaraswamy Distribution: An Insight into User Experience Analysis

Lucjan Janowski, Natalia Cieplińska, Bogdan Ćmiel; AGH University of Krakow; Krakow, Poland

Abstract

This paper proposes a novel aggregation method using the Kumaraswamy distribution to analyze partial metric values, particularly in the evaluation of video quality. Through a weighted mean aggregation procedure, we unravel the underlying effects on the data. The three experiments analyzed in this paper demonstrates the method's efficacy regardless of the time aggregation, ranging from days, minutes, and frames. This approach, grounded in the Kumaraswamy distribution, offers a robust analytical tool to understand how individual metric values amalgamate, affecting overall user perceptions and experience.

Introduction

Modeling often requires the aggregation of partial values [1]. A quintessential example of this is found in the domain of Human Visual Systems, where the endeavor is to propose a metric that predicts video quality. Here, the scores acquired for each frame [2, 3, 4] must be aggregated to produce a single value per video or shot. A parallel challenge arises when aggregating metric values measured per second to represent a single user experience [5].

The aggregation procedure can be carried out through various aggregation operators [1]. Typically, algorithms are chosen to achieve the best fit, yet the rationale behind why a particular operator furnishes the most apt fit often remains elusive. Unlocking this mystery is particularly crucial when examining user behavior. For example, several psychological phenomena such as the primacy effect, the serial-position effect, and the recency effect have been delineated [6] that could influence the way aggregated values are perceived or analyzed.

In this paper, we propose employing the Kumaraswamy distribution [7] as a means to discern the presence and intensity of these effects in specific data sets. We subjected our proposed analysis to a subjective tests, in which we endeavored to predict user scores based on the quality scores assigned for each day, second, or frame.

Aggregation Procedure

We consider data represented as a vector of metric values m_i. Our objective is to employ an aggregation operator for synthesis. As exhaustively described in [1], a myriad of methods can be used for this purpose. One of the solutions in the video metric domain is presented in [4] where the authors propose the weighted mean without much explanation. We are also focused on the weighted mean, expressed by the equation:

$$\mathbb{C}_{\mathbf{w}}(m_1, \cdots, m_N) = \frac{\sum_{i=1}^{N} w_i m_i}{\sum_{i=1}^{N} w_i} \tag{1}$$

However, in this paper, the main focus is on explaining the result obtained. The explanation is achieved by using the Kumaraswamy distribution to delineate the function describing w_i. This distribution is characterized by two parameters, a and b, which dictate the shape of the distribution. Specific values of a and b derive a specific function, which can be analyzed to better understand the underlying process.

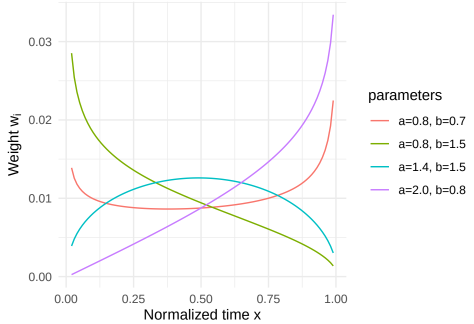

Figure 1. *Examples of different shapes obtained for 100 weights.*

Given that the Kumaraswamy distribution is defined over the interval $(0,1)$, an adaptation is required to extend it to indices ranging from 1 to N. Moreover, certain parameters of the Kumaraswamy distribution can yield arbitrarily high values near 0 or 1, which could potentially render the optimization algorithm unstable due to exceedingly large values. It is imperative to circumvent such occurrences while retaining the generalizability afforded by the Kumaraswamy distribution. Our analysis demonstrates that confining the focus to the interval $[0.01, 0.99]$ produces stable results. The mapping is thus expressed as:

$$w_i(a,b) = \frac{F_{a,b}(0.01 + 0.98\frac{i}{N}) - F_{a,b}(0.01 + 0.98\frac{i-1}{N})}{F_{a,b}(0.99) - F_{a,b}(0.01)} \tag{2}$$

where $F_{a,b}(x)$ denotes the CDF[1] of the Kumaraswamy distribution given by:

$$F_{a,b}(x) = 1 - (1 - x^a)^b \tag{3}$$

The weights generated by equation (2) can, depending on the parameters a and b, model how different the influence of metric m_i depends on i. The generated shapes can be related to the known psychology effects of recency, primacy, or position-effect [6] (see Figure 1).

The versatility of the Kumaraswamy distribution paves the way for diverse models. For example, $F_{0.8,1.5}(x)$ epitomizes the primacy effect, where the first values are given greater significance. On the contrary, $F_{2.0,0.8}(x)$ encapsulates the recency effect, but with a slightly stronger effect than the primacy effect

[1] Cumulative Density Function

described by $F_{0.8,1.5}(x)$ (see Figure 1). As presented in Figure 1 different shapes can be obtained. The exact interpretation of the importance of specific parts of the video or experience can be easily deduced based on the shape of the function. Moreover, it is easy to compare the shapes and strength of the effect obtained for different groups.

To find the optimal aggregation, we must find parameters a and b. We can do it by optimization of the equation:

$$\min_{a,b} \text{AIC}(U = \alpha\mathbb{C}_{a,b}(m) + \beta) \tag{4}$$

where AIC is an Akaike Information Criterion [8] of the $U = \alpha\mathbb{C}_{a,b}(m) + \beta$ model, α and β are model parameters estimated for each a and b, $\mathbb{C}_{a,b}(m)$ is an aggregation operator give by equation (1) where w_i is given by equation (2).

The essence of equation (4) lies in deriving a model, linear or ordinal, that depicts an answer as a linear combination of aggregated metric values. It is important to note that we maintain the original metric and opt for a straightforward model. This approach ensures that our optimization process selects the most effective temporal pulling of the metric, within the constraints imposed by weight flexibility. Essentially, this yields the optimal model for aggregating the metric. We operate under the assumption that the metric behaves consistently irrespective of its position within the sequence, allowing us to interpret weights as indicators of the significance of specific positions within the sequence.

In addition, we can estimate the confidence interval related to the weights by bootstrapping analysis. By selecting a bootstrap sample of user samples, we can estimate numerous weights functions. Next, we calculate the mean weights for any sample. Knowing the mean weights, we can calculate the distance from the mean to any other estimated weights. The final step is to remove all the weights with the largest distance, until we are left with not more than 95% of all estimated weights. The remaining weights generat 95% confidence interval.

Example of Analysis

We present three different cases of the analysis for three different time frames. It shows the flexibility of the method, which by nature is not time-dependent. The first study is long-term (LTS); we aggregated the values obtained each day to predict the value at the end of the week. The second study aggregates around 2.5 minutes of video. The last experiment is a typical laboratory experiment with sequences around 10 seconds.

LTS

At AGH, we conducted an experiment in which participants watched one video per day for six days of the week (from Monday to Saturday). On Sunday they did not watch any video, they were asked to summarize the whole week experience with a single ACR (Absolute Category Rating) 5-point discrete scale $\{1,2,3,4,5\}$ answer. The videos the testers watched were pre-downloaded and prepared to target specific and constant for the entire 40 seconds movie quality m_i.

In addition, we have run a laboratory experiment (lab) in which the testers watched the six videos from the whole week one by one and scored all movies with just a single ACR value. One of the research questions we have is: if the data collected in a lab study are different from those obtained in LTS.

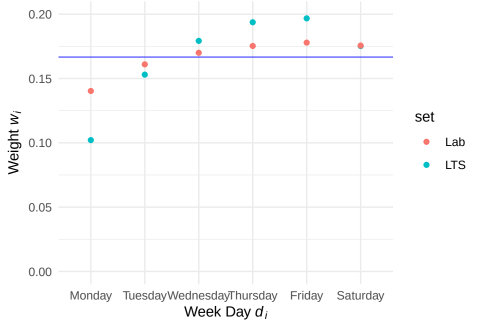

Figure 2. *Weights showing how strongly quality presented at this day influences the overall quality for both lab and long-term study. Blue line described equal influence, each day has the same importance.*

A detailed description of the LTS including more analysis is described in a separate publication, which is currently under review. Here, we present only results that utilize the described aggregation method.

To answer the above-mentioned research question, we compared the weights obtained for the lab and LTS study. If the weights are similar, we can say that the lab study described the behavior of LTS well enough. In Figure 2 we present the weights for each day of the week obtained in the lab[2] and LTS.

The results in Figure 2 show two interesting differences between lab and LTS. The first is natural; in the lab study people watched videos one by one; the first video has a stronger impact on the quality than the first video in the LTS. There is simply less time between watching the video and scoring. The second observation is not as obvious. For the lab study, the influence of the last four days is very similar. For LTS, Saturday has a significantly weaker influence than Friday. Analysis of such significant, but small, differences would not be possible with simple linear regression.

2.5 minutes

At AGH, we conducted an experiment in which participants watched a full episode of their choice on the Netflix service. Approximately every 2.5 minutes, they were asked to rate the quality of the video on the ACR scale, as specified by ITU-T Recommendation P.913 [9]. We developed a model to predict user ratings based on the aggregated VMAF (Video Multimethod Assessment Fusion) score, using the weighted aggregation function described in Section . The analysis helped us to understand which segments of the video significantly impacted perceived quality. Interestingly, we observed varying optimal aggregation functions for two distinct age groups, as depicted in Figure 3.

A detailed examination of the results falls beyond the scope of this paper; here, the experiment serves as a demonstrative application of the analytical tool, elucidating how quality perceptions are influenced in different demographics.

Classical Lab Experiment

As a classical subjective experiment, we used the LIVE Mobile Video Quality Database [10]. The videos are 15 seconds long.

[2] Of course, in lab study day of the week means the order a video was presented, not actual day of the week.

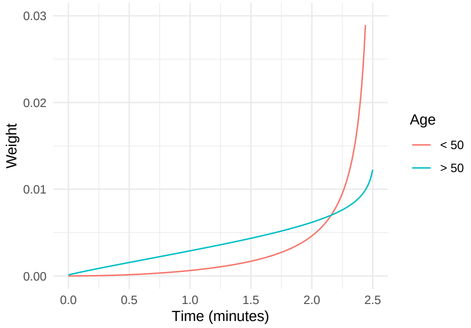

Figure 3. *Results obtained using the described method, showcasing that users older than 50 years consider a longer span of the watched sequence when rating quality.*

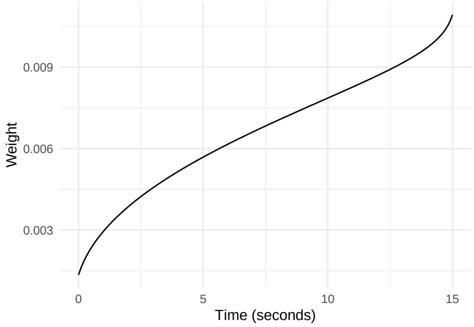

Figure 4. *Weights function obtained for 15 second long video. Even for a short video the function shows important recency effect.*

We obtained from Lukas Krasula from Netflix the VMAF scores calculated for each frame. Similarly to the previous analysis, we were able to see how important each frame (or second) is in a video. With a short video, our intuition suggests that time should not be an important factor, although previous research shows that it is not the case [4].

The optimal aggregation function is presented in Figure 4. The function obtained shows a nearly linear forgetting factor. It should be noted that with the presented model, a perfect linear function is possible with parameters $a = 1, b = 1$. This means that the curvatures at the end are more optimal than a straight line. We have found two possible explanations. The first is subject-based. Subjects could be distracted more at the start of the sequence and better recalling the last frame. The second possible explanation is based on compression. Perhaps the metric predicts the start of the sequence less precisely, since the compression algorithm, targeting a specific bitrate, is not yet stable. We think the second explanation is less probable, but we cannot exclude it.

The aggregation taking into account time influence yields better results. The AIC change from 140 for the mean to 102 in the case of Kumaraswamy aggregation. The change is statistically significant and easy to implement.

Summary

This paper addresses a fundamental challenge in modeling, the aggregation of partial values to yield a singular, comprehensive metric. This is a recurrent issue, especially in fields like video quality assessment where metrics from individual frames or seconds need to be amalgamated into a single score representing the overall quality or user experience. The paper sheds light on the existing aggregation operators, highlighting that often, the best fit algorithm is chosen without a thorough understanding of why it provides an optimal fit.

Our proposed solution to this challenge hinges on taking advantage of the Kumaraswamy distribution to discern the underlying effects and their intensities present in the data. We have outlined a weighted mean aggregation procedure utilizing the Kumaraswamy distribution, which is meticulously adapted to cater to the specific data range. The Kumaraswamy distribution parameters a and b are instrumental in determining the shape of the distribution, thus influencing the result of the aggregation.

A real-world application of this method is demonstrated through an analysis of three different experiments ranging from aggregating days, minutes, and frames. The analysis, supported by a predictive model that relates VMAF scores to user ratings, elucidated how different segments of the video were crucial in influencing perceived quality. A notable discovery was the variation in optimal aggregation functions across different age demographics, affirming the method's potential in uncovering user behavior patterns.

The innovative use of Kumaraswamy distribution for aggregation presents a robust tool for an in-depth understanding of how individual metric values coalesce to form an overall perception, especially in user experience-centric domains. Although the detailed results of the experiments are beyond the scope of this paper, the analysis serves as a testament to the utility and versatility of the proposed method in obtaining actionable insights from aggregated metric data. Future work is needed to better determine differences in the aggregation made by uses for different time intervals.

Acknowledgments

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014-2021 under project 2019/34/H/ST6/00599.

References

[1] Vicenç Torra and Yasuo Narukawa. *Modeling decisions: information fusion and aggregation operators*. Springer Science & Business Media, 2007.

[2] Anastasia Mozhaeva, Lee Streeter, Igor Vlasuyk, and Aleksei Potashnikov. Full reference video quality assessment metric on base human visual system consistent with psnr. In *2021 28th Conference of Open Innovations Association (FRUCT)*, pages 309–315, 2021.

[3] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2):2, 2016.

[4] Kalpana Seshadrinathan and Alan C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1153–1156, 2011.

[5] Natalia Cieplińska, Lucjan Janowski, Katrien De Moor, and Michał Wierzchoń. Long-term video qoe assessment studies: A systematic review. *IEEE Access*, 10:133883–133897, 2022.

[6] R.C. Atkinson and R.M. Shiffrin. Human memory: A pro-

posed system and its control processes. volume 2 of *Psychology of Learning and Motivation*, pages 89–195. Academic Press, 1968.

[7] MC Jones. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical methodology*, 6(1):70–81, 2009.

[8] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[9] ITU-T. Itu-t p.913 (06/2021): Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. *ITU: Geneva, Switzerland*, 2021.

[10] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo de Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):652–671, 2012.