Voting Aggregation Leads to (Interval) Median

Olga Kosheleva[1], and Vladik Kreinovich[2]

[1]Department of Teacher Education
[2]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Formulation of the Problem

- For many real-real problems, there are several different decision making tools.
- Each of these tools has its advantages: otherwise, it would not be used.
- To combine the advantages of different tools, it therefore desirable to aggregate their results.
- One of the most widely used methods of aggregating several results is voting:
  - if the majority of results satisfy a certain property,
  - then we conclude that the actual value has this property.
- *Example:* if most classifiers classify the disease as pneumonia, we conclude that it is pneumonia.

2. What We Do and Why This Is Non-Trivial

- *What we do:* we analyze how voting can be used to aggregate several numerical estimates.
- *Observation:* voting is closely related to the AI notion of a "typical" object of a class.
- *Example:* what is an intuitive meaning of a term "typical professor"?
  - if most professors are absent-minded,
  - then we expect a "typical" professor to be absent-minded as well.
- *Problem:* no one is perfectly typical, e.g., due to their specific research area.
- This is why in AI, there is an ongoing discussion on how best to describe typical (not abnormal) objects.

3. Main Definition

- *We have:* n estimates $x_i = (x_{i1}, \ldots, x_{iq})$ $(1 \le i \le n)$ for the values of q quantities.
- *We want:* to combine these estimates into a single estimate x, so that:
  - if the majority of x_i satisfies a property P,
  - then x should satisfy also this property.
- *Let $q \ge 1$, let $\mathcal{S}$ be a class of subsets of $\mathbb{R}^q$.*
- *We say that $x \in \mathbb{R}^q$ is a* possible $\mathcal{S}$-aggregate *of $x_1, \ldots, x_n \in \mathbb{R}^q$ if for every $S \in \mathcal{S}$:*
  - *if the majority of x_i are in this set,*
  - *then x should be in this set.*
- *The set of all possible $\mathcal{S}$-aggregates is called the $\mathcal{S}$-*aggregate *of the elements $x_1, \ldots, x_n$.*

4. What If We Allow All Properties?

- Let $U \stackrel{\text{def}}{=} 2^{\mathbb{R}^q}$ be the set of *all* subsets of $\mathbb{R}^q$.
- *For every $q \geq 1$ and $n \geq 3$, if all n elements $x_1, \ldots, x_n$ are all different, then their U-aggregate is empty.*
- *Proof:*
  - All x_i belong to $X = \{x_1, \ldots, x_n\}$, so $x \in X$ hence $x = x_i$ for some i.
  - But most elements x_j are different from x_i, i.e., $x_j \in -\{x_i\}$, thus, $x \neq x_i$.
  - Contradiction shows that no such x is possible, i.e., that the U-aggregare is empty.
- Thus, the aggregation problem is indeed non-trivial.

5. When Is the U-Aggregate Non-Empty?

- *Let $q \geq 1$.*
- *When $n = 1$ then the U-aggregate set of x_1 is $\{x_1\}$.*
- *When $n = 2$, then the U-aggregate set is $\{x_1, x_2\}$.*
- *For all odd $n \geq 3$:*
  - *if the majority of $x_1, \ldots, x_n$ are equal to each other, then the U-aggregate set is this common element.*
- *For all even $n \geq 4$:*
  - *if the majority of $x_1, \ldots, x_n$ are equal to each other, then the U-aggregate set is this common element;*
  - *if half of x_i are equal to a, and all others ae equal to $b \neq a$, then the U-aggregate is $\{a, b\}$.*
- *In all other cases, the U-aggregate set is empty.*

6. 1-D Interval-Based Voting Aggregation Leads to (Interval) Median

- Let I denote the set of all intervals $[a, b]$.
- For every sequence $x_1, \ldots, x_n$, let $x_{(1)} \leq \ldots \leq x_{(n)}$ denote the result of sorting the numbers x_i in increasing order.
- When $n = 2k + 1$ for some integer k, then by a *median*, we mean the value $x_{(k+1)}$.
- When n is even, i.e., when $n = 2k$ for some integer k, then by a *median*, we mean the interval $[x_{(k)}, x_{(k+1)}]$.
- The median will also be called an *interval median*.
- **Resyult:** *For every sequence of numbers* $x_1, \ldots, x_n$, *the* I*-aggregate set is equal to the median.*

7. Discussion

- Median is the most *robust* aggregation, i.e., the least vulnerable to possible outliers.
- Thus, median is often used in data processing.
- Median is used in econometrics, as a more proper measure of "average" ("typical") income than the mean.
- Otherwise, a single billionaire living in a small town:
  - increases its mean income
  - without affecting the living standards of its inhabitants.

8. What If We Require Strong Majority?

- What if we only require $x_i \in S$ when the proportion of values $x_i \in S$ exceeds a certain threshold $t > 0.5$?
- The set of all possible t-$\mathcal{S}$-aggregates is called the *t-$\mathcal{S}$-aggregate set.*
- *For every $x_1, \ldots, x_n$, and for every t, the t-I-aggregate set is the interval $[x_{(n-k+1)}, x_{(k)}]$, where $k = \lceil t \cdot n \rceil$.*

9. An Alternative Derivation of the Interval Median

- Interval median can be also derived from other natural conditions:
- That it is a continuous function of $x_1, \ldots, x_n$.
- That it is invariant with respect to arbitrary strictly increasing or strictly decreasing re-scalings:
  - such re-scalings which make physical sense;
  - e.g., we can measure sound energy in Watts or in decibels – which are logarithmic units.
- That this is the narrowest such operation:
  - else we could, e.g., take an operation returning the whole range $\left[\min\limits_i x_i, \max\limits_i x_i\right]$.

10. Formalizing Scale-Invariance

- We say that a mapping $A(x_1, \ldots, x_n) = [\underline{a}(x_1, \ldots, x_n), \overline{a}(x_1, \ldots, x_n)]$ is *scale-invariant* if:
  - for each strictly increasing $f(x)$:
  $$\underline{a}(f(x_1), \ldots, f(x_n)) = f(\underline{a}(x_1, \ldots, x_n)) \text{ and}$$
  $$\overline{a}(f(x_1), \ldots, f(x_n)) = f(\overline{a}(x_1, \ldots, x_n));$$
  - for each strictly decreasing continuous $f(x)$:
  $$\underline{a}(f(x_1), \ldots, f(x_n)) = f(\overline{a}(x_1, \ldots, x_n)) \text{ and}$$
  $$\overline{a}(f(x_1), \ldots, f(x_n)) = f(\underline{a}(x_1, \ldots, x_n));$$

11. Alternative Derivation: Result

- Let $n \geq 1$ be fixed.
- By an *aggregation operation*, we mean a mapping that maps each tuple $x_1, \ldots, x_n$ into an interval

  $A(x_1, \ldots, x_n) = [\underline{a}(x_1, \ldots, x_n), \overline{a}(x_1, \ldots, x_n)]$ so that:

  1. this operation is *continuous*, i.e., both functions $\underline{a}(x_1, \ldots, x_n)$ and $\overline{a}(x_1, \ldots, x_n)$ are continuous;
  2. this operation is *scale-invariant*;
  3. A is *the narrowest*: for every operations $B(x_1, \ldots, x_n)$ satisfying Properties 1 and 2,
     * if $B(x_1, \ldots, x_n) \subseteq A(x_1, \ldots, x_n)$ for all tuples,
     * then $B(x_1, \ldots, x_n) = A(x_1, \ldots, x_n)$ for all tuples.
- **Result:** *Interval median is the only aggregation operation in this sense.*

12. 1-D Case: Can We Expand Beyond Intervals?

- In our analysis, instead of closed intervals, we can consider general *convex subsets* of the real line.
- This includes closed, open, semi-open intervals, and intervals with infinite endpoints.
- Can we go beyond intervals?
- It turns out that we cannot:
- **Result:** *($n = 3$ or $n \ge 5$):*
  - *Let a class $\mathcal{S}$ contain, in addition to all the intervals, a non-convex set S_0.*
  - *Then, there exists values $x_1, \ldots, x_n$ for which the $\mathcal{S}$-aggregate set is empty.*

13. Multi-D Interval-Based Voting Aggregation

- Let B denote the set of all the boxes

$$[a_1, b_1] \times \ldots \times [a_q, b_q].$$

- **Result:** *For every sequence of tuples $x_1, \ldots, x_n$, the B-aggregate set is the box $M_1 \times \ldots \times M_q$, where*
  - *M_i is the interval median of the i-th components $x_{1i}, \ldots, x_{ni}$.*
- For the class P of all convex polytopes, P-aggregate can be empty: $x_1 = (0, 0, 0, \ldots, 0)$,

$$x_2 = (0.1, 0.9, 0, \ldots, 0), \quad x_3 = (1, 1, 0, \ldots, 0).$$

- Here, B-aggregate is the median $(0.1, 0.9, 0, \ldots, 0)$.
- However, the majority of x_i (namely, x_1 and x_3) belong to the segment $S = \{(x, x, 0, \ldots, 0) : 0 \le x \le 1\}$.
- Alas, the median does not belong to this segment.

14. Acknowledgments

This work was supported in part:

- by the National Science Foundation grants:
  - HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
  - DUE-0926721, and
- by an award from Prudential Foundation.

15. Proof of the Main 1-D Result

- Let us first prove that every possible U-aggregate x should belong to the median set.
- If $n = 2k + 1$, then the majority of $x_{(i)}$ belong to $[x_{(1)}, x_{(k+1)}]$: namely, $x_{(1)} \leq \ldots \leq x_{(k+1)}$.
- Thus, every possible U-aggregate x must belong to the same interval, and thus, we must have $x \leq x_{(k+1)}$.
- Similarly, the majority of elements $x_{(i)}$ belong to $[x_{(k+1)}, x_{(n)}]$: namely, $x_{(k+1)} \leq \ldots \leq x_{(n)}$.
- Thus, every possible U aggregate x must belong to the same interval, and so, we must have $x \geq x_{(k+1)}$.
- From $x \leq x_{(k+1)}$ and $x \geq x_{(k+1)}$, we conclude that $x = x_{(k+1)}$, i.e., x coincides with the median.
- If $n = 2k$, then the majority of $x_{(i)}$ belong to $[x_{(1)}, x_{(k+1)}]$: namely, $x_{(1)} \leq \ldots \leq x_{(k+1)}$.

16. Proof of the Main 1-D Result (cont-d)

- Thus, every possible U aggregate x must belong to the same interval, and so, we must have $x \leq x_{(k+1)}$.
- Similarly, the majority of $x_{(i)}$ belong to the interval $[x_{(k)}, x_{(n)}]$: namely, $x_{(k)} \leq \ldots \leq x_{(n)}$.
- Thus, every possible U aggregate x must belong to the same interval, and thus, we must have $x \geq x_{(k)}$.
- So, we conclude that $x_{(k)} \leq x \leq x_{(k+1)}$, i.e., that x is indeed an element of the median interval $[x_{(k)}, x_{(k+1)}]$.
- To complete the proof, let us prove that every element of the interval median is indeed a possible I-aggregate.
- For this, we need to show that:
  - if an interval $[a, b]$ contains the majority of $x_{(i)}$,
  - then $[a, b]$ contains the interval median.

17. Proof of the Main 1-D Result (cont-d)

- We need to prove that:
  - if an interval $[a, b]$ contains the majority of $x_{(i)}$,
  - then $[a, b]$ contains the interval median.
- Let us prove it by considering two possible situations: when n is odd and when n is even.
- Let us show that in the odd case $n = 2k + 1$, if the interval $[a, b]$ contains the majority of the elements $x_{(i)}$, then $x_{(k+1)} \in [a, b]$, i.e., $a \le x_{(k+1)}$ and $x_{(k+1)} \le b$.
- We will prove both inequalities by contradiction.
- If $a > x_{(k+1)}$, then the interval $[a, b]$ cannot contain any of the $k + 1$ elements $x_{(1)} \le \ldots \le x_{(k+1)}$.
- Thus, $[a, b]$ contains $\le$ k remaining elements $x_{(k+2)}, \ldots, x_{(n)}$ – which do not form a majority.

18. Proof of the Main 1-D Result (cont-d)

- Similarly, if $b < x_{(k+1)}$, then the interval $[a, b]$ cannot contain any of the $k + 1$ elements $x_{(k+1)} \le \ldots \le x_{(n)}$.
- Thus, this interval contains $\le k$ remaining elements $x_{(1)}, \ldots, x_{(k)}$, which also do not form a majority.
- So, if the interval $[a, b]$ contains the majority of elements $x_{(i)}$, then it must contain the median $x_{(k+1)}$.
- So, the median is a possible I-aggregate of the values $x_1, \ldots, x_n$.
- Let us show that in the even case, when $n = 2k$:
  - if the interval $[a, b]$ contains the majority of the elements $x_{(i)}$,
  - then $[x_{(k)}, x_{(k+1)}] \subseteq [a, b]$, so $a \le x_{(k)}$ & $x_{(k+1)} \le b$.
- We will prove both inequalities by contradiction.

19. Proof of the Main 1-D Result (final part)

- If $a > x_{(k)}$, then the interval $[a, b]$ cannot contain any of the k elements $x_{(1)} \leq \ldots \leq x_{(k)}$.
- So, it must contain no more than k remaining elements $x_{(k+1)}, \ldots, x_{(n)}$ – which do not form a majority.
- Similarly, if $b < x_{(k+1)}$, then the interval $[a, b]$ cannot contain any of the k elements $x_{(k+1)} \leq \ldots \leq x_{(n)}$.
- So, it must contain no more than k remaining elements $x_{(1)}, \ldots, x_{(k)}$, which also do not form a majority; thus:
  - if the interval $[a, b]$ contains the majority of elements $x_{(i)}$,
  - then it must contain the median $[x_{(k)}, x_{(k+1)}]$.
- So, every element from the interval median is a possible I-aggregate of the values $x_1, \ldots, x_n$. Q.E.D.

20. Proof of the Multi-D Result

- Let us first prove that every possible B-aggregate tuple belongs to the median box $M_1 \times \ldots \times M_q$.
- Let us fix one of the dimensions i and consider $x_{1i}, \ldots, x_{ni}$.
- For all $j \neq i$, consider the smallest intervals $[A_j, B_j] \stackrel{\text{def}}{=} \left[\min\limits_k (x_{kj}), \max\limits_k (x_{kj})\right]$ containing all x_{kj}.
- For each possible B-aggregate tuple $x = (e_1, \ldots, e_q)$, the desired property holds for all the boxes of the type

$$B = [A_1, B_1] \times \ldots \times [A_{i-1}, B_{i-1}] \times [a_i, b_i] \times [A_{i+1}, B_{i+1}] \times \ldots \times$$

- Since all other intervals forming this box are the largest possible, $x_j \in B \Leftrightarrow x_{j1} \in [a_i, b_i]$.

21. Proof of the Multi-D Result (cont-d)

- Thus, for these boxes:
  - the definition of a possible B-aggregate of the tuples $x_1, \ldots, x_n$ implies that
  - the i-th component e_i of the tuple x is a possible I-aggregate of the components $x_{1i}, \ldots, x_{ni}$.
- We already know that this implies that e_i belongs to the interval median M_1 of these components.
- Vice versa, let us prove that every tuple $x \in M_1 \times \ldots \times M_q$ is a possible B-aggregate.
- Indeed, let $x = (e_1, \ldots, e_q) \in M_1 \times \ldots \times M_q$, and let us assume that the majority of x_i belong to the box

$$B = [a_1, b_1] \times \ldots \times [a_q, b_q].$$

- This implies, for every component i, that the majority of the values $x_{1i}, \ldots, x_{ni}$ belong to the interval $[a_i, b_i]$.

22. Proof of the Multi-D Result (final part)

- For every component i, the majority of the values $x_{1i}, \ldots, x_{ni}$ belong to the interval $[a_i, b_i]$.
- We already know, from the 1-D case, that this implies that $e_i \in [a_i, b_i]$ for every i.
- Thus, we indeed have

$$x = (e_1, \ldots, e_q) \in [a_1, b_1] \times \ldots \times [a_q, b_q] = B.$$

- The proposition is proven.