Interval Techniques for Processing Educational Data

Olga Kosheleva[1], Vladik Kreinovich[2], Luc Longpré[2], Mourat Tschoshanov[1], and Gang Xiang[2]

[1]Department of Teacher Education
[2]Depatment of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA

olgak@utep.edu, vladik@utep.edu, longpre@utep.edu
mouratt@utep.edu, gxiang@utep.edu

1. Formulation of the Problem

- Teaching is very important, and teaching is not always very effective.
- There exist many different pedagogical techniques.
- It is important to experimentally compare their effectiveness.
- *Traditional approach:* statistical techniques.
- *Problem:*
  - traditional techniques are tailored to processing numbers;
  - in education, we often have intervals (A means [90,100]) or fuzzy-type perceptions like "understands well".
- *Conclusion:* we need new techniques to process such interval and fuzzy statistical data.

2. Problems with the Traditional Approach: Example

- How do we select a teaching method?
- *Main criterion:* good average results m: e.g., good average grade.
- *Also important:* ensure that the results are consistently good, i.e., standard deviation σ is low.
- *Example:*
  - in one method, all the students got Bs,
  - in the other method, half of the students got Bs and half of the students got As.
- *Traditional approach:* $A = 4$, $B = 3$, so:
  - In the first method, $m = 3$ and $\sigma = 0$.
  - In the second method, $m = 3.5$ and $\sigma = 0.5$.
- *Conclusion:* second method is less stable.

3. Problems with the Traditional Approach: Example cont-d

- In reality:
  - in the first method, half of the students got 80, half 88; and
  - in the second method, half of the students got 89, and half 91.
- Then:
  - In the first method, the average is $m = \dfrac{80+88}{2} = 84$, and $\sigma = 4$.
  - In the second method, the average is $m = \dfrac{89+91}{2} = 90$, and $\sigma = 1 \ll 4$.
- *Conclusion:* second method is actually *more* stable.
- *What is necesssary:* we want methods that would:
  - take interval uncertainty into account and thus
  - provide guaranteed answers to questions like:

    "Is the first method better than the second one?".

4. Interval Approach Is Needed

- *Main problem:* letter grade ℓ represents an *interval* $\mathbf{x} = [\underline{x}, \overline{x}]$ of possible values of the number grade.
- *Examples:*
  - the letter grade A represents the interval $[90, 100]$;
  - the letter grade B represents the interval $[80, 90]$;
  - the letter grade C represents the interval $[70, 80]$.
- *Main objective:* given a set of letter grades $\ell_1, \ldots, \ell_n$, to compute a certain statistical characteristic C such as:
  - average, standard deviation,
  - correlation with other characteristics (such as the family income or the amount of time that a student spends on homeworks).
- *Traditional approach:* the statistical characteristic is defined in terms of numerical values, as $C = C(x_1, \ldots, x_n)$.
- *Examples:* the population average $m = \dfrac{x_1 + \ldots + x_n}{n}$, the population variance $V = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - m)^2$.

5. Interval Approach Is Needed (cont-d)

- *Specific feature of educational data:*
  - instead of the *exact* values x_i,
  - we often only know the *intervals* $\mathbf{x}_i$ corresponding to the letter grade ℓ_i.
- *Problem:*
  - for different possible values $x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n$,
  - we get different values of the corresponding characteristic C.
- *Objective:* given:
  - the characteristic $C(x_1, \ldots, x_n)$ and
  - the intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of possible values of $x_1, \ldots, x_n$,

  find the range of possible values $C(x_1, \ldots, x_n)$ of the desired characteristic when $x_i \in \mathbf{x}_i$.
- *Objective reformulated in mathematical terms:* find the interval

$$\mathbf{C} = \{C(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

6. This Problem Is a Particular Case of the General Problem of Interval Computations

- The need to perform computations under interval uncertainty occurs in many areas of science and engineering.
- In many such areas, we therefore face the following problem:
  - we know:
    - n intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and
    - an algorithm $y = f(x_1, \ldots, x_n)$ that transforms n real numbers (inputs) into a single number y (result of data processing);
  - we must estimate the range of possible values of y, i.e., the interval

$$\mathbf{y} = \{f(x_1, \ldots, x_n) \mid x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- This problem is called the main problem of *interval computations*.
- *Conclusion:* the problem of processing educational data under interval uncertainty is a particular case of the more general problem of interval computations.

7. How We Can Process Interval Data: General Description

- *General case:* many efficient techniques are known.
- *Statistical case:* several algorithms have been developed for the case when the the function $f(x_1, \ldots, x_n)$ is one of the standard statistical characteristics such as average m or standard deviation V;.
- *Specific feature of average:* average m is a monotonic function of each of its variables.
- *Conclusion:*
  - its value is the largest when each x_i attains the largest possible value $x_i = \overline{x}_i$, and
  - its value is the smallest when the variance attains its smallest possible value $\underline{x}_i$.
- Thus, for the average m, the interval takes the form $[\underline{m}, \overline{m}]$, where

$$\underline{m} = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}; \quad \overline{m} = \frac{\overline{x}_1 + \ldots + \overline{x}_n}{n}.$$

8. Computing Variance Under Interval Uncertainty

- *Bad news:* in general, for interval uncertainty, computing the range $[\underline{V}, \overline{V}]$ of variance V is NP-hard.
- *Specific case:* for educational data, intervals can only "touch".
- *Good news:* in such case, there is an efficient algorithm.
- *Algorithm for computing $\overline{V}$:*
  - First, we sort the grades into an increasing sequence for which $\underline{x}_1 \le \underline{x}_2 \le \ldots \le \underline{x}_n$ and $\overline{x}_1 \le \overline{x}_2 \le \ldots \le \overline{x}_n$.
  - For every k from 1 to n,
    * we pick $x_i = \underline{x}_i$ for $i \le k$ and $x_i = \overline{x}_i$ for $i > k$;
    * then, we compute the average $m = \dfrac{x_1 + \ldots + x_n}{n}$ of the selected x_i, and
    * check whether this average satisfies the inequality

$$\underline{x}_k \le m \le \overline{x}_{k+1}.$$

    * If it does, then the population variance of the corresponding sequence $x_1, \ldots, x_n$ is exactly the desired upper bound $\overline{V}$.

9. Computing Variance Under Interval Uncertainty (cont-d)

- To compute the lower bound $\underline{V}$, similarly, for every k,
  - we select:
    - $x_i = \overline{x}_i$ when $\overline{x}_i \le \underline{x}_k$, and
    - $x_i = \underline{x}_i$ when $\underline{x}_i \ge \overline{x}_k$.
  - We then compute the average m of the selected x_i and check whether this average satisfies the inequality

$$\underline{x}_k \le m \le \overline{x}_k.$$

  - If it does, then:
    - we assign $x_i = m$ for all the un-assigned value i, and
    - the population variance of the corresponding sequence $x_1, \ldots, x_n$ is exactly the desired lower bound $\underline{V}$.

10. Numerical Example: Computing $\overline{V}$

- *Situation:* sorted grades are C, B, and A, so $\underline{x}_1 = 70$, $\overline{x}_1 = 80$, $\underline{x}_2 = 80$, $\overline{x}_2 = 90$, $\underline{x}_3 = 90$, $\overline{x}_3 = 100$.
- For $k = 1$, we pick:
  - $x_1 = \underline{x}_1 = 70$,
  - $x_2 = \overline{x}_2 = 90$, and
  - $x_3 = \overline{x}_3 = 100$.
- Here, $m = \dfrac{x_1 + x_2 + x_3}{3} = 86\dfrac{2}{3}$.
- Here, $\underline{x}_1 = 70 \le m \le \overline{x}_2 = 90$.
- So, the upper bound $\overline{V}$ is equal to the population variance $\dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - m)^2$ of the values $x_1 = 70$, $x_2 = 90$, and $x_3 = 100$.
- Conclusion: $\overline{V} = 102\dfrac{2}{9}$.

11. Numerical Example: Computing $\underline{V}$

- *Situation:* sorted grades are C, B, and A, so $\underline{x}_1 = 70$, $\overline{x}_1 = 80$, $\underline{x}_2 = 80$, $\overline{x}_2 = 90$, $\underline{x}_3 = 90$, $\overline{x}_3 = 100$.
- For $k = 1$, we assign $x_2 = \underline{x}_2 = 80$ and $x_3 = \underline{x}_3 = 90$.
- Their average $m = 85$ is outside the interval $[\underline{x}_1, \overline{x}_1] = [70, 80]$, so we have to consider the next k.
- For $k = 2$, we assign $x_1 = \overline{x}_1 = 80$ and $x_3 = \underline{x}_3 = 90$.
- Their average $m = 85$ of these two values satisfies the inequality $\underline{x}_2 = 80 \le m \le \overline{x}_2 = 90$.
- Hence we assign $x_2 = 85$, and compute $\underline{V}$ as the population variance of the values $x_1 = 80$, $x_2 = 85$, and $x_3 = 90$.
- Conclusion: $\underline{V} = 16\dfrac{2}{3}$.

12. Fuzzy Approach: In Brief

- *The main idea behind fuzzy uncertainty:*
  - instead of just describing which objects (in our case, grades) are possible,
  - we also describe, for each object x, the degree $\mu(x)$ to which this object is possible.
- For each degree of possibility α, we can determine the set of objects that are possible with at least this degree of possibility.
- *Comment:* this set is called the α-*cut* $\{x \mid \mu(x) \geq \alpha\}$ of the original fuzzy set.
- *How we can process fuzzy data: general idea.*
  - Instead of an interval $\mathbf{x}_i$, we have a family of nested intervals $\mathbf{x}_i(\alpha)$ – α-cuts of the given fuzzy sets.
  - *Objective:* compute the fuzzy number corresponding to this the desired characteristic $C(x_1, \ldots, x_n)$.
  - *How:* for each α, apply the interval algorithm to the α-cuts $\mathbf{x}_i(\alpha)$ of the corresponding fuzzy sets.
  - The resulting nested intervals form the desired fuzzy set for C.

13. Need for Combining Probabilistic, Interval, and Fuzzy Uncertainty

- In the case of *interval* uncertainty,
  - we consider all possible values of the grades, and
  - we do not make any assumptions about the probability p of different values within the corresponding intervals.
- *In reality:* in many cases, we have commonsense ideas about p.
- *Example 1:*
  - *fact:* a student has almost all As but only one B;
  - *conclusion:* this is a strong student;
  - *prediction:* this B is at the high end of the B interval.
- *Example 2:*
  - *fact:* a student has almost all Cs but only one B;
  - *conclusion:* this is a weak student;
  - *prediction:* this B is at the lower end of the B interval.
- It is desirable to take such arguments into account when processing educational data.

14. Simplest Case: Normally Distributed Grades

- *Situation:* actual number grades are normally distributed, with an (unknown) mean m and an unknown standard deviation σ.
- *In precise terms:* the cumulative probability distribution (cdf) $F(x) \stackrel{\text{def}}{=} \text{Prob}(\xi < x)$ has the form $F_0\left(\dfrac{x-m}{\sigma}\right)$, where $F_0(x)$ is the cdf of the standard Gaussian distribution with 0 mean and unit standard deviation.
- *Our objective:* determine the values m and σ.
- *Case of number grades:* if we knew the values of the number grades x_i, then we could apply the above statistics and estimate m and $\sigma = \sqrt{V}$.
- *Case of letter grades:* based on the letter grades, we can find, for the threshold values 60, 70, etc., the frequency with which we have the grade smaller that this threshold.
- *Notations:* denote by f the proportion of F grades, by d the proportion of D grades.
- *Conclusion:* the frequency of $x < 60$ is f, the frequency of $x < 70$ is $f + d$, the frequency of $x < 80$ is $f + d + c$, etc.

15. Simplest Case: Towards an Algorithm

- *Idea:* probability ≈ frequency:

$$F_0\left(\frac{60-m}{\sigma}\right)\approx f;\quad F_0\left(\frac{70-m}{\sigma}\right)\approx f+d;$$

$$F_0\left(\frac{80-m}{\sigma}\right)\approx f+d+c;\quad F_0\left(\frac{90-m}{\sigma}\right)\approx f+d+c+b.$$

- *Simplification:* let us denote by $\psi_0(t)$ the function that is inverse to $F_0(t)$.
- *Result:* the first equality takes the form $60-m/\sigma\approx\psi_0(f)$, i.e., $\sigma\cdot\psi_0(f)+m\approx 60$.
- *Result:* to find the unknowns m and σ, we get a system of linear equations:

$$\sigma\cdot\psi_0(f)+m\approx 60;\quad \sigma\cdot\psi_0(f+d)+m\approx 70;$$

$$\sigma\cdot\psi_0(f+d+c)+m\approx 80;\quad \sigma\cdot\psi_0(f+d+c+b)+m\approx 90.$$

- *How we can solve it:* e.g., Least Squares.
- *Comment:* if the distribution is non-Gaussian, and we know its shape, i.e., $F(x)=F_0((x-m)/\sigma)$ for known $F_0(t)$, then we can similarly find m and σ.

16. Simplified Case When All the Grades Are $\geq$ C

- *Real-life situation:* in many cases, only C and above is an acceptable grade.
- In such situations, $f = d = 0$ and $c + b + a = 1$, so we get a simplified system of two linear equations with two unknowns:

$$\sigma \cdot \psi_0(c) + m = 80; \quad \sigma \cdot \psi_0(c + b) + m = 90.$$

- *Simplification:* subtracting the first equation from the second one, we conclude that $\sigma = \dfrac{10}{\psi_0(b + c) - \psi_0(c)}$.
- *Further simplification;* this formula can be further simplified if the distribution $F_0(x)$ is symmetric (e.g., Gaussian distribution is symmetric), i.e., for every x,
  - the probability $F_0(-x)$ that $\xi \leq -x$ is equal to
  - the probability $1 - F_0(x)$ that $\xi \geq x$.
- *Simplified result:*

$$\sigma = -\frac{10}{\psi_0(a) + \psi_0(c)}; \quad m = 80 + \frac{10}{1 + \dfrac{\psi_0(a)}{\psi_0(c)}}.$$

17. Relation to Fuzzy Logic

- *Reminder:*

$$\sigma = -\frac{10}{\psi_0(a) + \psi_0(c)}; \quad m = 80 + \frac{10}{1 + \dfrac{\psi_0(a)}{\psi_0(c)}}.$$

- *Conclusion 1:* σ is an increasing function of the sum $\psi_0(a) + \psi_0(c)$.
- *Conclusion 2:* m is monotonically increasing with the ratio $\psi_0(a)/|\psi_0(c)|$.
- *Explanation:* $\psi_0(a)$ monotonically depends on the proportion a of grades in the A range:
  - the more grades are in the A range and
  - the fewer grades are in the C range,
  - the larger the average grade m.
- So m should be kind of monotonically depending on the degree to which is is true that we have A grades and not C grades.
- *Comment:* the operations of:
  - sum as "or" and
  - ratio as "a and not c"

  naturally appear in fuzzy logic.

18. Selecting an "Or" Operation

- The degree of belief a in a statement A can be estimated as proportional to the number of arguments in favor of A.
- In principle, there exist infinitely many potential arguments.
- So, in general, it is hardly probable that
  - when we pick a arguments out of infinitely many and then
  - we pick b out of infinitely many,

  the corresponding sets will have a common element.
- Thus, it is reasonable to assume that every argument in favor of A is different from every argument in favor of B.
- Under this assumption, the total number of arguments in favor of A and arguments in favor of B is equal to $a + b$.
- *Conclusion:* the natural degree of belief in $A \vee B$ is proportional to $a + b$.

19. Selecting an "And" Operation

- Different experts are reliable to different degrees.
- Our degree of belief in a statement A made by an expert is $w \,\&\, a$, where:
  - w is our degree of belief in this expert, and
  - a is the expert's degree of belief in the statement A.
- What are the natural properties of the "and"-operation?
- First, since $A \,\&\, B$ means the same as $B \,\&\, A$, we have $a \,\&\, b = b \,\&\, a$.
- Second, when an expert makes two statements B and C, then our resulting degree of belief in $B \vee C$ can be computed in two different ways:
  - We can first compute *his* degree of belief $b \vee c$ in $B \vee C$, and then use the "and"-operation to generate our degree of belief $w \,\&\, (b \vee c)$.
  - We can also first generate our degrees $w \,\&\, b$ and $w \,\&\, c$, and then use an "or"-operation to combine these degrees, arriving at $(w \,\&\, b) \vee (w \,\&\, c)$.

20. Selecting an "And" Operation (cont-d)

- *Reminder:*
  - We can first compute *his* degree of belief $b \vee c$ in $B \vee C$, and then use the "and"-operation to generate our degree of belief $w \,\&\, (b \vee c)$.
  - We can also first generate our degrees $w \,\&\, b$ and $w \,\&\, c$, and then use an "or"-operation to combine these degrees, arriving at $(w \,\&\, b) \vee (w \,\&\, c)$.
- It is natural to require that both ways lead to the same degree of belief, i.e., that the "and"-operation be distributive with respect to $\vee$.
- *Reasonable:* $w \,\&\, a$ is (non-strictly) increasing function.
- It can be shown that every commutative, distributive, and monotonic operation $\& : R \times R \to R$ has the form $a \,\&\, b = C \cdot a \cdot b$ for some $C > 0$.
- This expression can be further simplified if we introduce a new scale of degrees of belief $a' \stackrel{\text{def}}{=} C \cdot a$; in the new scale, $a \,\&\, b = a \cdot b$.

21. Selecting Implication and Negation

- *Selecting implication – idea:* an implication $A \to B$ is a statement C such that if we add C to A, we get B.
- *Reformulation:* in other words, $(a \to b) \,\&\, a = b$.
- *Resulting selection of implication:* since $a \,\&\, b = a \cdot b$, we get $a \to b = b/a$.
- *Negation – idea:* $\neg A$ can be viewed as a particular case of implication, $A \to F$, for a false value F.
- *Question:* how do we select the value e_0 corresponding to "false" F?
- *Idea:* we know that "false" and "false" is "false".
- *Reformulation:* in other words, $e_0 \,\&\, e_0 = e_0$.
- *Resulting selection of "false":* since $a \,\&\, b = a \cdot b$, this implies that $e_0 = 1$.
- *Resulting selection of negation:* negation $A \to F$ is equal to $a \to e_0$, i.e., to $1/a$.
- *Conclusion:* "A and not B" means $a \cdot (1/b) = a/b$; this is exactly the operation we wanted.