Selecting the Most Representative Sample is NP-Hard: Need for Expert (Fuzzy) Knowledge

J. Esteban Gamez[1], François Modave[1], and Olga Kosheleva[2]
Departments of [1]Computer Science and [2]Teacher Education
University of Texas, El Paso, TX 79968, USA
contact email olgak@utep.edu

1. Outline

- One of the main applications of fuzzy is to formalize the notions of "typical", "representative", etc.
- The main idea behind fuzzy: formalize expert knowledge expressed by words from natural language.
- In this talk, we show that
  - if we do not use this knowledge, i.e., if we only use the data,
  - then selecting the most representative sample becomes computationally difficult (NP-hard).
- Thus, the need to find such samples in reasonable time justifies the use of fuzzy techniques.

2. Introduction to the problem

- *In practice:* the population is often large, so we analyze a sample.
- *Examples:* poll, educational survey.
- *Idea:* the more "representative" the sample, the larger our confidence in the statistical results.
- *Requirement:* a representative sample should have the same averages as the population.
- *Example:* the same average age, average income, etc.
- *Additional requirement:* the sample should exhibit the same variety as the population.
- *Example:* the sample should include both poorer and reacher people.
- *Formalization:* a representative sample should have the same variance as the population.

3. Population: exact description

By a *population*, we mean a tuple

$$p \stackrel{\text{def}}{=} \langle N, k, \{x_{j,i}\}\rangle,$$

where:

- N is an integer; this integer will be called the population size;
- k is an integer; this integer is called the *number of characteristics*;
- $x_{j,i}$ $(1 \le j \le k, 1 \le i \le N)$ are real numbers;
- the real number $x_{j,i}$ will be called the *value* of the j-th characteristic for the i-th object.

4. Statistical characteristics

- Let $p = \langle N, k, \{x_{j,i}\}\rangle$ be a population, and let j be an integer from 1 to k.
- By the *population mean* E_j of the j-th characteristic, we mean the value $E_j = \dfrac{1}{N} \cdot \sum\limits_{i=1}^{N} x_{j,i}$.
- By the *population variance* V_j of the j-th characteristic, we mean the value

$$V_j = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_{j,i} - E_j)^2.$$

- For every integer $d \geq 1$, by the *central moment* $M_j^{(2d)}$ *of order* $2d$ of the j-th characteristic, we mean the value

$$M_j^{(2d)} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_{j,i} - E_j)^{2d}.$$

5. Sample

- Let N be a population size.
- By a *sample*, we mean a non-empty subset $I \subseteq \{1, 2, \ldots, N\}$.
- For every sample I, by its *size*, we mean the number of elements in I.
- By the *sample mean* $E_j(I)$ of the j-th characteristic, we mean the value $E_j(I) = \frac{1}{n} \cdot \sum_{i \in I} x_{j,i}$.
- By the *sample variance* $V_j(I)$ of the j-th characteristic, we mean the value $V_j(I) = \frac{1}{n} \cdot \sum_{i \in I} (x_{j,i} - E_j(I))^2$.
- For every $d \geq 1$, by the *sample central moment* $M_j^{(2d)}(I)$ *of order* $2d$ of the j-th characteristic, we mean the value

$$M_j^{(2d)}(I) = \frac{1}{n} \cdot \sum_{i \in I} (x_{j,i} - E_j(I))^{2d}.$$

6. Statistics

- Let $p = \langle N, k, \{x_{j,i}\}\rangle$ be a population, and let I be a sample.
- By an *E-statistics tuple* corresponding to p, we mean a tuple $t^{(1)} \stackrel{\text{def}}{=} (E_1, \ldots, E_k)$.
- By an *E-statistics tuple* corresponding to I, we mean a tuple $t^{(1)}(I) \stackrel{\text{def}}{=} (E_1(I), \ldots, E_k(I))$.
- By an *(E, V)-statistics tuple* corresponding to p, we mean a tuple $t^{(2)} \stackrel{\text{def}}{=} (E_1, \ldots, E_k, V_1, \ldots, V_k)$.
- By an *(E, V)-statistics tuple* corresponding to I, we mean a tuple $t^{(2)}(I) \stackrel{\text{def}}{=} (E_1(I), \ldots, E_k(I), V_1(I), \ldots, V_k(I))$.
- For every integer $d \geq 1$, we can similarly define a statistics tuple of order $2d$.

7. How to describe closeness

- By a *distance function*, we mean a mapping ρ that maps tuples t and t' into a real value $\rho(t, t')$ s.t.
  - $\rho(t, t) = 0$ for all tuples t and
  - $\rho(t, t') > 0$ for all $t \neq t'$.
- *Example:* Euclidean metric between the tuples $t = (t_1, t_2, \ldots)$ and $t' = (t'_1, t'_2, \ldots)$:

$$\rho(t, t') = \sqrt{\sum_j (t_j - t'_j)^2}.$$

8. Formulation of the problem

- Let ρ be a distance function.
- *E-sample selection problem* corresponding to ρ:
  - *Given:*
    * a population $p = \langle N, k, \{x_{j,i}\}\rangle$, and
    * an integer $n < N$.
  - *Find:* a sample $I \subseteq \{1, \ldots, N\}$ of size n for which the distance $\rho(t^{(1)}(I), t^{(1)})$ is the smallest possible.
- *(E, V)-sample selection problem* corresponding to ρ:
  - *Given:*
    * a population $p = \langle N, k, \{x_{j,i}\}\rangle$, and
    * an integer $n < N$.
  - *Find:* a sample $I \subseteq \{1, \ldots, N\}$ of size n for which the distance $\rho(t^{(2)}(I), t^{(2)})$ is the smallest possible.

9. Main results

- For every distance function ρ, the corresponding E-sample selection problem is NP-hard.
- For every distance function ρ, the corresponding (E, V)-sample selection problem is NP-hard.
- For every distance function ρ and for every $d \geq 1$, the $(2d)$-th order sample selection problem is NP-hard.

10. Auxiliary result

- *In our proofs:* we considered the case when the desired sample contains half of the original population.
- *In practice:* samples usually form a much smaller portion of the population.
- *A natural question:*
  - fix $2P \gg 2$, and
  - look for samples which constitute the $(2P)$-th part of the original population.
- *Result:* the resulting problems of selecting the most representative sample are still NP-hard.

11. Proof: main idea

- *Reminder:* NP-hard means that we can reduce every problem from a certain class NP to this one.
- *Usual proof:* reduce a known NP-hard problem to our problem.
- *Why this works:* transitivity of reduction.
- *Known NP-hard problem:* subset sum problem
  - *given:* positive integers $s_1, \ldots, s_m$,
  - *find:* $\varepsilon_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{m} \varepsilon_i \cdot s_i = 0$.
- *Reduction:* $N = 2n$, $k = 2$, $n = m$, and:
  - $x_{1,i} = s_i$ and $x_{1,m+i} = -s_i$ for all $i = 1, \ldots, m$;
  - $x_{2,i} = x_{2,m+i} = 2^i$ for all $i = 1, \ldots, m$.
- *We will show:* $\rho(t(I), t) = 0 \Leftrightarrow$ the original instance of the subset sum problem has a solution.

12. Proof (cont-d)

- *Reminder:* $x_{1,i} = s_i$ and $x_{1,m+i} = -s_i$ for $i = 1, \ldots, m$.
- *Reminder:* $x_{2,i} = x_{2,m+i} = 2^i$ for $i = 1, \ldots, m$.
- *Population as a whole:* $E_1 = 0$ and
  $E_2 = \dfrac{2 + 2^2 + \ldots + 2^m}{m}$.
- Since $|I| = m$, for $E_2(I) = E_2$ to be true, we must have $\sum\limits_{i \in I} x_{2,i} = 2 + 2^2 + \ldots + 2^m$.
- All terms in RHS are divisible by 4 except for 2.
- All $x_{2,i}$ are divisible by 4 except for $x_{2,1}$ and $x_{2,m+1}$, so I must have exactly one of them.
- Similarly, I must have exactly one of i and $m + i$.
- So, corr. value $x_{1,j(i)}$ is $\varepsilon_i \cdot s_i$ for some $\varepsilon_i \in \{-1, 1\}$.
- Thus, $E_1(I) = E_1 = 0$ means that $\sum\limits_{i=1}^{m} \varepsilon_i \cdot s_i = 0$. QED.