

최대 절단 문제에서 간선기반 이진 인코딩을 통한 대리모델 유전알고리즘

유서아, 김용혁

광운대학교 컴퓨터과학과

9380187@naver.com, yhdfly@kw.ac.kr

A Surrogate-assisted Genetic Algorithm through an Edge-based Binary Encoding for the MAX-CUT

Seo-Ah Yu, Yong-Hyuk Kim

KwangWoon Univ., Dept. Comp Sci.

요 약

본 논문은 그래프 문제에 적용할 수 있는 ‘최대 절단 문제(Max-Cut Problem)’기반의 새로운 학습 데이터 인코딩을 통해 기계학습 모델을 훈련시키고, 이 모델을 활용하여 유전알고리즘의 적합도 계산 부분을 기계학습 모델로 대체하는 대리모델 유전알고리즘을 설계한다. 이 논문은 최대 절단 문제를 중점적으로 다루고 있으며, 이 문제는 그래프 이론에서 중요한 문제 중 하나이다. 본 논문에 사용한 최대 절단 문제 기반의 인코딩 방식은 기계학습모델의 근사 성능에 긍정적 영향을 미친다. 향상된 기계학습모델의 근사 성능이 대리모델 유전알고리즘의 최적 해를 찾는 능력에도 긍정적인 영향을 미침을 보인다.

I. 서론

본 논문에서는 최대 절단 문제를 기반으로 한 학습 데이터 인코딩을 통해 기계학습 모델을 훈련시키고, 이 모델을 이용하여 유전알고리즘의 적합도 계산 부분을 대체하는 대리모델 유전알고리즘을 설계한다. 이로써 기계학습 모델의 근사 성능 향상이 대리모델 유전알고리즘이 최적 해를 찾는 능력에도 긍정적 영향을 미침을 실험적으로 입증한다.

최대 절단 문제는 무 방향 그래프에서 간선을 잘라 두 부분으로 분할했을 때, 절단되는 간선의 개수가 최대가 되도록 분할하는 문제이다. 이 문제는 NP-난해(NP-Hard) 문제로 알려져 있고, 최적 해를 찾기 위해서는 지수 시간이 필요하다. 따라서 이러한 NP-난해 문제를 근사 알고리즘을 이용하여 효율적으로 근사하는 것이 중요하다. 대리모델 유전알고리즘은 유전알고리즘에 일부분을 기계학습 모델로 대체한 유전알고리즘으로, 본 논문에서는 유전알고리즘에서 가장 시간이 많이 걸리는 적합도 계산 부분을 기계학습 모델로 대체한다. 대리모델 유전알고리즘은 계산 비용을 감소시키거나, 빠른 수렴에 도움을 줄 수 있고, 다양한 문제에도 적용 가능하다는 장점이 있다[1]. 본 논문에 사용된 인코딩 방식은 정점의 정보만을 제공하는 기존의 인코딩 방식과 다르게 간선의 정보를 추가로 제공하여 그래프의 연결성을 명확히 나타낼 수 있다[2]. 이로 인해 기계학습 알고리즘들이 문제를 효과적으로 학습할 수 있도록 돕는다. 이 논문은 그래프 문제에서 효과적인 학습 데이터 인코딩 방식을 통해 기계학습 모델의 근사 성능을 향상시키고, 이 모델을 활용하여 설계한 대리모델 유전알고리즘이 최적 해를 찾는 능력이 뛰어난 것을 보인다. 기계학습 모델과 대리모델 유전알고리즘의 성능 간 관계를 밝힘으로써 기존 유전알고리즘의 단점을 보완할 대리모델 유전알고리즘의 중요성과 가능성을 강조한다. 또한, 다른 그래프 문제에도 적용할 수 있는 확장 가능성에 대한 논의를 제공하고 있다. 이는 그래프 이론과 대리모델 유전알고리즘의 융합으로 새로운 연구 영역의 가능성을 보인다.

II. 본론

본 논문에서는 서로 다른 밀도를 가지는 4가지의 그래프로 최대 절단 문제 기반 인코딩을 사용한 데이터 세트를 생성하였다. 기존의 인코딩을 VE(Vertex encoding), 본 논문에 사용된 인코딩을 VEE(Vertex edge encoding)으로 명명한다. VE는 공간배열로, 각 공간을 나타내는 n 개의 요소로 이루어져 있다. 배열의 각 요소는 0 또는 1 중에서 랜덤하게 선택된다. 예를 들어 “110”의 경우 정점 1과 2는 같은 공간에 정점 3은 다른 공간에 속하는 것을 의미한다. VEE는 정점의 개수 n 으로 $n \times n$ 크기의 간선행렬을 생성한다. 간선행렬의 인덱스 (i, j) 에 ‘1’로 표시되면, 정점 i 와 정점 j 사이에 간선이 존재하는 것을 의미한다. 두 정점 사이에 간선이 존재하는 경우, 공간배열에서 해당하는 두 정점의 요소를 XOR 연산하여 절단 행렬을 생성한다. 이렇게 생성한 절단행렬을 앞선 공간배열 뒤에 덧붙여 모델 학습데이터로 사용한다. 그래프는 간선의 개수가 정점의 0.5배인 G1, 1배인 G2, 2배인 G3, 3배인 G4로 구성하였으나, 본 논문에서는 공간상의 이유로 G1, G4의 결과만을 비교한다. 이 데이터 세트를 이용하여 세 가지 다른 기계학습 모델, 선형회귀(Linear regression, LR), 서포트 벡터 회귀(Support vector regression, SVR), 가우시안 프로세스 (Gaussian process, GP)를 훈련시켰다. 정점의 개수인 차원은 30, 100, 500 세 가지로 진행하였다. 학습된 기계학습 모델로 대리모델 유전알고리즘을 설계하고, 가장 우수한 해를 최적해로 설정하였다. 본 논문에 적용된 인코딩의 유용성을 밝히기 위해 VE로 학습된 대리 모델과 VEE로 학습된 대리 모델의 최적 해를 비교했다. VE의 결과와 VEE의 결과의 차이가 유의미한지 판단하기 위해 t-검정을 수행하였다. 또한, 대리모델 유전알고리즘의 가능성을 강조하기 위해 일반 유전알고리즘의 최적해와 대리모델 유전알고리즘의 해를 비교한다. 본 논문의 유전알고리즘은 일 점 교차를 이용하고, 돌연변이 비율은 0.015이다. 자손의 상위 50개를 부모의 하위 50개와 대체하는 방식으로 세대교체를 한다. 실험은 30회 반복하여 평균을 기준으로 분석함으로써 실험의 일관성을 가졌다.

III. 결론

본 논문에서는 먼저 각 모델 별 VE와 VEE의 근사성능 차이를 관찰하고 이 모델을 이용하여 설계한 대리모델 유전알고리즘의 최적 해를 비교한다. 근사성능의 평가지표는 RMSE이다. 표 1을 보았을 때, 정점의 개수가 적은 30차원의 일부를 제외하고 모든 경우에서 VEE가 VE보다 작은 RMSE를 보인다. 이는 새로운 최대 절단 문제 기반의 인코딩이 기계학습 훈련에 긍정적 영향을 주는 것이라고 볼 수 있다. VEE의 RMSE를 비교하였을 때, LR모델의 근사성능이 월등히 좋았으며, 그다음으로 GP, 비교적 안 좋은 성능을 보인 것은 SVR이었다. 하지만 SVR이 다른 것과 비교적으로 낮은 성능일 뿐, 작은 오차로, 근사가 잘 이루어지지 않았다고 볼 수 없다. 표 2의 진한(Bold)부분은 VE보다 좋은 최적 해를 보인 VEE의 최적해 값을 나타내고 있다. G1부터 G4를 통틀었을 때 36가지 경우에서 26개의 경우가 VE보다 VEE가 좋은 최적 해를 찾았으므로 약 72%경우에서 VEE가 VE보다 대리 모델이 최적 해를 찾는 능력에 도움을 준다고 할 수 있다. 특히 근사에서 월등한 성능을 보인 LR모델의 경우 일반 유전알고리즘보다 더 좋은 최적 해를 찾는 경우도 있었다. 이는 대리모델 유전알고리즘이 일반 유전알고리즘의 단점을 보완하기 위해 대체제로 사용될 수 있는 가능성을 보여준다.

결론적으로, 새로운 최대 절단 문제 기반의 인코딩(VEE)이 기계학습 훈련에 긍정적인 영향을 미치며, 대리모델 유전알고리즘이 최적 해를 효과적으로 찾는데 기여할 수 있음을 실험적으로 입증한다.

향후 다른 그래프 문제로의 확장성을 확인하기 위해, 최대 부분 그래프 문제(Max-clique problem) 등에 본 실험을 적용해 볼 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1048466).

참 고 문 헌

- [1] Seo-Ah Yu, & Kim Yong-hyuk (2023). Investigation on the Performance of Surrogate-assisted Genetic Algorithms According to Changes in Learning Model and Data. 아시아태평양융합연구교류 논문지, 9(6), 29-38, <http://dx.doi.org/10.47116/apjcri.2023.06.03>
- [2] Seo-Ah Yu, & Yong-Hyuk Kim (2023). Fitness Approximation of Machine Learning Models through a Novel Binary Encoding in Graph Problems. Korea Artificial Intelligence Conference.

표 1. G1, G4 그래프의 각 기계학습 모델별 근사 오차 비교

G1				
RMSE		30	100	500
SVR	VE	1.37E-01	1.21E+00	7.57E+00
	VEE	1.55E-01	4.07E-01	7.99E-01
LR	VE	2.29E+00	3.71E+00	7.81E+00
	VEE	4.12E-15	8.60E-15	3.50E-14
GP	VE	9.22E-07	7.37E-02	7.60E+00
	VEE	2.15E-08	2.06E-04	9.32E-03

G4				
RMSE		30	100	500
SVR	VE	2.96E-01	2.70E+00	1.78E+01
	VEE	4.77E-01	8.30E-01	3.91E+00
LR	VE	4.21E+00	8.17E+00	1.83E+01
	VEE	1.38E-14	2.97E-14	1.24E-13
GP	VE	1.05E-05	3.92E-01	1.79E+01
	VEE	5.19E-05	2.32E-01	2.24E+00

표 2. G1, G4 그래프를 이용한 각 대리모델 유전알고리즘과 일반 유전알고리즘의 최적해 비교

		30	100	500		
GA	Best	23	58	209		
	Ave	22.93	56.77	203.13		
	Var	6.22e-02	7.12e-01	9.65e+00		
	Dev	2.49e-01	8.44e-01	3.11e+00		
SVR	VE	Best	21	42	147	
		Ave	19.17	38.50	132.77	
		Var	8.06e-01	5.52e+00	6.58e+01	
		Dev	8.98e-01	2.35e+00	8.11e+00	
	VEE	Best	21	44	158	
		Ave	19.00	39.60	149.17	
		Var	7.33e-01	3.77e+00	1.23e+01	
		Dev	8.56e-01	1.94e+00	3.50e+00	
p-value		4.72e-01	5.70e-02	2.27e-12		
LR	VE	Best	0	2	83	
		Ave	0.00	0.77	71.37	
		Var	0.00e+00	3.79e-01	2.86e+01	
		Dev	0.00e+00	6.16e-01	5.34e+00	
	VEE	Best	23	58	209	
		Ave	22.97	56.67	203.27	
		Var	3.22e-02	6.22e-01	6.20e+00	
		Dev	1.80e-01	7.89e-01	2.49e+00	
	p-value		1.16e-62	7.96e-90	6.29e-54	
	GP	VE	Best	21	45	142
			Ave	18.92	40.80	128.64
			Var	1.03e+00	3.12e+00	4.32e+01
Dev			1.02e+00	1.77e+00	6.57e+00	
VEE		Best	22	43	158	
		Ave	19.12	40.24	151.84	
		Var	1.23e+00	3.62e+00	9.65e+00	
		Dev	1.11e+00	1.90e+00	3.11e+00	
p-value		5.18e-01	2.96e-01	3.76e-17		

		30	100	500		
GA	Best	75	237	951		
	Ave	73.77	231.63	935.07		
	Var	1.85e+00	1.21e+01	6.63e+01		
	Dev	1.36e+00	3.48e+00	8.14e+00		
SVR	VE	Best	67	189	785	
		Ave	61.57	169.40	756.20	
		Var	3.18e+00	8.50e+01	2.75e+02	
		Dev	1.78e+00	9.22e+00	1.66e+01	
	VEE	Best	65	185	827	
		Ave	62.10	177.07	803.40	
		Var	2.61e+00	1.22e+01	1.11e+02	
		Dev	1.62e+00	3.49e+00	1.05e+01	
p-value		2.38e-01	1.67e-04	1.90e-17		
LR	VE	Best	0	17	516	
		Ave	0.00	9.53	474.73	
		Var	0.00e+00	1.37e+01	4.00e+02	
		Dev	0.00e+00	3.70e+00	2.00e+01	
	VEE	Best	75	238	960	
		Ave	74.10	229.57	938.93	
		Var	2.02e+00	1.58e+01	6.63e+01	
		Dev	1.42e+00	3.98e+00	7.97e+00	
	p-value		2.39e-51	7.81e-86	4.55e-50	
	GP	VE	Best	67	184	790
			Ave	63.04	176.92	752.64
			Var	4.04e+00	1.23e+01	3.30e+02
Dev			2.01e+00	3.51e+00	1.82e+01	
VEE		Best	66	185	827	
		Ave	62.28	176.44	803.40	
		Var	2.84e+00	1.66e+01	7.98e+01	
		Dev	1.69e+00	4.07e+00	8.93e+00	
p-value		1.62e-01	6.64e-01	3.07e-14		