Maximum Entropy in Support of Semantically Annotated Datasets

Paulo Pinheiro da Silva,
Vladik Kreinovich, and Christian Servin

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
emails paulo@utep.edu,
vladik@utep.edu, christians@miners.utep.edu

1. Checking Whether Two Datasets Represent the Same Data: Formulation of the Problem

- *In the semantic web:* data are often encoded in Resource Description Framework (RDF).
- *In RDF:* every piece of information is represented as a triple consisting of a *subject*, a *predicate*, and an *object*.
- *Example:* a predicate *hasGravityReading*.
- *Problem:* in different datasets D', D'' the same predicate *hasGravityReading* may not mean the same thing.
- *Existing solution:* use semantics.
- *Remaining problem:* concepts may still be slightly different.
- *Possible solution:* compare values $x'_1, \ldots, x'_n \in D'$ and $x''_1, \ldots, x''_n \in D''$ measured at the same locations.

2. Need to Take Uncertainty into Account

- *Problem* (reminder): check whether the predicate means the same in databases D' and D''.
- *Solution* (reminder): compare values $x'_1, \ldots, x'_n \in D'$ and $x''_1, \ldots, x''_n \in D''$ measured at the same locations.
- *Ideal case* (of exact values): if $\Delta x_i \stackrel{\text{def}}{=} x'_i - x''_i = 0$ for all i, the predicate means the same in D' and D''.
- *Problem:* due to measurement errors, the measurement result x'_i differs from the actual (unknown) value x_i:

$$\Delta x'_i \stackrel{\text{def}}{=} x'_i - x_i \neq 0.$$

- *Hence:* $\Delta x_i = (x'_i - x_i) - (x''_i - x_i) = \Delta x'_i - \Delta x''_i \neq 0.$
- *Traditional assumption:* $\Delta x'_i$ are normally distributed, with 0 mean and known standard deviation σ'_i.
- *Conclusion:* $\sigma_i^2 = (\sigma'_i)^2 + (\sigma''_i)^2 + 2r_i \cdot \sigma'_i \cdot \sigma''_i$, where $r_i \in [-1, 1]$ is the correlation between $\Delta x'_i$ and $\Delta x''_i$.

3. First Idea: Assume Independence

- *Reminder:* $\sigma_i^2 = (\sigma_i')^2 + (\sigma_i'')^2 + 2r_i \cdot \sigma_i' \cdot \sigma_i''$, with the unknown correlation r_i.
- *Usual approach:* assume independence: $r_i = 0$ and $(\sigma_i)^2 = (\sigma_i')^2 + (\sigma_i'')^2$.
- *Informal justification:*
  - *all we know:* $r_i \in [-1, 1]$;
  - *information is invariant* w.r.t. $T : r_i \to -r_i$;
  - *conclusion:* the selected r_i must be invariant: $Tr_i = r_i$, so $-r_i = r_i$, and $r_i = 0$.
- *Formal justification:* the Maximum Entropy approach.
- χ^2 *criterion:* if $\sum\limits_{i=1}^{n} \frac{(\Delta x_i)^2}{(\sigma_i')^2 + (\sigma_i'')^2} \le \chi^2_{n,\alpha}$, then the two datasets D' and D'' describe the same quantity.

4. An Alternative Idea: Worst-Case Estimations

- *Reminder:* $\sigma_i^2 = (\sigma_i')^2 + (\sigma_i'')^2 + 2r_i \cdot \sigma_i' \cdot \sigma_i''$, with the unknown correlation r_i.
- *Previous approach:* assume independence ($r_i = 0$).
- *Problem:* measurement errors may be correlated.
- *Property:* if data fit for some values σ_i, then it fits for larger values σ_i as well.
- *Resulting solution:* check the largest possible values σ_i.
- *Fact:* σ_i is largest when $r_i = 1$; then $\sigma_i^2 = (\sigma_i' + \sigma_i'')^2$.
- *New χ^2 criterion:* if $\displaystyle\sum_{i=1}^{n} \frac{(\Delta x_i)^2}{(\sigma_i' + \sigma_i'')^2} \le \chi^2_{n,\alpha}$, then the two datasets D' and D'' describe the same quantity.
- *Comment:* if this inequality is not satisfied, then the datasets describe somewhat different quantities.

5. Conclusion

- *Question:* are the semantically equivalent quantities in two databases D' and D'' actually the same?
- *Input:*
  - semantically annotated measurement results $x'_1, \ldots, x'_n \in D'$ and $x''_1, \ldots, x''_n \in D''$;
  - information about the measurement uncertainty: st.dev. σ'_i and σ''_i.
- *Case of independent measurement errors:* D' and D'' represent the same data $\Leftrightarrow \sum\limits_{i=1}^{n} \frac{(\Delta x_i)^2}{(\sigma'_i)^2 + (\sigma''_i)^2} \le \chi^2_{n,\alpha}$.
- *Alternative situation:* measurement errors may be correlated.
- *Recommendation:* D' and D'' represent the same data $\Leftrightarrow$ a weaker inequality holds: $\sum\limits_{i=1}^{n} \left(\frac{\Delta x_i}{\sigma'_i + \sigma''_i} \right)^2 \le \chi^2_{n,\alpha}$.

6. Acknowledgments

This work was supported in part:

- by National Science Foundation grant HRD-0734825, and EAR-0225670 and DMS-0532645 and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.