Regular paper | Journal of the KIECS. pp. 895-900, vol. 18, no. 5, Oct. 31. 2023, t. 121, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.5.895

임계값 적용에 기반한 저 복잡도 그래프 신호 샘플링 알고리즘

김윤학*

Low-Complexity Graph Sampling Algorithm Based on Thresholding

Yoon-Hak Kim*

요 약

그래프에서 전체 노드중 일부 노드를 선택하고 선택된 노드에서 발생하는 신호로부터 원 신호를 복원하는 저 복잡도를 갖는 샘플링 기술에 대해 연구한다. 이를 위해, 매 단계에서 한 개의 노드를 선택하는 탐욕 알고리즘을 기반으로, 기존의 방식인 최소 비용값을 갖는 노드를 찾기 위해 전체노드를 탐색 및 계산하는 방식을 취하지 않고 임계값을 설정하여 임계값 이하의 값을 갖는 노드를 선택하도록 하여 탐색 및 비용함수 계산비용을 줄이는 방식을 제안한다. 복원성능 저하를 막기 위해 정확한 임계값 설정이 요구되며, 이를 위해 임계값을 구하기 위한 매 단계에서 샘플링 집합 크기와의 관계식을 정립한다. 다양한 그래프상에서 복원성능 및 복잡도 (실행시간) 평가를 수행하여, 기존 방식 대비 복원성능을 유지하면서 평균 1.3배 이상 빠른 실행시간이 보임을 확인했다.

ABSTRACT

We study low-complexity graph sampling which selects a subset of nodes from graph nodes so as to reconstruct the original signal from the sampled one. To achieve complexity reduction, we propose a graph sampling algorithm with thresholding which selects a node with a cost lower than a given threshold at each step without fully searching all of the remaining nodes to find one with the minimum cost. Since it is important to find the threshold as close to a minimum cost as possible to avoid degradation of the reconstruction performance, we present a mathematical expression to compute the threshold at each step. We investigate the performance of the different sampling methods for various graphs, showing that the proposed algorithm runs 1.3 times faster than the previous method while maintaining the reconstruction performance.

키워드

Graph Signal Processing, Graph Sampling, Greedy Algorithm, Thresholding
그래프 신호 처리, 그래프 샘플링, 탐욕 알고리즘, 임계값 적용

* 교신저자 : 조선대학교 전자공학부
• 접 수 일 : 2023. 08. 16
• 수정완료일 : 2023. 09. 13
• 게재확정일 : 2023. 10. 17
• Received : Aug. 16, 2023, Revised : Sep. 13, 2023, Accepted : Oct. 17, 2023
• Corresponding Author : Yoon Hak Kim
Dept. Electronic Engineering, Chosun University,
Email : yhk@chosun.ac.kr

Ⅰ. 서론

네트워크 시스템에서 발생하는 데이터는 고차원적이고 비정형적 구조를 갖는 특성을 보이며, 이러한 네트워크 데이터를 표현하고 처리하기 위해 그래프 신호처리 분야에 대한 다양한 연구가 폭넓게 이루어지고 있다[1-2]. 그래프 샘플링은 많은 수의 노드를 갖는 네트워크 시스템에서 주어진 응용목적에 맞는 최적의 부분 노드들을 선택하는 기술로 그래프 신호처리 분야에서 중요한 위치를 차지하고 있으며, 샘플링된 신호로부터 오차를 최소화하면서 원 신호를 복원하기 위한 다양한 샘플링 기술들이 개발되고 있다[3-10]. 복잡도를 줄이기 위해 직접적인 비용함수인 복원 오차 대신 상한값을 최소화하면서 노드들을 선택하는 탐욕적(greedy) 반복 알고리즘이 제시되었고[3-6], 샘플링을 위한 행렬 구성에 필요한 고유값 분해(eigen-decomposition) 없이 샘플링을 하는 저 복잡도 기술이 개발되었다[7-8]. 복원성능을 위해 복원 오차를 비용함수로 하는 QR분해에 기반한 탐욕적 샘플링 알고리즘이 제안되었으며[9], 최근에는 가중치를 갖는 그래프 신호에 관한 샘플링 방식에 대해 연구가 진행되었다[10].

본 논문에서는 매 단계 한 개의 최적 노드를 선택하는 저 복잡도 탐욕(greedy) 알고리즘을 제안한다. 기존의 탐욕 알고리즘은 노드 선택시 남아있는 모든 노드에 대해 비용함수를 계산하고 이들 중 최소의 비용함수 값을 갖는 노드를 선택하게 되는데, 이는 노드 수가 많은 경우 막대한 탐색 및 계산비용을 초래하게 된다. 본 논문에서는 현 단계에서 다음 단계의 최적 노드가 갖는 비용함수 값에 대한 예측을 통해 임계값을 설정하여, 다음 단계에서 임계값보다 낮은 비용함수 값을 갖는 노드 발견 시 탐색을 중단하고 해당 노드를 최적 노드로 선택하여 복잡도를 개선하는 방식을 제안한다. 이 경우 임계값 설정에 대한 정확도가 복원성능에 영향을 주게 되는데, 이를 위해 임계값을 구하기 위한 매 단계에서 샘플링 집합 크기와의 관계식을 정립한다.

본 논문에서는 기존에 제안한 QR 분해에 기반한 샘플링 기술[9]에 제안 알고리즘을 적용하여, 다양한 그래프상에서 다른 샘플링 방식과 성능을 비교 평가한다. 기존 기술 대비 복원성능을 유지하면서 복잡도(실행시간)를 개선하는 효과를 보임을 확인한다.

본 논문은 2장에서 그래프 및 복원오차에 대한 설명과 문제를 정립하며, 3장에서는 복잡도를 개선하기 위한 임계값 설정 방식과 이를 적용한 샘플링 알고리즘을 제안한다. 다양한 그래프상에서 다른 알고리즘과의 성능평가가 4장에서 수행되고, 5장의 결론으로 마무리한다.

Ⅱ. 문제정립

N개의 노드를 갖는 네트워크 시스템이 i번째 노드와 j번째 노드의 연결이 가중치 e_{ij}를 갖는 경우, 그래프 $G(V,E)$로 표현가능하며, 여기서 $V=\{1,\cdots,N\}$이고, $E=\{i,j,e_{ij}\}$를 의미한다. 이 시스템에서 전체 노드들의 신호값은 그래프 신호 $\boldsymbol{f}=[f_1 \cdots f_N]^T \in \boldsymbol{R}^{\boldsymbol{N}}$로 나타낼 수 있으며, f_i 는 i번째 노드에서의 신호값을 나타낸다. 비정형적 구조를 갖는 그래프 신호의 변화는 $N\times N$ 행렬인 그래프 라플라시안 $\boldsymbol{L}$(combinatorial graph Laplacian 또는 normalized Laplacian) 등을 통해 표현할 수 있으며, 그래프 신호는 그래프 라플라시안의 고유벡터와의 선형결합 $\boldsymbol{f}=\boldsymbol{Uc}$ 으로 표현가능하다. 여기서 $\boldsymbol{U}$는 N개의 고유벡터 $\boldsymbol{u_i}$을 열벡터(column vector)로 갖는 고유벡터 행렬 $\boldsymbol{U}=[\boldsymbol{u}_1 \cdots \boldsymbol{u}_N]$ 이고, $\boldsymbol{c}=\boldsymbol{U}^{-1}\boldsymbol{f}$ 는 그래프 푸리에 변환(graph Fourier transform, GFT)을 의미한다[1]. 그리고, ω-대역폭제한 그래프 신호인 경우 식 (1)과 같이 선형결합으로 나타낼 수 있다:

$$\boldsymbol{f}=\sum_{i=1}^{r} c_i \boldsymbol{u_i}=\boldsymbol{U_{VR}}\boldsymbol{c_R} \quad \cdots (1)$$

여기서, c_i는 그래프 푸리에 변환 $\boldsymbol{c}$의 i번째 요소이며 $c_i=0, |\lambda_i|>\omega, \forall_{i>r}$ 이고 λ_i는 i번째 고유값(eigenvalue)이다. $\boldsymbol{c_R}=[c_1 \cdots c_r]^T$ 은 $r\times 1$ 열벡터이고, 아래첨자 $R=\{1,...,r\}$은 $\boldsymbol{c}$로 부터 $\boldsymbol{c_R}$를 구성하는데 사용한 요소들의 인덱스 집합을 나타낸다. 본 논문에서는 행렬과 벡터에 사용되는 아래첨자를 동일한 방식으로 해석한다. $\boldsymbol{U_{VR}}=[\boldsymbol{u}_1 \cdots \boldsymbol{u}_r]$ 또한 아래첨자인 집합 V와 R에 속해있는 인덱스에 해당되는 $\boldsymbol{U}$의 행과 열로 구성된 $N\times r$행렬을 표현한다.

그래프 샘플링에서 주목할 점은, N개의 행벡터에서 r개의 독립적인 행벡터(independent row vector)를 $\boldsymbol{U}_{VR}$ 로부터 선택할 수 있는데, 이렇게 선택된 행벡터의 인덱스에 해당하는 노드들로 샘플링 집합 S를 구성할 수 있으며, 이때 ω-대역폭 제한 무잡음 그래프 신호인 경우 샘플링된 신호 $\boldsymbol{f}_S$로부터 복원오차(Mean Squared Error, MSE) 없이 복원이 가능하다[4]. 또한 $\boldsymbol{f}_S$의 측정잡음이 서로 독립이고 정규분포(normal distribution) $N(0,\sigma^2\boldsymbol{I}), \sigma=1$를 갖는 가산성(additive) 잡음이라고 가정한 경우, 선택된 |S|개의 독립적인 행벡터로 구성된 $|S|\times r$ 행렬 $\boldsymbol{U}_{SR}$를 사용하여 식 (2)와 같이 복원오차를 구할 수 있다[4]:

$$MSE = tr\left[(\boldsymbol{U}_{SR}^{T}\boldsymbol{U}_{SR})^{-1}\right] \quad \cdots (2)$$

결국, 그래프 샘플링 문제는 전체 N개의 행벡터로부터 r개의 독립적인 행벡터를 선택하는 다양한 조합의 수가 존재하는데, 이 중에 어떻게 복원오차 (2)를 최소화하는 조합을 선택하는가 하는 문제가 된다. 전체 N개의 노드에서 r개의 노드를 한 번에 선택하는 방식은 막대한 계산비용을 초래함으로, 한 번에 한 개의 최적 노드를 선택하는 탐욕적 접근방식을 취하게 되는데 이를 수식적으로 식 (3), (4)와 같이 나타낼 수 있다:

$$j^* = \arg\min_{\boldsymbol{u}_j, j\in S_i^C} tr\left(\left(\boldsymbol{U}_{S_i R}^{T}\boldsymbol{U}_{S_i R}\right)^{+}\right) \quad \cdots (3)$$

$$S_i^* = S_{i-1}^* + \{j^*\},\ i=1,...,|S| \quad \cdots (4)$$

수식 (3)은 i번째 단계에서 남아있는 전체노드 $S_i^C = V - S_{i-1}^*$로부터 복원오차를 최소화하는 최적의 노드 j^*번째 노드를 선택하는 작업을 수식적으로 표현하며, $\boldsymbol{U}_{S_i R}^{T} = [\boldsymbol{u}^{(1)}\cdots\boldsymbol{u}^{(i-1)}\boldsymbol{u}_j]$ 이고 $\boldsymbol{u}^{(i)}$는 i번째 단계에서 선택된 최적의 행벡터를 의미한다. 수식 (4)는 최적의 노드로 선택된 j^*번째 노드를 i번째 단계에서의 샘플링 집합 S_i^*에 추가하는 작업을 의미한다.

III. 저 복잡도 샘플링 알고리즘

샘플링 집합 선택은 조합문제로 막대한 계산비용을 요구하며, 이를 해결하기 위해 사용되는 탐욕 알고리즘은 상당 수준 계산량을 줄였으나, N이 큰 경우 여전히 높은 수준의 복잡도를 가지게 된다. 이는 매 단계에서 하나의 최적의 노드를 선택하기 위해 남아있는 모든 노드에 대해 비용함수를 계산하는 데에 그 원인이 있다 하겠다.

본 논문에서는 매 단계에서 수행되는 노드 탐색을 위해 임계값을 설정하여 이보다 작은 노드가 발견되면 즉시 탐색을 종료하도록 하여 복잡도를 개선하는 전략을 제안한다. 이를 위해서는 정확한 임계값 설정이 요구되는데, 임계값은 전체노드의 수와 샘플링 집합의 크기, 그리고 비용함수와 연관이 있다는 것에 주목할 필요가 있다. 또한, 매 단계가 진행됨에 따라, 최적의 노드가 갖는 비용함수 값은 증가하는 경향을 보이게 되며, 이는 다음 단계에서 최적의 노드를 선택하기 위한 임계값 설정에 영향을 주게 된다. 그림 1은 기존의 샘플링 알고리즘[9]을 수행하여 매 단계에서

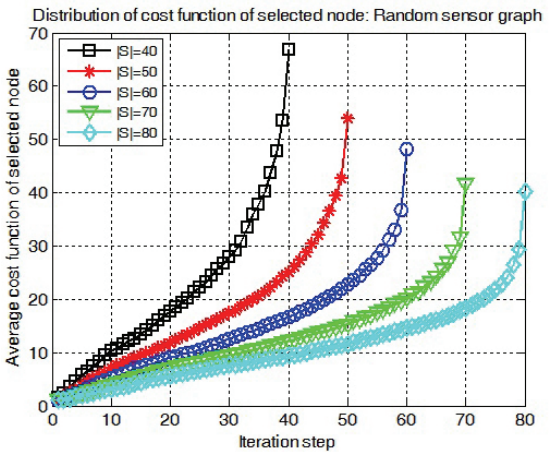

그림 1. 랜덤센서 그래프에서 샘플링 집합크기 변화에 따른 최적 노드의 비용함수 분포

Fig. 1 Distribution of cost function of the selected node by varying |S| for Random sensor graph

선택된 최적의 노드(selected node)가 갖는 비용 함수값을 보여준다. 즉, N=1,000개의 노드를 갖는 랜덤센서 그래프(Random Sensor Graph)를 30회 발생하여 각각의 발생된 그래프에 대해 r=|S|=40,...,80개의 노드를 선택하는 경우, 매 단계(가로축)에서 최적의 노드가

갖는 평균 비용 함수값(세로축)을 샘플링 집합의 크기를 변화하면서 나타내었다. 그림 1에서 알 수 있듯이, 단계가 진행됨에 따라 비용함수 값이 증가하고 샘플링 집합의 크기 |S|가 클수록 상대적으로 비용함수 값이 감소한다는 것을 확인할 수 있다. 그림 1은 본 논문에서 제안하는 임계값 적용을 위한 임계값 결정에 중요한 척도가 된다. 즉, i 단계에서 최적의 노드를 선택하기 위한 임계값 T_i를 이전 단계에서 선택된 노드의 비용함수 값 C_{i-1}과 샘플링 집합의 크기 |S|와의 관계식으로 식 (5), (6)과 같이 근사화할 수 있다:

$$T_i = C_{i-1} + \Delta C_i \quad \cdots (5)$$

$$\Delta C_i \propto \frac{C_{i-1}}{|S|^2} \quad \cdots (6)$$

여기서, 증가분 $\Delta C_i > 0$이며 임계값 설정에 중요한 설계 인자라 할 수 있다. 본 논문에서 제안하는 저복잡도 샘플링 알고리즘을 정리하면 다음과 같다.

단계 1: 샘플링 집합 $S_o = \varnothing$, $i = 1$로 초기화한다.

단계 2: 식 (7), (8)을 통해 최적의 노드를 구하고 샘플링 집합 S를 업데이트한다:

$$\text{if } tr\left(\left(\boldsymbol{U}_{S_i R}^T \boldsymbol{U}_{S_i R}\right)^+\right)_{\boldsymbol{u}_j, j \in S_i^C} \le T_i,\ j^* = j \quad \cdots (7)$$

$$S_i^* = S_{i-1}^* + \{j^*\},\ i = 1,...,|S| \quad \cdots (8)$$

단계 3: $i = i+1$하고, $S = S_r^*$이 될 때까지 단계 2, 3을 반복한다.

본 논문에서 제안하는 샘플링 기술의 복원성능은 임계값의 정확도와 밀접한 관계가 있으며, 또한 임계값이 클수록 복잡도를 상당 수준 개선할 수 있는 반면에, 복원 성능저하를 초래하게 된다. 본 논문에서는 고정 임계값을 적용하는 방식(수식 (5)에서 $\Delta C_i = 0$)과, 수식 (5)와 (6)을 사용하여 단계마다 임계값을 적응시키는 방식에 대해 성능평가를 진행하게 된다.

IV. 실험 및 분석

다양한 샘플링 기술의 성능을 평가하기 위해 다음 3개의 그래프를 발생하여 복원오차 및 복잡도(실행시간) 평가를 진행하였다:

1)랜덤센서 그래프(Random sensor graph, RSG)

2)랜덤정형 그래프(Random regular graph, RRG)

3)랜덤 Erdös-Rényi 그래프(Random Erdös-Rényi graph, RERG)

총 노드 수 N=1,000을 갖는 그래프를 각각 30회 발생하여 샘플링 기술의 평균 복원오차 및 실행시간을 통해 성능을 비교하였다. 랜덤정형 그래프 발생시 6개의 주변노드와 일정하게 연결되도록 하였고, 랜덤 Erdös-Rényi 그래프의 경우 0.05의 확률로 주변노드와 연결되도록 발생되었다. 그래프 $G(V,E)$ 및 그래프 라플라시안 $\boldsymbol{L}$(combinatorial Laplacian), 고유벡터 행렬 $\boldsymbol{U}$ 및 GFT $\boldsymbol{c}$는 GSP 툴박스를 사용하여 생성되었다[11]. 본 실험에서는 다음의 샘플링 기술들의 성능평가가 수행되었다:

1) 열 방향 가우시안 제거법(column-wise Gaussian elimination)에 기반한 효율적인 방식(Efficient sampling method, ESM[4])

2) 복원 오차를 비용함수로 하는 QR분해에 기반 방식(QR factorization-based method, QRM[9])

3) 고정식 임계값에 기반한 제안 알고리즘(Proposed method with fixed Th)

4) 적응식 임계값에 기반한 제안 알고리즘(Proposed method with adaptive Th)

샘플링 기술의 복원 성능평가를 위해 그래프 신호의 발생이 필요하며, 이는 식 (9)의 다 변수 가우시안 분포를 사용하여 발생되었다[12]:

$$p(\boldsymbol{f}) \propto \exp\left(-\boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}\right) = \exp\left(-\boldsymbol{f}^T (\boldsymbol{L} + \delta \boldsymbol{I}) \boldsymbol{f}\right) \quad \cdots (9)$$

여기서, δ는 공분산 행렬 $\boldsymbol{K} = (\boldsymbol{L} + \delta \boldsymbol{I})^{-1}$ 의 존재를 위해 작은 값(=0.01)을 취하게 된다. 또한 그래프 신호의 잡음 세기는 가산성 정규분포 $N(0,\sigma^2)$, $\sigma = 0.1$로 설정되었다.

본 실험에서는 다양한 그래프상에서 서로 다른 샘플링 기술을 사용하여 샘플링 집합 S를 구성하였고, 이를 통해 샘플링된 신호를 얻은 후 원 신호와의 평균 복원오차를 계산하였다. 그림 2, 3, 4에서 샘플링 집합의 크기 r=|S|를 40,...,80까지 변화시키면서 샘플링 기술의 복원 성능을 비교 평가하였다. 모든 그래프에서 적응식 임계값에 기반한 제안 알고리즘은 QRM

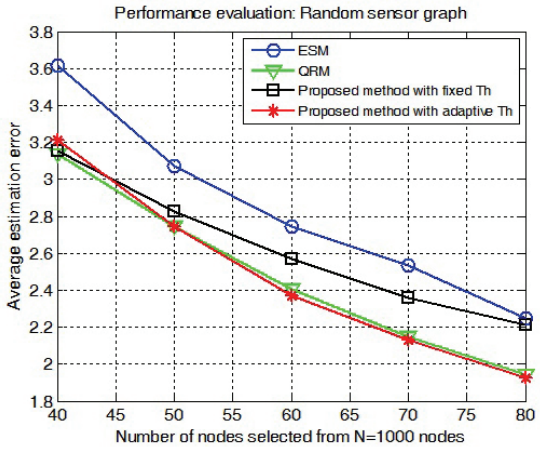

그림 2. RSG에서 샘플링 기술의 복원 성능평가
Fig. 2 Evaluation of reconstruction performance for RSG

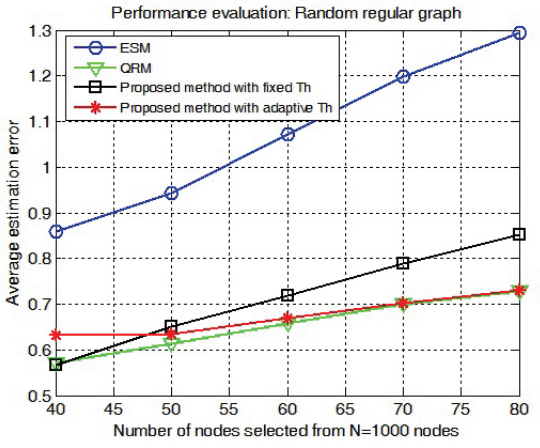

그림 3. RRG에서 샘플링 기술의 복원성능평가
Fig. 3 Evaluation of reconstruction performance for RRG

대비 복원성능 저하가 거의 발견되지 않았으며 고정식에 비해 우수한 복원성능이 보임을 알 수 있다. 또한 복잡도 평가를 위해 샘플링 기술의 실행시간을 비교하였으며, 그림 5는 랜덤센서 그래프에서의 샘플링 집합 크기에 따른 실행시간을 보여준다. 적응식과 고정식은 실행시간에 큰 차이가 없고, 제안 알고리즘의 샘플링 속도가 QRM 대비 향상된 성능이 보임을 알 수 있다. 실험에 사용된 모든 그래프에서 QRM 대비 적응식 제안 알고리즘이 복원성능저하 없이 평균 1.3배 이상 빠른 샘플링 속도가 보임을 확인하였다.

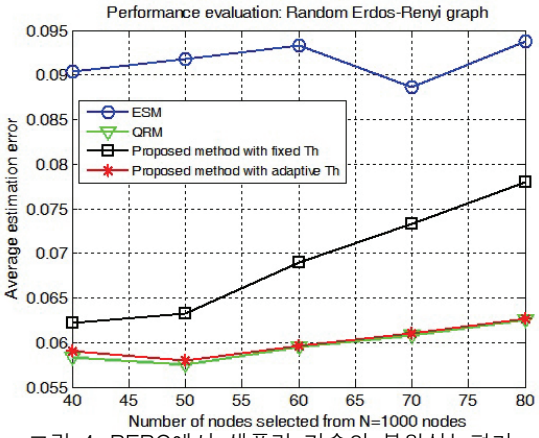

그림 4. RERG에서 샘플링 기술의 복원성능평가
Fig. 4 Evaluation of reconstruction performance for RERG

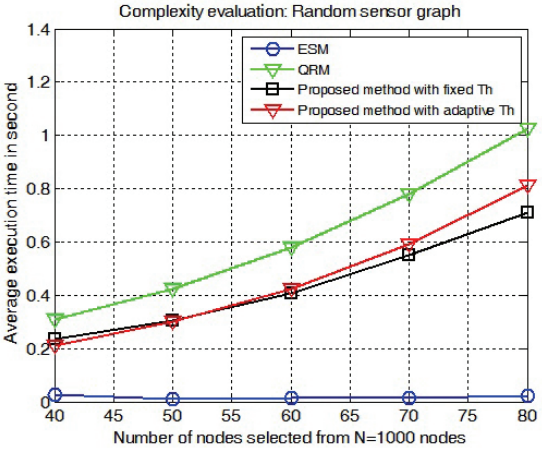

그림 5. RSG에서 샘플링 기술의 복잡도 평가
Fig. 5 Complexity evaluation for RSG

V. 결론

그래프 신호 샘플링을 위한 저 복잡도 기술에 대한 연구가 진행되었으며, 이를 위해 매 단계에서 최소의 비용함수 값을 갖는 노드 대신 임계값을 설정하여 임계값 보다 작은 비용함수 값을 갖는 노드를 선택하게 하여 복잡도를 개선하였다. 이 경우 발생할 수 있는 복원성능 저하를 막기 위해 정확한 임계값 설정이 요구되며, 이를 위해 적응식 임계값 설정방식에 대해 제안하였다. 다양한 그래프 상황에서 기존 샘플링 방법

대비 복원성능을 유지하면서 평균 1.3배 이상 빠른 실행시간이 보임을 확인하였다. 제안 샘플링 기술은 실용적 분야에 응용할 수 있으리라 판단되며, 향후 다양한 응용분야를 위한 저 복잡도 샘플링 기술에 대한 연구를 지속적으로 수행할 계획이다.

References

[1] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine,* vol. 30, no. 3, 2013, pp. 83-98.

[2] A. Ortega, P. Frossard, J. Kovaevic, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: overview, challenges and applications," *Proceedings of the IEEE*, vol. 106, no. 5, 2018, pp. 808-828.

[3] A. Anis, A. Gadde, and A. Ortega, "Towards a sampling theorem for signals on arbitrary graphs," *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3864-3858.

[4] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, 2016, pp. 3775-3789.

[5] S. Chen, R. Varma, A. Sandryhaila, and J. Kovaevic, "Discrete signal processing on graphs: sampling theory," *IEEE Transactions on Signal Processing*, vol. 63, no. 24, 2015, pp. 6510-6523.

[6] Y. Kim, "Efficient sampling of graph signals with reduced complexity," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 17, no. 2, Apr. 2022, pp. 367-374.

[7] A. Sakiyama, Y. Tanaka, T. Tanaka, and A. Ortega, "Eigendecompostion-free sampling set selection for graph signals," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, 2019, pp. 2679-2692.

[8] F. Wang, G. Cheung, and Y. Wang, "Low-complexity graph sampling with noise and signal reconstruction via Neumann series," *IEEE Transactions on Signal Processing*, vol. 67, no. 21, 2019, pp. 5511 - 5526.

[9] Y. Kim, "QR factorization-based sampling set selection for bandlimited graph signals," *Signal Processing*, vol. 179, 2021, pp. 1-10.

[10] Y. Kim, "Sampling set selection algorithm for weighted graph signals," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 17, no. 1, Feb. 2022, pp. 1023-1030.

[11] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, "GSPBOX: A toolbox for signal processing on graphs," *Information Theory,* 2014.

[12] N. D. Lawrence, "A unifying probabilistic perspective for spectral dimensionality reduction: insights and the new models," *Journal of Machine Learning Research,* vol. 13, no. 1, May 2012, pp. 1609-1638.

저자 소개

Yoon-Hak Kim

1992: BS degree in Electronic Engineering, Yonsei University
1994: MS degree in Electronic Engineering, Yonsei University
2007: Ph.D. degree in Electrical Engineering, University of Southern California.
2012 - Present: Professor, Chosun Univeristy