A First Step Towards Even More Sparse Encodings of Probability Distributions

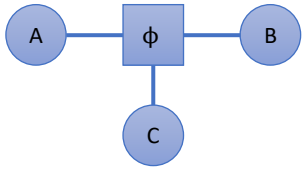
Florian Marwitz, Tanya Braun, Ralf Möller

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR INFORMATIONSSYSTEME

Probability Distributions

Model $G, P_G = \frac{1}{Z}\prod_{f \in gr(G)} f$

- Model real world scenarios with uncertainties
- Parfactors: Sparse encoding by factorization to encode (conditional) independences and first-order logic (FOL) to encode graph symmetries
- Alternative: MLNs with a sparse encoding using weighted FOL formulas



Overview: Compact Formula Extraction

1. Reduction using Quantiles and Clustering
2. Formula Extraction (canonically)
3. Minimization, e.g. Quine-McCluskey

Problem: Exponential Number of Entries/Weights

- Canonical transformation of a parfactor into formulas
  - Adds one formula for each entry in parfactor
  - ➢ Exponential number of formulas

Solution: Compact Formula Extraction

- Idea: Reduction, fewer different numbers
  - Summarize multiple rows
  - Minimize corresponding formula
  - Keep distance at most ϵ
- Quantiles
  - Calculate q-quantiles
  - Map each number to mean of quantile
  - Increase q until distance is at most ϵ
- Clustering
  - Cluster numbers in potential
  - Map each number to mean of cluster

Example

Example Factor

$a \in \{0,1\}$	$b \in \{0,1\}$	$c \in \{0,1\}$	$\phi(a,b,c)$
0	0	0	1
0	0	1	4.7
0	1	0	4.8
0	1	1	4.9
1	0	0	5
1	0	1	5.1
1	1	0	5.2
1	1	1	5.3

- Canonically
  - ➢ Eight formulas
- Quartiles
  - Odd row and following
  - ➢ Four formulas
- Clustering
  - Cluster 1: First row
  - Cluster 2: Other rows
  - ➢ Two formulas

Empirical Results

- Datasets Smokers and artificial, in which parfactor i=1…9 has i−1 ones and 9−i twos

Results

Test	S1	S2	A1	A2
σ	0.5	1	0.1	0.2
ϵ	0.3	0.3	0.05	0.1
d_{noisy}	0.1	0.27	0.03	0.065
d_{mapped}	0.01	0.32	0.015	0.08
#	2	2	1-2	1-2
L	3	4-7	1-4	1-4
E	0.004	0.31	0.01	0.021

- Distance and Reconstruction
  - d_{noisy}, d_{mapped}: Hellinger distance to noised and mapped model
- Sparsity and Error
  - Formulas per parfactor (#), atoms per formula (L), mean abs. error (E)