JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 2, pp. 31-39, February 2020
https://doi.org/10.9708/jksci.2020.25.02.031

BST-IGT Model: Synthetic Benchmark Generation Technique Maintaining Trend of Time Series Data

Kyung Min Kim*, Jong Wook Kwak*

*Student, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea
*Professor, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea

[Abstract]

In this paper, we introduce a technique for generating synthetic benchmarks based on time series data. Many of the data measured on IoT devices have a time series characteristic that measures numerical changes over time. However, there is a problem that it is difficult to model the data measured over a long period as generalized time series data. To solve this problem, this paper introduces the BST-IGT model. The BST-IGT model separates the entire data into sections that can be easily time-series modeled, collects the generated data into templates, and produces new synthetic benchmarks that share or modify characteristics based on them. As a result of making a new benchmark using the proposed modeling method, we could create a benchmark with multiple aspects by mixing the composite benchmark with the statistical features of the existing data and other benchmarks.

▸ **Key words**: time series data generation, IoTs, performance evaluation, benchmark, ARIMA

[요 약]

본 논문에서는 시계열 데이터를 기반으로 합성 벤치마크를 생성하는 기법을 소개한다. IoT 기기에서 측정되는 많은 데이터는 시간에 따른 수치 변화를 측정하는 시계열적 특성이 있다. 하지만 긴 기간 동안 측정되는 데이터를 일반화된 시계열 데이터로 모델링하기 힘든 문제점이 존재한다. 이런 문제를 개선하기 위해 본 논문에서는 BST-IGT 모델을 소개한다. BST-IGT 모델은 전체 데이터를 시계열 모델링이 쉬운 구간으로 분리하여 생성 데이터를 템플릿으로 수집하고 이를 기반으로 특성을 공유하거나 변형되는 새로운 합성 벤치마크를 생성한다. 제안된 모델링 기법을 이용하여 신규 벤치마크를 생성한 결과, 기존 데이터의 통계적 특성을 유지하는 합성 벤치마크와 다른 벤치마크와의 혼합으로 여러 특성을 가지는 벤치마크의 생성을 수행할 수 있었다.

▸ **주제어**: 시계열 데이터 생성, 사물인터넷, 성능 평가, 벤치마크, ARIMA

First Author: Kyung Min Kim, Corresponding Author: Jong Wook Kwak
*Kyung Min Kim (bl43a@ynu.ac.kr), Dept. of Computer Engineering, Yeungnam University
*Jong Wook Kwak (kwak@yu.ac.kr), Dept. of Computer Engineering, Yeungnam University
• Received: 2020. 01. 28, Revised: 2020. 02. 10, Accepted: 2020. 02. 11.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

각종 센서와 정보통신기술의 발달로 인해 사물인터넷(Internet of Things)의 성능 향상을 위한 연구 또한 급격하게 증가하고 있다. 이러한 연구의 성능 평가와 비교를 위해서 사물인터넷 센서에서 수집된 여러 데이터가 활용된다.

UC Irvine Machine Learning Repository와 같은 공용 데이터 세트 저장소는 기계 학습 연구를 위해 사물인터넷 센서에서 측정될 수 있는 여러 변량의 데이터 세트를 제공한다. 또한, 본 목적인 기계 학습 목적 이외에도 사물인터넷 기기의 성능 평가를 위해 데이터 세트를 참조하여 활용하는 사례가 있다. 그 외의 많은 연구는 연구자가 직접 구성한 하드웨어를 이용하여 수집된 데이터를 기반으로 성능 평가에 활용하는 경우가 많다 [1-4].

사물인터넷 센서로 측정된 데이터들은 기본적으로 시계열의 특성을 가질 것으로 기대되지만, 센싱 주기가 짧고 변화율이 높으므로 기존의 일반적인 시계열 모델로 해석하는 데 어려움이 따른다. 기존의 연구에서 이러한 데이터의 해석과 재생산을 위해 Markov-Chain 모델이나 Mixed Gaussian Model을 사용하여 특정 파라미터를 유추해내어 적합한 모델을 찾는 연구가 진행되었다. 그러나 Markov-Chain으로 해석한 모델은 선택해야 할 파라미터가 많고 최종적으로 유사한 데이터를 생산해낸 것을 변용하는 데 제약이 있다 [5-7].

본 논문에서는 이러한 필요성에 따라 주어진 경험적 데이터를 자동 회귀 적산 평균 모델(autoregressive integrated moving average, ARIMA)을 이용하여 시계열 모델링을 수행하고 합성 벤치마크를 생성하는 기법을 제안한다. 특히 가속도 센서, 자이로 센서와 같은 전체 경향에서 일반화된 시계열 모델을 생산해내기 힘든 문제에 대하여, 구간을 분리하여 벤치마크의 템플릿과 구간 모델을 생성하고 수집한다.

제안 시스템에서 파생되는 합성 벤치마크는 용도와 변형 정도에 따라 (1) 모방 벤치마크, (2) 준-모방 벤치마크, (3) 변형 벤치마크로 분류하여 유사한 경향을 보이는 모델의 검증이나 특정 부분값에서 이상치 혹은 변형이 일어난 사례에 대해서도 성능 평가를 할 수 있도록 구성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 시계열 모델을 추정하는 배경지식과 합성 벤치마크를 만들기 위한 기반 연구를 소개한다. 3장에서는 본 논문의 동기를 소개하고 4장에서 본 연구에서 제안하는 합성 벤치마크 생성 방법에 대해 구체적으로 기술한다. 5장에서는 제안하는 합성 벤치마크 생성 기법을 이용하여 생성된 벤치마크의 사례를 제시하고 생성 결과에 대해 논의한다. 마지막으로 6장에서는 결론에 관해 서술한다.

II. Background Work

1. Time Series Data Pattern

시계열(time series) 데이터는 관측치가 시간적 순서를 가진 데이터이다. 이 데이터는 변수 간의 상관성이 존재하는 데이터를 의미하며 독립 동일 분포(independent identically distribution, i.i.d)의 성질을 갖거나 불규칙한 데이터는 시계열 표본으로 선택하지 않는다. 시계열 데이터에서 나타나는 대표적인 특성들을 Fig. 1에 나타냈다.

추세(Trend)는 전체 구간에 대해서 센싱된 데이터가 장기적으로 증가하거나 혹은 감소하는 형태를 의미한다. 추세는 선형 혹은 비선형적일 수 있으며 아래의 Fig. 1의 (a)와 같은 형태가 추세를 나타내는 데이터의 대표적인 예시이다.

계절성(Seasonality)은 해, 일, 시간 등 특정한 시기에 의한 계절성 요인이 시계열에 영향을 줄 때 계절성 패턴이 나타난다. 계절성은 기본적으로 상수 형태로 나타나며 아래 Fig. 1의 (b)는 계절성 요소가 시계열 데이터에 영향을 끼치는 경우이다.

주기성(Cycle)은 고정된 빈도가 아닌 형태로 값이 주기에 따라 증가하거나 감소하는 모습을 보일때 주기가 나타난다고 표현한다. 특정한 요소의 순환 패턴과 관련이 있는 방향으로 나타나며, 계절적 패턴의 크기보다 변동성이 크

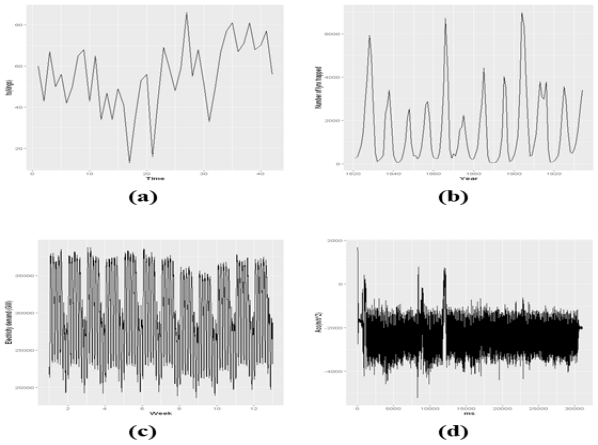

Fig. 1. Characteristics of Time Series Data

게 나타나는 특성이 존재한다. Fig. 1의 (c)는 주기성이 나타나는 시계열 데이터의 예시이다.

이와는 별도로, 시계열 데이터의 일반적인 정의에 부합하지만 특정한 추세나 계절성, 주기성이 보이지 않는 패턴이 존재한다. Fig. 1의 (d)에 나타난 시계열 데이터는 특정한 경향성을 보이지 않고 무작위적인 요동이 나타난다. 따라서 이러한 경우 기본적으로 예측 모델을 생성하는 데 어려움이 존재한다 [8-9].

2. Time Series Data Model Estimation

시계열 데이터의 모델링은 과거 특정 시간의 데이터를 통해 현재 및 미래의 추세를 예측하는 데 사용된다. 널리 활용되는 시계열 모델 방식에는 지수 평활(Exponential smoothing)을 기반으로 하는 ETS(Error Trend Seasonality) 모델과 자기 회귀(Autoregression)를 기반으로 하는 ARIMA 모델이 있다 [10].

지수 평활은 과거 관측값의 가중 평균을 적용하여 시계열 데이터의 적응 조정을 진행한다. 지수 평활에서 과거 관측값은 오래된 값일수록 지수적으로 가중치를 감소시켜 가장 최근 관측값이 높은 가중치를 가지도록 한다.

지수평활법은 일반적인 예측에서 우수한 성능을 나타내지만, 계절성이나 추세가 있는 시계열 데이터에 대해서 잘 예측하지 못한다. 따라서 추세의 특성 및 계절성의 정보 기준을 미리 조합하는 ETS 모델이 제안되었다 [8].

자기 회귀 모델은 과거의 패턴을 기반으로 전 시점의 시계열 데이터가 현 시점의 시계열 데이터에 대해 자기상관(autocorrelation)이 있음을 가정하여 시계열 추정을 수행한다. 따라서 차수 p의 AR 모델은 아래의 식 (1)을 이용하여 나타낼 수 있다.

$$\hat{y_t} = \phi_0 + \sum_{i=1}^{k} \phi_i y_{t-1} + \epsilon_t \tag{1}$$

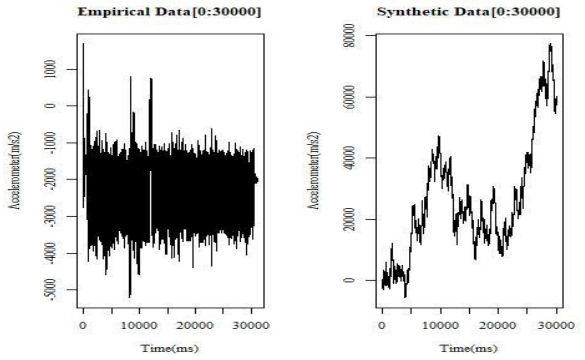

Fig. 2. ARIMA time series estimation for entire data

$$\hat{y_t} = \theta_0 + \sum_{i=1}^{k} \theta_i \epsilon_{t-1} + \epsilon_t \tag{2}$$

식 (1)에서 추정치 $\hat{y_t}$를 계산하기 위해서 기존 종속변수들이 합산되는 형태의 회귀 모델을 의미하고 ϵ_t는 오차를 의미한다. 여기에서, 오차값 ϵ_t를 식 (2)의 이동평균 모델을 이용하여 추론하면 ARMA 모델을 얻을 수 있다 [11].

ARMA 모델은 시간의 추이와 관계없이 평균 및 분산이 불변하거나 시점 간 공분산이 특정 시점에 귀속되지 않는 정상성(stationary)이 있는 시계열에 한해 적용할 수 있다. 따라서 원래의 시계열에서 연이은 관측값의 차이를 구하는 차분(difference)을 ARMA 모델에 합하여 비정상성에서 동작할 수 있는 식 (3)의 ARIMA(p, d, q) 모델을 얻을 수 있다.

$$\hat{y'_t} = \sum_{i=1}^{k} \phi_i y'_{t-1} + \sum_{i=1}^{k} \theta_i \epsilon_{t-1} + \epsilon_t \tag{3}$$

ARIMA(p, d, q) 모델은 3개의 파라미터로 이루어지는 회귀 모델로써, p는 자기회귀 부분의 차수, d는 1차 차분이 포함된 분량, q는 이동 평균의 지수로 구성된다. 적절한 p와 q값을 결정하기 위해 자기상관계수(autocorrelation function, ACF)와 편자기상관계수(partial autocorrelation function, PACF) 그래프를 이용하여 파라미터를 한정할 수 있다 [12-16].

III. Motivation

본 논문의 제안은 시계열 데이터의 합성 벤치마크 생성을 위한 데이터 모델링 과정에서 전체 센서 데이터에 대한일관적인 모델을 제시하기가 난해하다는 문제점에서 출발한다.

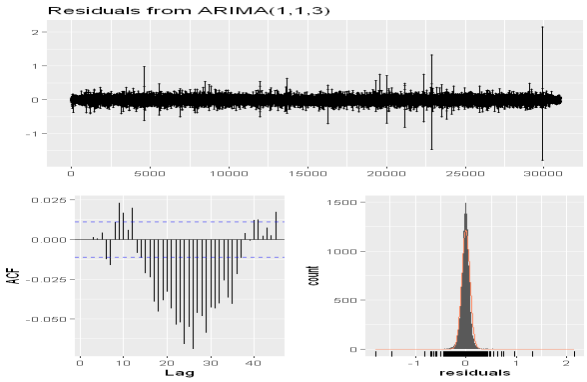

Fig. 3. residuals for the entire data

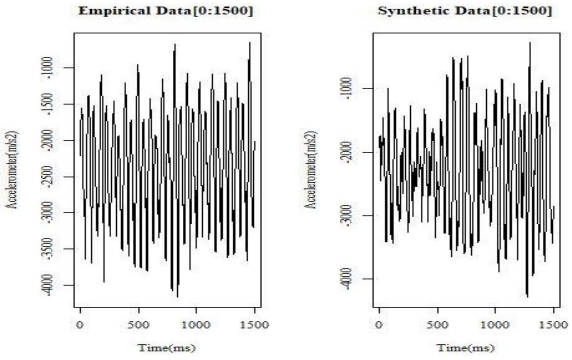

Fig. 4. ARIMA time series estimation for partial data

Fig. 2는 전체 데이터 세트를 그래프로 나타낸 것으로 우리의 목적은 이러한 데이터 경향을 보이는 유사 데이터 집합을 새로 생성하는 것이다. 가속도 센서는 시간의 변화에 따른 가속도의 변화 값을 나타내므로 일반적으로 시계열 데이터로 생각될 수 있다. 그러나 앞선 장에서 언급했듯이 이러한 형태의 데이터는 경향성, 계절성 및 주기성이 나타나지 않아 일반적인 시계열 모델링을 이용한 예측이 힘들며 Fig. 2의 우측 그래프와 같이 시계열 모형인 ARIMA 모형을 이용하여 벤치마크의 모델을 해석하고 재생산하려 하면 큰 분산과 불규칙한 경향으로 인해 시계열 모델링 결과 유사한 경향을 보이는 모델을 생성해내지 못한다. 결과 모델에 대한 추가 분석을 통해 자기상관계수가 적정 구간에서 절삭되지 않는 형태로 나타나 모형이 적절하게 적합 되지 않은 모습을 Fig. 3에서 확인할 수 있다.

반면, 측정 구간에서 일부 구간을 샘플링해낼 경우 다음의 Fig. 4의 좌측 그래프와 같이 부분적인 시계열 형태로 추출할 수 있다. 추출된 데이터를 활용하여 ARIMA를 이용해 시계열 모델링을 수행할 경우 원본 시계열 데이터의 경향성을 갖는 합성 데이터를 생성할 수 있다. Fig. 4의 우측 그래프는 주어진 모델을 이용하여 시계열 데이터 추정으로 생성을 한 결과이다. 또한, Fig. 5는 ARIMA를 이용하여 모델링 분석하여 재구축한 결과로 잔차가 정규분포를 따르고 자기상관계수가 적당히 구간 안에 포함되어 모델 적합이 잘 되었음을 확인할 수 있다.

이러한 관측 결과를 기반으로, 일반적인 시계열 모델을 이용하여 모델링하기 곤란한 실제 벤치마크를 부분으로 분해하여 시계열 모델링을 수행하고 이를 조합하여 새로운 합성 벤치마크를 생성하는 시스템을 제안한다.

IV. Methodology

1. BST-IGT Model Methodology

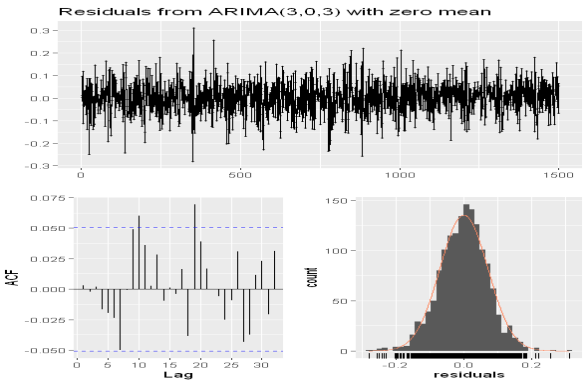

Fig. 5. residuals for the partial data

본 논문에서는 기존 벤치마크를 Benchmark Strucuture Template(이하 BST)과 Interval Generating Template(이하 IGT)로 분해하고 이를 재구축하여 기존 데이터의 경향성을 유지 혹은 새로운 합성 벤치마크를 생성하는 시스템을 제안한다. 생성 과정은 크게 구분하여 (1) 분해 (2) 재구축 (3) 생성모델 검증 과정으로 구성된다. 제안 기법을 통해 합성 벤치마크를 생성하기 위한 과정을 아래 Fig. 6에 나타내었다.

Fig 6. (a) 입력으로 기존의 경험적인 데이터가 삽입된다. 입력 벤치마크는 전체 벤치마크 혹은 벤치마크의 부분으로 구성될 수 있다.

수집된 데이터를 기반으로 BST (b-1)과 IGT (b-2)로 분해하는 Decomposition 과정을 수행한다. 이 과정에서 생성되는 BST과 IGM은 실측값이 아닌 적합 모델과 평균, 분산, 최대, 최소값의 기술통계량을 갖는 Generating Model의 형태이다.

Generating (c-1) 과정은 분해 과정에서 생성된 BST과 IGT를 합성하여 새로운 합성 벤치마크를 생성하는 과정이다. 부가적으로 필요에 따라 생성 모델에 대한 Manipulating (c-2)을 통해 기존의 BST 혹은 IGT의 일부 통계량을 조정하여 새로운 특성을 부여한다. 그리고 Mapping(c-3) 과정에서 BST와 IGT를 연결하여 새로운 합성 벤치마크를 생성한다.

검토 과정은 재구축된 모델의 특성을 정의하고 검증하는 과정으로 생성 모델의 유형에 따라 모방 합성 벤치마크(Imitation synthetic benchkmark, ISB) (d-1), 준-모방 합성 벤치마크(Semi-imitation synthetic benchmark, SSB) (d-2), 변형 합성 벤치마크(Variation synthetic benchmark, VSB) (d-3)로 구분하여 분류된다. 분류된 벤치마크의 특성은 실험 및 성능 평가의 용도에 따라 다르게 활용될 수 있다.

2. Decomposition

이 과정은 수집된 경험적 데이터를 라벨링된 벤치마크 단위로 분해하여 라벨링된 벤치마크를 생성하기 위한 BST와 부분 구성 요소를 생성하는 IGT를 수집하는 과정이다. 한 벤치마크는 하나의 BST 구조를 가지며 여러 IGT 데이터를 포함할 수 있다. IGT와 BST는 공통적으로 Table 1과 같은 데이터로 기술된다.

Decomposition을 통해 생성되는 BST와 IGT는 실제 데이터값이 아닌 데이터를 표현하기 위한 템플릿 모델이다. 데이터로는 구간 내 데이터의 빈도, 검증을 위한 유형

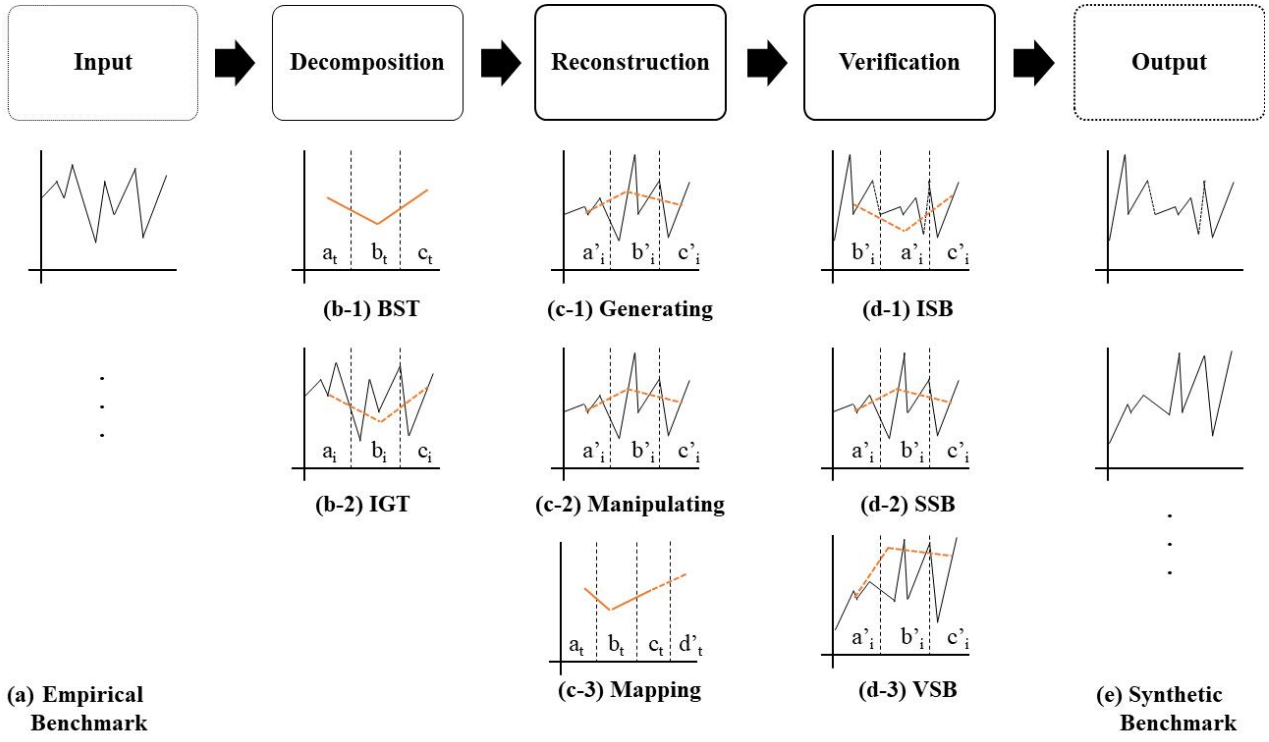

Fig. 6. Summary of the Synthetic Benchmark Generation Process

분류와 데이터 생성을 위한 기반 모델 및 기본적인 기술통계량인 최대값, 최소값, 평균 및 표준편차로 구성되어 있다. IGT는 개별 실제 데이터가 갖는 분포를 생성하기 위한 모델이며 추정 구간의 축소가 이루어져 피팅에 영향을 끼치는 계절성이 없어지는 부수 효과가 있다. BST는 전체 벤치마크의 특성과 IGT의 위치를 정의하기 위한 모델이다.

Table 1. BST and IGT Structure

Variable	Type	Description
freq (opt)	Integer	The number of data in the interval
type	Integer	Type 1 / Type 2 / Null
model	string	ARIMA(p, q, d) / distribution type
max	float	Maximum value in the interval
min	float	Minimum value that appears in the interval
mean	float	Arithmetic mean of the interval
sd	float	Standard deviation of the interval
tag (opt)	string	Supplementary Indicators for Evaluation

3. Reconstruction

재구축 과정은 분해 과정에서 수집된 BST과 IGT를 기반으로 새로운 합성 벤치마크를 생성한다. 이 과정은 세분화하여 (1) Generating과 (2) Manipulating 그리고 (3) Mapping 과정으로 나눌 수 있다.

Generating: BST, 및 IGT에서 기술되어있는 ARIMA 모델 및 분포 모델과 주어진 기본 기술통계량으로 합성 데이터를 생성하는 과정이다. 이 과정에서 사용되는 IGT간의 시계열 특성은 고려되지 않으며 랜덤 시드(Random seed)를 통해 각 생성은 고유한 데이터 모형을 구축한다.

Manipulating: 필요에 따라 IGT의 특성을 세부적으로 조정한다. 평행이동, 단순곱을 통해 개별 IGT의 생성 모형을 변형하여 원본 모델에 변화를 부여한다. 이 과정은 생성하는 벤치마크의 분류에 따라 호출되지 않을 수 있으며 별도로 Mapping 중에도 값의 비율 조정을 위해 호출될 수 있다.

Mapping: BST의 특성을 조정하고 합성 벤치마크를 생성한다. BST를 대상으로 시계열 모델링 추정을 통해 합성 벤치마크의 구간 길이를 조정하거나 한 개 이상의 다른 실제 벤치마크와의 합성을 이 과정에서 수행한다. Mapping 과정까지 완료될 경우 한 개의 합성 벤치마크가 생성되며, 다음 과정인 Verification에서 생성된 벤치마크의 특성을 분류한다.

4. Verification

재구축된 벤치마크는 BST과 IGT의 특성에 따라 분류하기 위하여 조건을 부여한다. 기존의 데이터에서 수집된 BST와 IGT는 추정된 모델에 따라 다음의 두 가지 조건 중 하나를 만족한다.

Type I: 합성 모델이 ARIMA 모델로 추정될 경우 충족한다. 이 경우 ARIMA 모델에서 추정된 조건에 의하여 같은 상관 계수를 기반으로 생성된 데이터임이 증명되므로, 원본 모델 m과 추정 모델 m'는 통계적으로 동일한 모형임이 만족된다. 또한, 기존 모델 m이 가지는 시계열 데이터의 특성이 유지된다.

Type II: 합성 모델이 ARIMA 모델로 추정되지는 않았지만, 유사 분포를 특정할 수 있을 때 충족된다. 원본 모델 m이 갖는 통계적 특성을 특정할 수 있어서 생성 모델 m'는 통계적으로 동일한 모형으로 생성될 수 있지만, 원본 데이터가 가지는 시계열 경향은 이 경우 무시된다.

BST와 IGT가 충족하는 조건의 엄밀성에 따라 생성되는 합성 벤치마크의 유형을 정의한다. 조건의 충족 유형에 따라 상위 충족 요건은 하위 충족 요건을 부분집합으로 가져

Table 2. Type Conditions for Synthetic Benchmark Validation

Model	BST(e)	IGT(e)
Imitation	e＝Type I	e＝Type I
Semi-Imitation	For all e ≤ Type I, Exist e Type II,	
Variation	e≤Type II	e≤Type II

특정 요소에 대해 조건이 엄격할수록 생성 벤치마크는 제안한 기법과 같은 모델로 생성된 벤치마크임이 보장된다. 이는 Table 2에 나타나는 것과 같이 유형화할 수 있다.

모방 합성 벤치마크는 BST와 IGT의 모든 요소가 Type I 으로 구성되었을 경우 이를 ISB라 지칭한다. 각 개별 요소가 전부 원본 모델 m과 통계적으로 유의미하게 생성된 모델이며 전체 벤치마크의 구성 또한 원본 모델과 같은 상관 계수를 가지고 생성되었다. 따라서 해당 조건으로 판정된 벤치마크는 특정 요소에 대한 성능 평가에서 동일하거나 유사한 결과를 도출할 수 있다.

준 모방 합성 벤치마크는 BST 혹은 IGT가 모두 Type II 이상이면서 둘 중 하나가 Type I에 해당할 경우이다. BST가 Type II에 해당할 경우 해당 벤치마크의 개별 구성 요소는 시계열성과 유사성을 만족하나 전체 합성 벤치마크 구성의 일부분에서 시계열 특성은 유지되지 않는 경우를 의미한다. 연장 및 축소된 BST 모델이 이에 해당할 수 있다. 반대로 BST는 Type I을 만족하지만, IGT는 Type II 일 경우는 전체 합성 벤치마크 구조는 원본 모델과 같은 시계열 및 통계적 특성을 유지하지만 부분 요소에서 일치하는 구간 모델을 정의할 수 없어 시계열 특성이 무시된 임의 분포로 대체된 모형을 의미한다. 이는 해당 구간의 기술 통계량을 갖는 IGT가 존재하지 않을 때 미리 정의된 분포에서 결측값을 근사화하는 경우 발생할 수 있다. 이와 같은 합성 벤치마크는 SSB라 지칭하며 성능 평가에서 원본 모델과 같은 수준의 결과가 측정되지는 않지만 비교 지표간의 추세는 유지되리라 기대할 수 있다.

변형 합성 벤치마크는 재구축된 모델의 BST와 IGT가 모두 기존 모델의 시계열 및 통계적 분포를 따르지 않고 생성된 데이터 세트를 의미한다. 재구축 과정에서 BST와 IGT 모두에 Manipulating이 가해지거나, Mapping 과정에서 단일한 벤치마크가 아닌 2개 이상의 합성 벤치마크를 구성하도록 Mapping 되었을 경우 생성된 벤치마크는 VSB로 분류된다. 기존의 데이터 유형에서 완전히 다른 새로운 유형의 데이터를 생성해내는 과정이므로 성능 평가에서는 원본 모델과 전혀 다른 결과값을 나타낼 수 있다.

V. Experimental Result

1. Overview

본 장에서는 제안한 합성 벤치마크 생성 모델을 이용하여 실제 임의의 합성 벤치마크를 생성하고 그 과정에서 생성되는 데이터와 결과에 대하여 평가한다. 학습 모델은 UCI에서 제공되는 경험적 벤치마크를 사용하여 이와 유사한 성질을 갖는 벤치마크의 생성을 확인한다. 단, 같은 시계열 모델을 따르는 것은 ARIMA 모델을 공유하는 것 이외에 확실한 비교가 힘든 문제가 있어 시각화된 데이터를 제안하기 위하여 전체 데이터에서 연속된 일부 구간을 추출하여 활용한다.

2. Synthetic Benchmark Generation

우선 개별 구성 요소인 IGT 모델의 생성에 대하여 먼저 검증을 수행한다. 검증을 위해 각각의 Empirical Data들에 대한 IGT 모델을 추정하고 그리고 이를 기반으로 데이터를 생성해냈다. 그 결과는 Fig. 7과 Fig. 8에 각각 나타내었다.

Fig. 7은 Empirical Benchmark 1에 대한 Decomposition으로 IGT 모델 a와 b, 그리고 c 모두 ARIMA 모델을 이용하여 유사 시계열 모델을 생성한 Type I 모델에 해당한다. 전체적으로 경험 데이터와 합성 데이터가 유사한 평균과 표준편차를 가지고 있으며 유사한 시계열 특성을 나타내는 것이 확인된다.

Fig. 8은 Empirical Benchmark 2를 사용하여 모델링을 수행한 결과이다. IGT a와 b의 경우 ARIMA 모델을 활용하여 시계열 데이터 생성을 하였으나, c의 경우 ARIMA 모델로 시계열 해석 결과 적합한 그래프를 찾지 못하였다. 따라서해

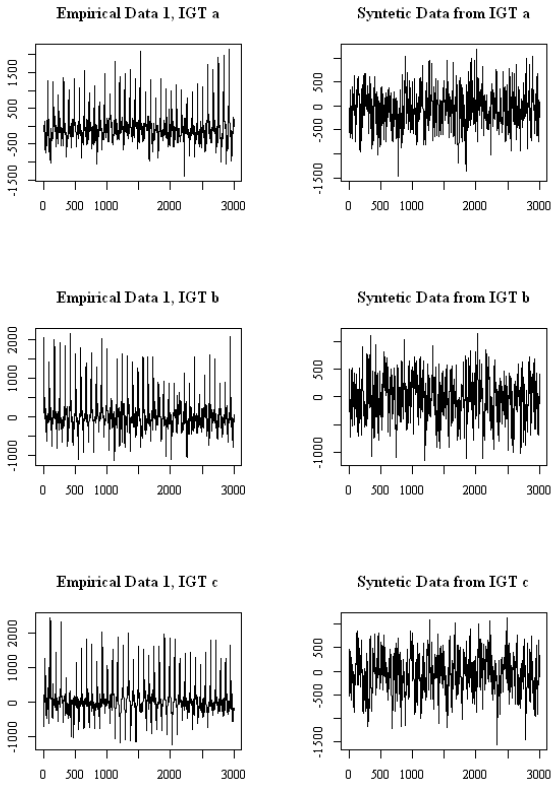

Fig. 7. IGT Modeling of Empirical Benchmark 1

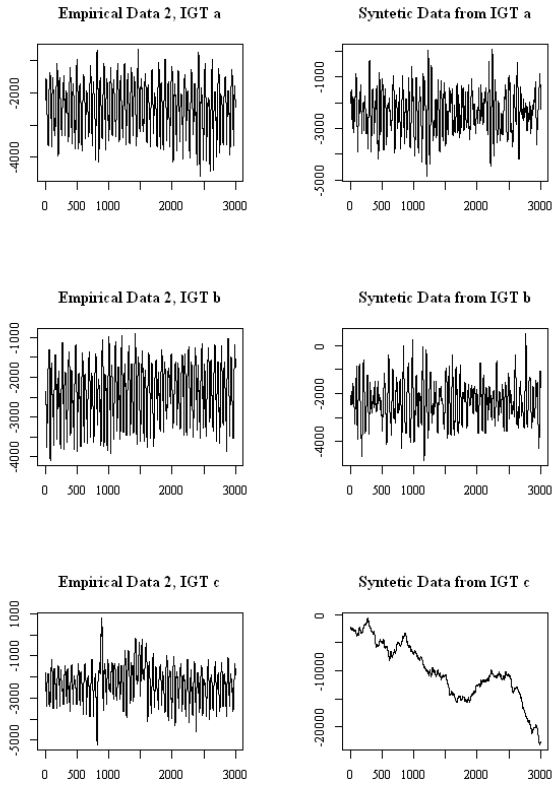

Fig. 8. IGT Modeling of Empirical Benchmark 2

당 IGT는 Type II 모델로 정의되며 Fig. 9와 같이 유사 분포를 찾아서 임의 분포 배열을 생성한다. 생성 모델과 원본 데이터에 대한 Kolmogrov-Smirnov 검정 결과, p-value 0.12로 두 난수가 같은 분포라는 귀무가설을 기각할 수 없다. 따라서 원본 시계열 데이터의 경향이 유지되지 않더라도 통계적으로는 유사한 분포가 생성되었음이 확인된다.

다음은 입력된 벤치마크로 생성한 합성 데이터들을 유형별로 나열하고 원본 데이터와의 시각적 차이를 검토한다. 생성 데이터의 모형을 Fig. 10, 그리고 데이터 분포의 밀도 그래프를 Fig. 11에 나타내었다.

Fig. 10 의 (a) Empirical Benchmark 1은 모든 선택 구간에서 ARIMA를 이용한 시계열 모델링이 수행된다. 이와 같은 경우 전체 IGT는 Type I형으로 분류되어 이미 수집된 ARIMA 모델을 이용하여 추정되었고, BST에는 어떠한 변형도 없는 모방 합성 벤치마크로 확인할 수 있다.

(b)는 Empirical Benchmark 1을 기반으로 주어진 IGT와 BST에 어떠한 조정도 가하지 않고 합성한 결과 원본 모형과 2% 이내의 차이를 가지는 기술 통계량을 갖는 ISB가 생성되었다.

Fig. 10의 (c)에서 나타나는 Empirical Benchmark 2는 두 구간에서 ARIMA를 이용한 모델 적합이 이루어졌으나, 한 구간에서 시계열 모델로 해석할 수 없는 형태의

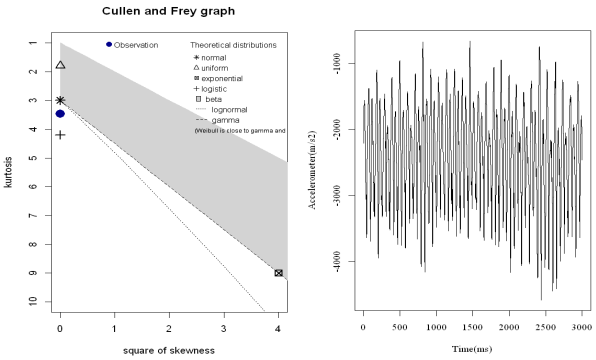

Fig. 9. Generate random data with similar distribution

데이터가 나타났다. 그로 인해서 Fig. 10의 (d)와 같은, 마지막 구간을 유사 분포로 대체하여 전체적인 기술 통계량과 분포는 유사하지만 원본 시계열 데이터 경향은 다소 손실된다. 이는 SSB로 분류될 수 있다.

Fig. 10의 (e)와 (f)는 중앙값을 재정렬하여 조정을 거친 BST를 기반으로 IGT를 분포에 따라 확장하고 Empirical Data 1과 Empirical Data 2의 속성을 섞은 변형 합성 벤치마크를 나타낸 것이다.

3. Discussion

ARIMA를 이용한 부분적인 시계열 추정은 통계적으로 유의미하게 동일하면서 경향성을 어느 정도 모방하는 합

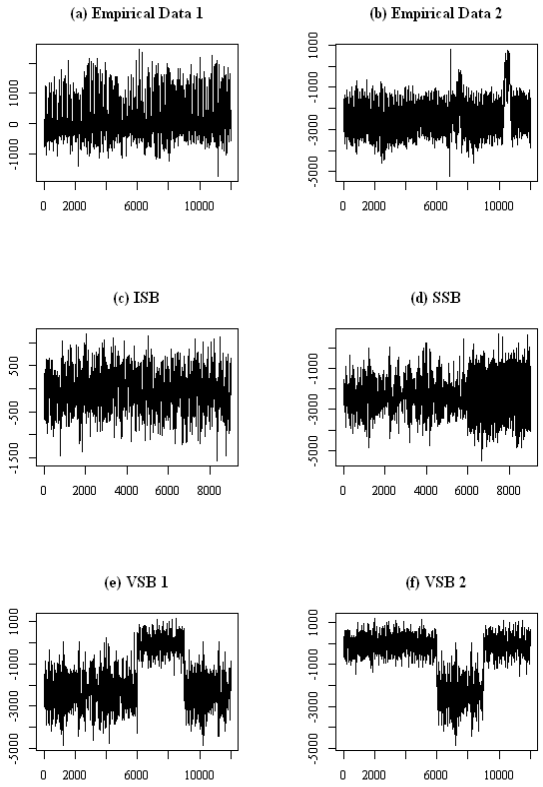

Fig. 10. Generated synthetic benchmark model

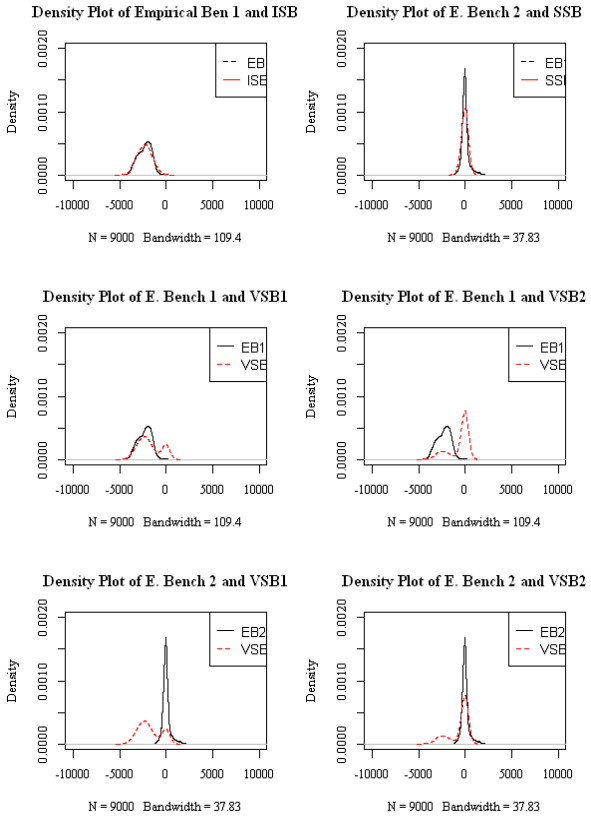

Fig. 11. Comparison of Density Between Benchmarks

성 데이터를 생성해낼 수 있지만 Fig. 7의 Empirical Data와 Synthetic Data의 비교에서 나타나듯 완전한 특성을 반영하는 것은 다소 어려움이 따르는 것으로 확인된다. 그러나 최종적인 합성 벤치마크의 비교를 나타낸 Fig. 11을 참조하면, ISB는 원본 벤치마크와 유사한 수치 분포을 가지며 SSB의 경우 밀도의 수치는 차이가 있으나 분포 모형은 동일한 형태로 작성된다. VSB는 각각 원형이 되는 벤치마크와 비교하였을 때 원본 데이터의 분포 성질을 일부 유지하는 것으로 확인된다.

제안 기법을 통한 벤치마크 생성으로 이후 성능 평가에서 합성 벤치마크에 기대할 수 있는 유사한 데이터나 합성 데이터의 생성은 기존의 Markov-chain 모델 등 장기간의 데이터 수집과 학습이 필요한 모델에 비해 간편하게 데이터를 확장할 수 있는 장점이 있어 여러 성능 평가의 참고 지표로써 활용이 기대된다.

VI. Conclusions

본 논문에서는 시계열 모델링을 통하여 합성 벤치마크를 만들기 위한 기법을 소개하였다. BST-IGT 모델은 전체 벤치마크에서 부분적으로 나타나는 시계열을 추정 모델을 기반으로 하여 이를 생성 가능한 정도의 단위로 분리하여 수집하고 합성 벤치마크를 생성한다. 또한 ARIMA 모델로 시계열 피팅이 정상적으로 이루어지지 않을 경우 유사 분포를 활용하여 생성되는 합성 벤치마크의 유형과 엄밀성을 정의하였다. 제안된 기법을 활용하여 신규 벤치마크를 생성하였을 때 통계적으로 유사한 속성을 갖는 합성 벤치마크가 생성되는 것을 확인하였으며 동시에 시각적으로도 유사한 경향을 가지는 것을 확인하였다.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2017R1D1A1A09000654).

REFERENCES

[1] Verma, S., Kawamoto, Y., Fadlullah, Z. M., Nishiyama, H., & Kato, N., "A survey on network methodologies for real-time

analytics of massive IoT data and open research issues", IEEE Communications Surveys & Tutorials. 19(3), pp. 1457-1477, 2017. DOI: 10.1109/COMST.2017.2694469

[2] Borgomeo, E., Farmer, C. L., & Hall, J. W., "Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments", Water Resources Research. 51(7), pp. 5382-5405, 2015. DOI: 10.1109/COMST.2017.2694469

[3] Arlitt, M., Marwah, M., Bellala, G., Shah, A., Healey, J., & Vandiver, B., "Iotabench: an internet of things analytics benchmark", Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering, pp. 133-144, January 2015. DOI: 10.1145/2668930.2688055

[4] Dua, D. and Graff, C.,"UCI Machine Learning Repository", [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2019.

[5] Aljawarneh, S., Radhakrishna, V., Kumar, P. V., & Janaki, V., "A similarity measure for temporal pattern discovery in time series data generated by IoT", 2016 International conference on engineering & MIS (ICEMIS), pp. 1-4. September 2016. 10.1109/ICEMIS.2016.7745355

[6] Xu, X., Huang, S., Chen, Y., Browny, K., Halilovicy, I., & Lu, W., "TSAaaS: Time series analytics as a service on IoT", 2014 IEEE International Conference on Web Services, pp. 249-256. June 2014. DOI: 10.1109/ICWS.2014.45

[7] Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W., "A review on time series forecasting techniques for building energy consumption", Renewable and Sustainable Energy Reviews. 74, pp. 902-924, 2017. DOI: 10.1016/j.rser.2017.02.085

[8] De Livera, A. M., Hyndman, R. J., & Snyder, R. D., "Forecasting time series with complex seasonal patterns using exponential smoothing", J American Statistical Association. 106(496), pp. 1513-1527, 2011. DOI: 10.1198/jasa.2011.tm09771

[9] Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D., "Forecasting with exponential smoothing: the state space approach", Springer Science & Business Media, 2008. DOI: 10.1198/jasa.2011.tm09771

[10] Jain, Garima, and Bhawna Mallick, "A study of time series models ARIMA and ETS.", Available at SSRN 2898968, 2017.

[11] Choi, ByoungSeon, "ARMA model identification", Springer Science & Business Media, 2012.

[12] Fan, S., & Hyndman, R. J., "Short-term load forecasting based on a semi-parametric additive model",IEEE Transactions on Power Systems. 27(1), pp. 134-141, August 2011. DOI: 10.1109/TPWRS.2011.2162082

[13] Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J., "ARIMA models to predict next-day electricity prices", IEEE transactions on power systems. 18(3), pp. 1014-1020, August 2003. DOI: 10.1109/TPWRS.2002.804943

[14] Singh, S. N., and Abheejeet Mohapatra, "Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting", Renewable energy. 136, pp. 758-768, 2019. DOI: 10.1016/j.renene.2019.01.031

[15] Farhath, Z. A., Arputhamary,`` B., & Arockiam, L., "A Survey on ARIMA Forecasting Using Time Series Model", Int. J. Comput. Sci. Mobile Comput. 5, pp. 104-109, August 2016. DOI: 10.3390/sym11020240

[16] Drago, Carlo, and Elisabetta Massa, "Measuring and Forecasting Financial Advisory Demand using a Hybrid ETS-ANN Model", BORDERS WITHOUT BORDERS:: Systemic frameworks and their applications, 2019.

Authors

Kyung Min Kim received a B.S. degree in Department of Computer Engineering from Yeungnam University, Korea in 2017, a M.S. degree in Department of Computer Engineering from Yeungnam University, Korea, in 2019, respectively. He is currently a Ph.D. candidate in Department of Computer Engineering at Yeungnam University. His current research interests include advanced processor architecture, internet of things and non-volatile memory.

Jong Wook Kwak received a B.S. degree in Computer Engineering from Kyungpook National University, Daegu, Korea in 1998, a M.S. degree in Computer Engineering from Seoul National University, Seoul, Korea in 2001, and a Ph.D. degree in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea in 2006. From 2006 to 2007, he worked as a Senior Engineer in the SoC R&D Center, at Samsung Electronics Co., Ltd. During 2011~2012, he was a Guest Researcher at the Research Institute of Advanced Computer Technology, Seoul National University. During 2012~2013, he was a Visiting Scholar at the Georgia Institute of Technology, Atlanta, GA, USA. As a Head Director, he led DREAM Software Human Resource Training Center from 2014~2015. During 2018~2019, he was a Visiting Scholar at Arizona State University, Tempe, AZ, USA. He is currently a professor in the Department of Computer Engineering, Yeungnam University. His research interests include advanced processor architecture, low-power mobile embedded systems, and high performance parallel computing.