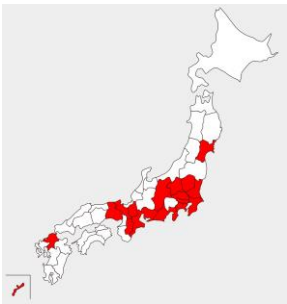
06 Is the data really biased?

Testing combinatorial correlation by decision diagrams

Abstract

We sometimes observe data with structures; population changes of cities on a map, traffic densities of roads on a traffic network, and reactions of sensors on a sensor network. Then, it is a natural question that the observations depend on the structure or not. Testing combinatorial correlation is a statistical method to answer the question: however, the test generally requires the exponential time because it considers all possible observations to evaluate the rarity of the current observation. In this research, we propose an efficient testing method using decision diagrams (DDs) that are a compact representation of a family of sets. We first compress the hypothesis patterns, which define the structure of the observations, by a DD and then construct another DD that compresses rare events to evaluate the rarity of the current observations. Our method reduces the testing time from 10^8 years to only 1 day in the case of testing binary observations on the Japanese prefecture map.

Combinatorial Correlation — Observations depend on Structure?



Ex1: Hotspot of Diseases

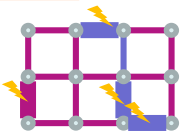
Red areas have so many patients.

White areas have a few patients.

This disease has

1. Hotspot (locality)? **Null Hypothesis**
2. No Hotspot (Bias)? **Alternative Hypothesis**

※ This observation is just an example.



Ex2: Sensor Network

Sensor Reaction

1. Intruder?
2. Noise?

Hist.	#
A B C	3
A B	5
B C	1
A	1
B	1

Ex3: Shopping History

Combination of A&B is

1. Popular?
2. Coincident?

Difficulty — Consider all possible observations to compute the P-value.

Combinatorial Correlation Testing

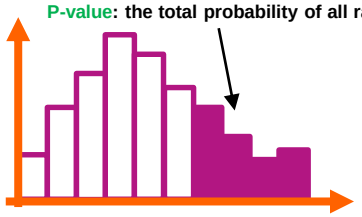
1. Get observations x and **hypothesis patterns** $\mathcal{F}$
2. Compute **Scan statistics** $K(x)$
3. Compute the **P-value** of $K(x)$ by **rare events** $\mathcal{W}$
4. Reject Alt. Hypo. if **P-value** ≦ Significance level (0.05)

$\mathcal{F}$ and $\mathcal{W}$ are exponentially huge. ⇒ Naïve computation takes 10^8 years!! Our method takes only 1 day!!

$p(K(x) = k)$

Rare Event: w **s.t.** $K(w) \geq K(x)$

P-value: the total probability of all rare events.



Histogram of K(x) considering all possible observations

Proposed method — Compute $K(x)$ and P-value on decision diagrams (DDs) of $\mathcal{F}$ and $\mathcal{W}$.

Scan Statistics

$$K(x) = \max_{S \in \mathcal{F}} \sum_{v \in S} x_v$$

Compute without Decompression

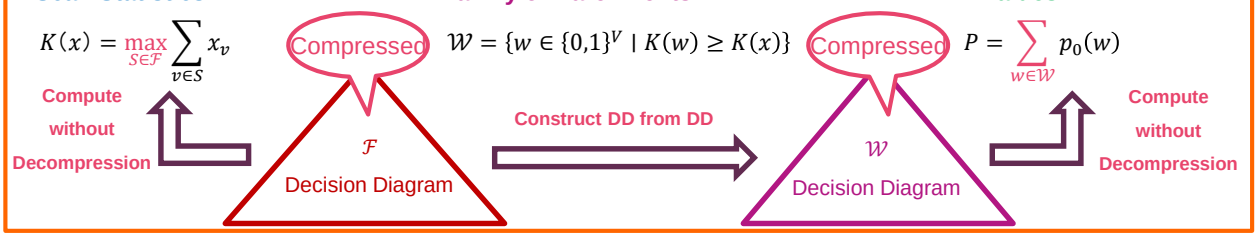
Family of Rare Events

$$\mathcal{W} = \{w \in \{0,1\}^V \mid K(w) \geq K(x)\}$$

P-values

$$P = \sum_{w \in \mathcal{W}} p_0(w)$$

Compute without Decompression



References

[1] M. Ishihata, T. Maehara: "Exact Bernoulli scan statistics by binary decision diagrams," *The 28th International Joint Conference on Artificial Intelligence (IJCAI-2019)*, 2019.

Contact

Masakazu Ishihata Email: cs-openhouse-ml@hco.ntt.co.jp

Learning and Intelligent Systems Research Group, Innovative Communication Laboratory

オープンハウス 2020