ISCAS 中国科学院软件研究所学术年会'2022 暨计算机科学国家重点实验室开放周
学术论文

Learning DMEs and DGMEs From Both Positive and Negative Examples

Yeting Li, Haiming Chen, Zixuan Chen
The Computer Journal, 2022, Yeting Li, 15801685206, liyt@ios.ac.cn

Background & Motivation

- XML documents with corresponding schemas on the Web only account for 30.2%, with the proportion of 24.5% for valid ones.
- We focus on the inference of disjunctive multiplicity schema (DMS).
- The central task of DMS inference is learning disjunctive multiplicity expression (DME). Previously, DME learning has been studied from positive examples.

Contributions

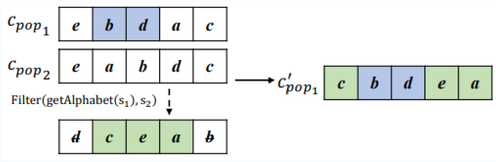
- We devise an algorithm called iDME based on a genetic algorithm for learning DMEs from positive examples S_+ and negative examples S_-.
- We propose a new subclass that allows numeric occurrences called disjunctive generalized multiplicity expressions (DGMEs). We provide an algorithm called iDGME for learning DGMEs.
- Results show that with only S_+, our algorithm can learn a DME that accepts all positive examples. And when given S_+ and S_-, we can learn DMEs or DGMEs with high accuracy.

Approach: Based on Genetic Algorithms

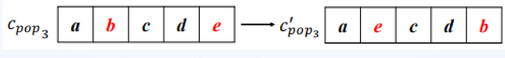
- encoding: DME $r \rightarrow (c_{pop}, m_{pop}, \tau)$
- measurement:

$$K(r, S_+, S_-) = \frac{|T_P| + |T_N|}{|S_+| + |S_-|} \qquad CC(r) = n! * \prod_{i=1}^{n} |D_i|$$

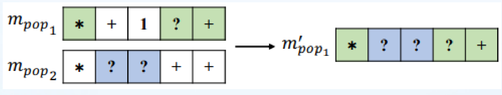
- c_{pop} crossover (iDME):



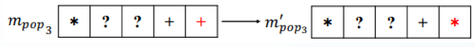
- c_{pop} mutate (iDME):



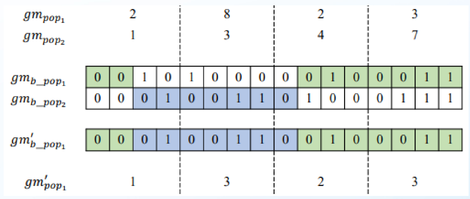
- m_{pop} crossover (iDME):



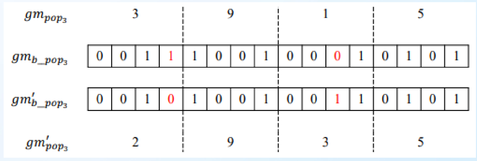
- m_{pop} mutate (iDME):



- m_{pop} crossover (iDGME):



- m_{pop} mutate (iDGME):



Evaluation

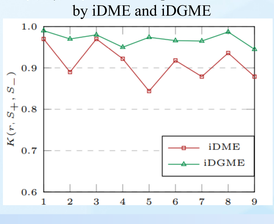
Success rate on synthetic target DMEs under different proportion of S_+ and S_-.

No.	# successes	rate	$\lvert S_+\rvert$	$\lvert S_-\rvert$
1	6	11.54%	25	0
2	10	19.23%	50	0
3	13	25.00%	75	0
4	20	38.46%	200	0
5	24	46.15%	250	0
6	25	48.08%	300	0
7	12	23.08%	25	25
8	16	30.77%	50	25
9	17	32.69%	75	25
10	25	48.08%	200	25
11	29	55.77%	250	25
12	30	57.69%	300	25
13	17	32.69%	25	75
14	20	38.46%	50	75
15	23	44.23%	75	75
16	32	61.54%	200	75
17	34	65.38%	250	75
18	36	69.23%	300	75
19	33	63.46%	200	100
20	35	67.31%	250	125
21	37	71.15%	300	150
22	35	67.31%	200	200
23	39	75.00%	250	250
24	43	82.70%	300	300

The learned DMEs of $learner^{+}_{DME}$ & iDME.

$\lvert\Sigma\rvert$	$\lvert S_+\rvert$	$learner^{+}_{DME}$	iDME	K_1	K_2
5	100	$(a^+\|e) \parallel (c^7\|d^7) \parallel b^*$	$(a^+\|e) \parallel (c^7\|d) \parallel b^*$	1	1
	500	$(a^+\|c^7\|c^*) \parallel b^+ \parallel d^7$	$(a^7\|c\|c^*) \parallel b^+ \parallel d^7$	1	1
	1000	$(a^7\|d^*) \parallel b^* \parallel c^* \parallel e^+$	$(a\|d^*) \parallel b^* \parallel c^* \parallel e^+$	1	1
10	100	$(a^7\|b^*\|c^7\|g^*\|h^*\|i^*\|j^*) \parallel d^+ \parallel e^7 \parallel f^*$	$(a\|b^+\|c\|g^+\|h^+\|i^+\|j^*) \parallel d^+ \parallel e^7 \parallel f^*$	1	1
	500	$(c^*\|g^*\|h^*\|i^*\|j^*) \parallel (a^*\|b^*\|d^*) \parallel e^* \parallel f^*$	$(a^+\|b^+\|c^+\|d^*) \parallel (e^+\|g^+\|j^*) \parallel (f^+\|h^+\|i^+)$	1	1
	1000	$(a^*\|b^*\|c^*\|d^*\|e^*\|f^*) \parallel g^* \parallel h^* \parallel i^* \parallel j^*$	$(a^+\|b^+\|c^+\|d^*) \parallel (e^+\|g^+\|j^*) \parallel (f^+\|h^*) \parallel i^*$	1	1
15	100	$(d^*\|e^*\|f^*\|l^*\|m^*\|n^*\|o^*) \parallel (f^*\|g^*\|i^*\|k^*) \parallel (a^*\|c^*\|h^*) \parallel (b^*\|j^*)$	$(a^+\|c^+\|h^+\|m^+\|o^*) \parallel (b^+\|d^+\|e^+\|j^+\|n^*) \parallel (f^+\|g^+\|i^+\|k^+\|l^*)$	1	1
	500	$(d^*\|e^*\|f^*\|l^*\|m^*\|n^*) \parallel (a^*\|c^*\|h^*\|o^*) \parallel (g^*\|i^*\|k^*) \parallel (b^*\|j^*)$	$(a^+\|c^+\|h^+\|m^+\|o^*) \parallel (b^+\|d^+\|e^+\|j^+\|n^*) \parallel (f^+\|g^+\|i^+\|k^+\|l^*)$	1	1
	1000	$(a^7\|b^7\|c^7\|d^7\|e^*) \parallel f^* \parallel g^7 \parallel h^* \parallel i^* \parallel j^* \parallel k^7 \parallel l^* \parallel m^* \parallel n^* \parallel o^+$	$(c\|d\|e^+\|k^7) \parallel (a\|b\|g^7) \parallel f^* \parallel h^* \parallel i^* \parallel j^* \parallel n^* \parallel m^* \parallel o^+ \parallel l^*$	1	1

$K(r, S_+, S_-)$ value of expressions inferred by iDME and iDGME



- If only S_+ are available, iDME may learn an over-generalized DME, and the addition of S_- is very useful for learning DMEs.
- With only S_+, both our algorithm and the state-of-the-art algorithm can learn a DME that accepts all S_+.
- The results learned by iDGME accept more S_+ and reject more S_- than that of iDME. Namely, iDGME can learn expressions more accurately.

Conclusion

We provided an algorithm iDME to learn a DME from S_+ and S_- based on genetic approaches. We extended the subclass DME to DGME which allows numeric occurrences, and developed the algorithm iDGME to learn a DGME.