Detecting Outliers under Interval Uncertainty: A New Algorithm Based on Constraint Satisfaction

Evgeny Dantsin and Alexander Wolpert
Department of Computer Science, Roosevelt University
Chicago, IL 60605, USA, {edantsin,awolpert}@roosevelt.edu

Martine Ceberio, Gang Xiang, and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA, {mceberio,vladik}@cs.utep.edu

1. Outlier Detection Is Important

- In many application areas, it is important to detect *outliers*, i.e., unusual, abnormal values.
- *In medicine:* outliers may mean disease.
- *In geophysics:* outlier may mean a mineral deposit.
- *In structural integrity testing:* outlier may mean a structural fault.
- *Traditional engineering approach* to outlier detection:
  - collect measurement results $x_1, \ldots, x_n$ corresponding to normal situations;
  - compute $E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ and $\sigma = \sqrt{V}$, where $V \stackrel{\text{def}}{=} M - E^2$ and $M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$;
  - a value x is classified as an outlier if it is outside the interval $[L, U]$, where $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$, $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$, and $k_0 > 1$ is pre-selected (most frequently, $k_0 = 2$, 3, or 6).

2. Outlier Detection Under Interval Uncertainty

- *In practice:* often, we only have intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of x_i.
- *Example:* the value $\widetilde{x}_i$ measured by an instrument with a known upper bound Δ_i on the measurement error means that

$$x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i].$$

- *Problem:* for different values $x_i \in \mathbf{x}_i$, we get different L and U.
- *Objective:* given $\mathbf{x}_i$ and k_0, compute

$$\mathbf{L} = [\underline{L}, \overline{L}] \stackrel{\text{def}}{=} \{L(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\};$$

$$\mathbf{U} = [\underline{U}, \overline{U}] \stackrel{\text{def}}{=} \{U(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- A value x is a *possible* outlier if it is outside one of the possible k_0-sigma intervals $[L, U]$, i.e., if $x \notin [\overline{L}, \underline{U}]$.
- A value x is a *guaranteed* outlier if it is outside all possible k_0-sigma intervals $[L, U]$, i.e., if , i.e., if $x \notin [\underline{L}, \overline{U}]$.

3. Which Approach Is More Reasonable?

- *Situation:* our main objective is not to miss an outlier.
  - *Example:* structural integrity tests.
  - *Clarification:* we do not want to risk launching a spaceship with a faulty part.
  - *Reasonable approach:* look for possible outliers.
- *Situation:* make sure that the value x is an outlier.
  - *Example:* planning a surgery.
  - *Clarification:* we want to make sure that there is a micro-calcification before we start cutting the patient.
  - *Reasonable approach:* look for guaranteed outliers.

4. Detecting Outliers Under Interval Uncertainty: What Is Known

- *Case of possible outliers:* there exist efficient algorithms for computing $\overline{L}$ and $\underline{U}$.
- *Case of guaranteed outliers:* the computation of $\underline{L}$ and $\overline{U}$ is, in general, NP-hard.
- *Technical result:* if $1 + (1/k_0)^2 < n$ (e.g., if $k_0 > 1$ and $n \geq 2$), then the maximum $\overline{U}$ of U (and the minimum $\underline{L}$ of L) is always attained at a combination of endpoints of $\mathbf{x}_i$.
- *Resulting algorithm:* compute $\overline{U}$ and $\underline{L}$ by trying all 2^n combinations of $\underline{x}_i$ and $\overline{x}_i$.
- *Specific case:* when all measured values $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ are definitely different from each other, in the sense that the "narrowed" intervals do not intersect

$$\left[\widetilde{x}_i - \frac{1+\alpha^2}{n} \cdot \Delta_i, \widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i\right],$$

where $\alpha = 1/k_0$ and $\Delta_i \stackrel{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width.

- *Good news:* in this case, we can compute $\overline{U}$ and $\underline{L}$ in feasible time.

5. What We Plan To Do

- *More general case:* no two narrowed intervals are proper subsets of one another.
- *In precise terms:* one of them is not a subset of the interior of the other.
- *Objective:* extend known efficient algorithms to this case.
- Since $\underline{L}(\mathbf{x}_i) = -\overline{U}(-\mathbf{x}_i)$, it suffices to be able to compute $\overline{U}$.
- *Main idea:* reduce the interval computation problem to the constraint satisfaction problem with the following constraints:
  - for every i, if in the maximizing assignment we have $x_i = \underline{x}_i$, then replacing this value with $x_i = \overline{x}_i$ will either decrease U or leave U unchanged;
  - for every i, if in the maximizing assignment we have $x_i = \overline{x}_i$, then replacing this value with $x_i = \underline{x}_i$ will either decrease U or leave U unchanged;
  - for every i and j, replacing both x_i and x_j with the opposite ends of the corresponding intervals $\mathbf{x}_i$ and $\mathbf{x}_j$ will either decrease U or leave U unchanged.

6. Algorithm

- *General idea:*
  - First, we sort of the values $\widetilde{x}_i$ into an increasing sequence.
  - Without losing generality, we can assume that
$$\widetilde{x}_1 \le \widetilde{x}_2 \le \ldots \le \widetilde{x}_n.$$
  - Then, for every k from 0 to n, we compute the value $V^{(k)} = M^{(k)} - (E^{(k)})^2$ of the population variance V for the vector $x^{(k)} = (\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$, and we compute $U^{(k)} = E^{(k)} + k_0 \cdot \sqrt{V^{(k)}}$.
  - Finally, we compute $\overline{U}$ as the largest of $n+1$ values $U^{(0)}, \ldots, U^{(n)}$.
- *Details:* how to compute the values $V^{(k)}$
  - First, we explicitly compute $M^{(0)}$, $E^{(0)}$, and
$$V^{(0)} = M^{(0)} - (E^{(0)})^2.$$
  - Once we know the values $M^{(k)}$ and $E^{(k)}$, we can compute
$$M^{(k+1)} = M^{(k)} + \frac{1}{n} \cdot (\underline{x}_{k+1})^2 - \frac{1}{n} \cdot (\overline{x}_{k+1})^2$$
and $E^{(k+1)} = E^{(k)} + \frac{1}{n} \cdot \underline{x}_{k+1} - \frac{1}{n} \cdot \overline{x}_{k+1}$.

7. Number of Computation Steps

- *Sorting:* requires $O(n \cdot \log(n))$ steps.
- Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V^{(0)}$ requires linear time $O(n)$.
- For each k from 0 to $n-1$, we need a constant number of steps to compute the next values $M^{(k+1)}$, $E^{(k+1)}$, and $V^{(k+1)}$ as

$$M^{(k+1)} = M^{(k)} + \frac{1}{n} \cdot (\underline{x}_{k+1})^2 - \frac{1}{n} \cdot (\overline{x}_{k+1})^2$$

  and $E^{(k+1)} = E^{(k)} + \frac{1}{n} \cdot \underline{x}_{k+1} - \frac{1}{n} \cdot \overline{x}_{k+1}$.
- *Computing* $U^{(k)} = E^{(k)} + k_0 \cdot \sqrt{V^{(k)}}$ also requires a constant number of steps.
- *Finally, finding* the largest of $n+1$ values $U^{(k)}$ requires $O(n)$ steps.
- *Overall:* we need

$$O(n \cdot \log(n)) + O(n) + O(n) + O(n) = O(n \cdot \log(n)) \text{ steps.}$$

- *Comment:* if the measurement results $\widetilde{x}_i$ are already sorted, then we only need linear time to compute $\overline{U}$.

8. Justification of the Algorithm

- *Known:* $\overline{U} = \max U$ is attained at a vector $x = (x_1, \ldots, x_n)$ in which each value x_i is equal either to $\underline{x}_i$ or to $\overline{x}_i$.
- *New result:* this maximum is attained at one of the vectors $x^{(k)}$ in which all the lower bounds $\underline{x}_i$ precede all the upper bounds $\overline{x}_i$.
- *How we prove it:* by reduction to a contradiction.
- *Assume:* the maximum is attained at a vector x in which one of the lower bounds follows one of the upper bounds.
- *Notation:* let i be the largest upper bound index followed by the lower bound.
- *Conclusion:* in x_{opt}, we have $x_i = \overline{x}_i$ and $x_{i+1} = \underline{x}_{i+1}$.
- *Following proof:* since maximum is attained at x, each replacing:
  - replacing x_i with $\underline{x}_i$;
  - replacing x_{i+1} with $\overline{x}_{i+1}$; and
  - replacing both

  leads to $\Delta U \le 0$; we trace these changes ΔU.
- We then conclude that one of the narrowed intervals is a proper subset of another – contradiction to our assumption.

9. Acknowledgments

This work was supported in part by:

- NASA under cooperative agreement NCC5-209,
- NSF grant EAR-0225670,
- NIH grant 3T34GM008048-20S1, and
- Army Research Lab grant DATM-05-02-C-0046.

The authors are thankful to the anonymous referees for valuable suggestions.