Graph based Tabular Financial Data Synthesis

2024년도 하계 세미나 - August 9th, Friday

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
Kwanggeun Kim

Table of Contents

Background

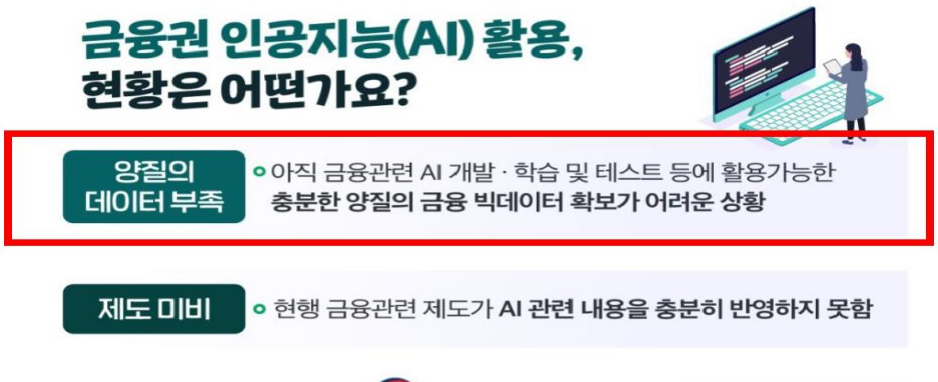
- 양질의 **금융 빅데이터 확보가 어려운 상황**임
  - 금융데이터는 개인정보와 민감정보가 포함되어 있어 **연구에 활용할 수 있는 실제 데이터가 많지 않고**
  - 나아가, 사기거래, 이상거래 등의 사건은 **발생 빈도가 낮아 모델이 학습하고 평가할 수 있는 데이터가 부족함**

→ 이에 대한 해결방안으로 **합성데이터가 주목**받고 있으며, 금융권의 AI 경쟁력 확보를 위해 **합성데이터의 연구는 반드시 선행될 필요**

매일경제 · 2017.03.22. · 네이버뉴스

한국거래소, AI 시장감시시스템 내년 4월 구축

한국거래소가 이 같은 미래 사회 청사진을 바탕으로 인공지능(**AI**)이 탑재된 차세대 **시장감시 시스템**으로 불공정 거래를 사전에 예방해 자본**시장**의 질적... **AI**와 빅데이터 **도입**으로 향후 불공정거래 찾아내기의 정확도가 ...



[참고] 금융위원회 공식블로그('22.8)

Background

- 금융 합성데이터 연구 및 활용에 대한 **국내·외 상황은?**
  - 독일 보험회사 Provinzial가 **보험상품 추천 AI의 학습을 위한 합성데이터 활용** 프로젝트를 수행하였고, 합성데이터로 학습시킨 AI가 원본데이터로 학습시킨 AI **대비 97% 정도 성능**을 보임
  - (2022 NeurIPS) Turning the Tables - Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation
  - 국내의 경우, **개인정보 유출 리스크**로 인한 가명정보 수준 데이터는 공개를 꺼려하고 있어, **해당 연구 및 활용 수준이 해외 대비 미진한 상황**

2022 10 Breakthrough Technologies

MIT 테크놀로지 리뷰 선정 '2022년 10대 기술'

1. 경구용 코로나19 치료제 (A pill for covid)
2. 실용적인 핵융합로 (fusion reactor)
3. 비밀번호를 대체하는 새로운 인증기술 (The End of Passwords)
4. 단백질 구조 예측용 AI (AI for Protein Folding)
5. 지분증명 (Proof of Stake)
6. 오래 지속되는 그리드 배터리 (Long-lasting Grid Batteries)
7. 인공지능을 위한 합성 데이터 (Synthetic Data for AI)

Synthetic Data Generation Model

- **통계 기반**
  - 원본 데이터의 통계적 특성을 이용하여 가정한 모집단 분포 추정 후 표본 추출하는 방식
  - E.g., Coupla 모형
- **머신러닝 기반**
  - 독립변수 집합을 지정해 종속변수 예측하는 모델을 이용한 방식
  - E.g., CART, SMOTE
- **생성AI 기반**
  - 딥러닝 생성형 AI 모델을 이용하여 합성데이터를 생성하는 방식
  - 테스트, 이미지 등 비정형 데이터 생성에 용이
  - 원하는 양의 합성데이터 생성에 용이
  - E.g., GAN

- **Modeling Tabular Data using Conditional GAN [CTGAN]**
  - **NeurIPS 2019**

CTGAN(Conditional Tabular GAN) 모델

- **Tabular Data의 예**

종목정보	시장	시가	고가	저가	종가	...
삼성전자	코스피	74,900	75,300	72,300	72,500	…
카카오	코스피	36,750	39,000	36,750	38,000	…
에코프로	코스닥	84,000	92,400	83,600	91,500	…

- **Tabular Data를 어떻게 GAN모델에 주입할 것인가?**
  - Mode-specific Normalization

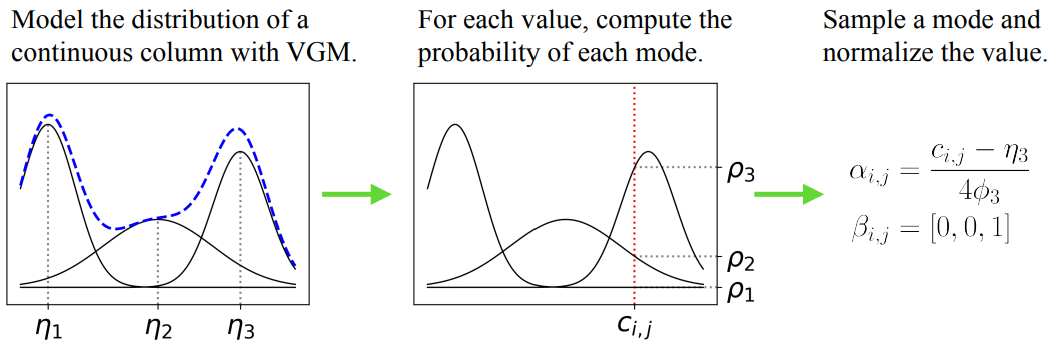


Figure 1: An example of mode-specific normalization.

The representation of a row become the concatenation of continuous and discrete columns

$$\mathbf{r}_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \ldots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus \mathbf{d}_{1,j} \oplus \ldots \oplus \mathbf{d}_{N_d,j},$$

where $\mathbf{d}_{i,j}$ is one-hot representation of a discrete value.

CTGAN(Conditional Tabular GAN) 모델

- **Conditional Generator and Training-by-Sampling**

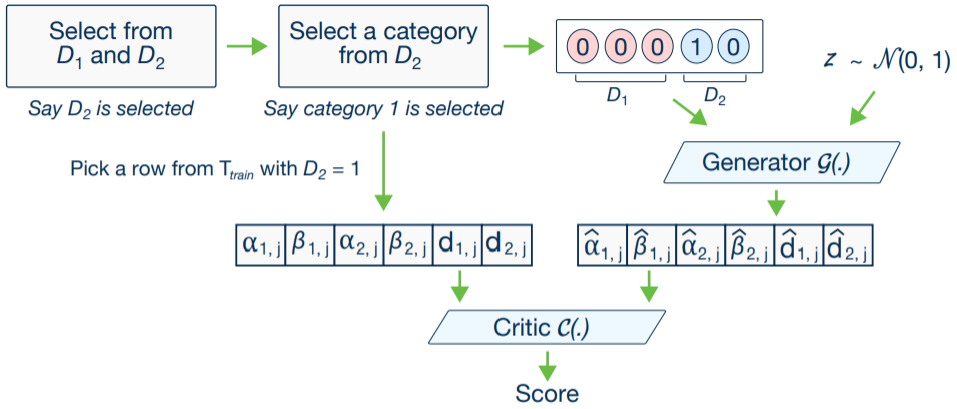


Figure 2: `CTGAN` model. The conditional generator can generate synthetic rows conditioned on one of the discrete columns. With training-by-sampling, the *cond* and training data are sampled according to the log-frequency of each category, thus `CTGAN` can evenly explore all possible discrete values.

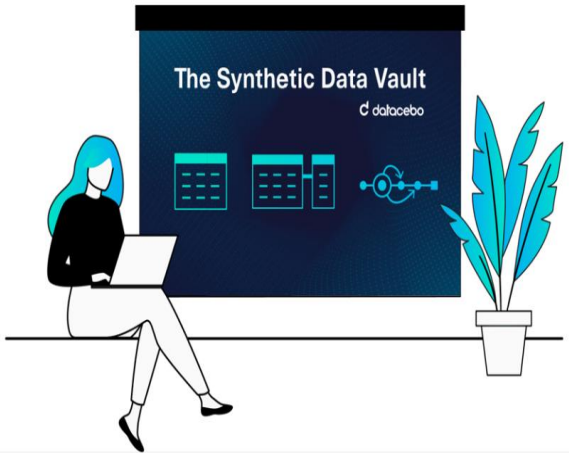
1. Create N_d zero-filled mask vectors $\mathbf{m}_i = [\mathbf{m}_i^{(k)}]_{k=1...|D_i|}$, for $i = 1, \ldots, N_d$, so the ith mask vector corresponds to the ith column, and each component is associated to the category of that column.
2. Randomly select a discrete column D_i out of all the N_d discrete columns, with equal probability. Let i^* be the index of the column selected. For instance, in Figure 2, the selected column was D_2, so $i^* = 2$.
3. Construct a PMF across the range of values of the column selected in 2, D_{i^*}, such that the probability mass of each value is the logarithm of its frequency in that column.
4. Let k^* be a randomly selected value according to the PMF above. For instance, in Figure 2, the range D_2 has two values and the first one was selected, so $k^* = 1$.
5. Set the k^*th component of the i^*th mask to one, i.e. $\mathbf{m}_{i^*}^{(k^*)} = 1$.
6. Calculate the vector $cond = \mathbf{m}_1 \oplus \cdots \mathbf{m}_{i^*} \oplus \mathbf{m}_{N_d}$. For instance, in Figure 2, we have the masks $\mathbf{m}_1 = [0, 0, 0]$ and $\mathbf{m}_{2*} = [1, 0]$, so $cond = [0, 0, 0, 1, 0]$.

합성데이터 관련 오픈소스들

명칭	공개 일자	합성 데이터 유형	특징
Gretel Synthetics	2020.3.	정형 데이터 (시계열)	• Tensorflow와 Pytorch 버전 모두 존재 • 우분투 OS에서만 동작 • 생성 기술로 DGAN 채택
SDV	2018.9.	정형 데이터 (테이블)	• JSON, DataFrame 등 다양한 데이터포맷 처리 지원 • 결측치 처리 등 데이터 품질 개선 기능 탑재 • 코플라(Copula) 통계모형, CTGAN 등 다양한 생성 기술을 채택하여 사용자의 선택을 지원
SmartNoise	2021.11.	정형 데이터 (테이블)	• Pytorch를 기반으로 구현 • 머신러닝 대표 라이브러리인 SKlearn API 구조로 설계 • CTGAN, PATE-GAN, PATE-CTGAN 등 다양한 GAN 확장 기술을 채택하여 사용자의 선택을 지원
TabGAN	2021.2.	정형 데이터 (테이블)	• 머신러닝 대표 라이브러리인 SKlearn API 구조로 설계 • 생성 기술로 CTGAN 채택 • 손쉽게 합성 학습용 데이터를 생성 및 검증할 수 있는 파이프라인 제공

Welcome to the SDV!

The **Synthetic Data Vault** (SDV) is a Python library designed to be your one-stop shop for creating tabular synthetic data. It is available to the public under the Business Source License. Additional plans are also available.



- **An Empirical Study of Utility and Disclosure Risk for Tabular Data Synthesis Models - In-Depth Analysis and Interesting Findings**
  - **BigComp ’24, Best Paper Award**

Utility and Disclosure Risk for Tabular Synthesis

- **Introduction**
  - 다양한 유형의 Tabular DataSet의 합성데이터를 **Utility 및 Disclosure Risk에 대해 최초로 심층 분석함**
- **DataSet 및 생성모델**
  - **데이터셋(3)**: Customer, Cencus, Ethereum
  - **생성모델(4):** Copula, CART, CTGAN, VAE
- **측정 지표**
  - **Utility Metrics:** 합성데이터가 원본데이터의 통계적 분포를 얼마나 잘 모사하는가를 측정하는 지표임
    - PC(Pair-wise correlation), Statistics, PMSE-based Score
  - **Disclosure Risk Metrics:** 합성 데이터가 원본 데이터로부터 개인 식별자를 유추할 수 있는 가능성을 얼마나 잘 줄였는지를 평가
    - GU, TCAP

Utility and Disclosure Risk for Tabular Synthesis

Dataset	Description	#Record	#Numerical	#Categorical	Sparsity Ratio	#Quasi-Identifier
Customer	Mart Customer Journey	2216	13	14	9.81%	4
Cencus	Sample from Census Bureau	32534	6	11	2.4%	6
Ethereum	Fraudulent and Valid Transactions	8981	33	16	53.63%	6

Dataset	Model	Volume	Utility Metrics			Disclosure Risk Metrics				
			Pair-wise Correlation	Statistics	PMSE-based Score	TCAP(T1)	TCAP(T2)	TCAP(T3)	TCAP(avg)	GU
Customer	Copula	same size	0.830	0.850	0.320	0.6600	0.005	0.002	0.222	0.005
		1.5 times	0.831	0.847	0.414	0.6740	0.005	0.003	0.227	0.009
		3 times	0.833	0.851	0.438	0.6810	0.005	0.003	0.230	0.010
	CART	same size	0.967	0.981	0.668	0.7193	0.004	0.001	0.241	0.007
		1.5 times	0.968	0.982	0.672	0.7366	0.004	0.001	0.247	0.016
		3 times	0.973	0.987	0.788	0.7383	0.005	0.002	0.248	0.024
	CTGAN	same size	0.833	0.869	0.380	0.6500	0.005	0.000	0.218	0.001
		1.5 times	0.834	0.869	0.460	0.6740	0.006	0.001	0.227	0.040
		3 times	0.835	0.869	0.599	0.7040	0.006	0.001	0.237	0.063
	VAE	same size	0.701	0.822	0.311	0.7010	0.004	0.001	0.235	0.005
		1.5 times	0.722	0.822	0.424	0.6890	0.004	0.002	0.232	0.017
		3 times	0.737	0.823	0.425	0.7030	0.005	0.000	0.236	0.020
Cencus	Copula	same size	0.875	0.856	0.447	0.1120	0.089	0.087	0.096	0.022
		1.5 times	0.869	0.861	0.400	0.1280	0.127	0.180	0.145	0.027
		3 times	0.875	0.862	0.453	0.2360	0.193	0.189	0.206	0.040
	CART	same size	0.884	0.994	0.996	0.1210	0.133	0.004	0.086	0.031
		1.5 times	0.884	0.996	1.000	0.2857	0.156	0.150	0.197	0.042
		3 times	0.887	0.997	1.000	0.3043	0.220	0.205	0.243	0.059
	CTGAN	same size	0.912	0.860	0.652	0.1070	0.094	0.003	0.068	0.017
		1.5 times	0.912	0.860	0.668	0.1850	0.097	0.003	0.095	0.023
		3 times	0.913	0.861	0.788	0.1760	0.089	0.002	0.089	0.031
	VAE	same size	0.870	0.857	0.550	0.1110	0.013	0.005	0.043	0.016
		1.5 times	0.869	0.863	0.602	0.1790	0.013	0.006	0.066	0.022
		3 times	0.886	0.871	0.613	0.1810	0.156	0.012	0.116	0.037
Ethereum	Copula	same size	0.837	0.780	0.007	0.0900	0.002	0.001	0.031	0.002
		1.5 times	0.846	0.793	0.116	0.0990	0.009	0.002	0.037	0.002
		3 times	0.848	0.814	0.123	0.1200	0.014	0.080	0.071	0.003
	CART	same size	0.905	0.951	0.152	0.0700	0.003	0.001	0.025	0.000
		1.5 times	0.906	0.950	0.152	0.0670	0.002	0.001	0.023	0.001
		3 times	0.907	0.952	0.260	0.0720	0.002	0.002	0.025	0.001
	CTGAN	same size	0.824	0.675	0.014	0.0200	0.001	0.000	0.007	0.000
		1.5 times	0.825	0.676	0.098	0.0230	0.006	0.001	0.010	0.000
		3 times	0.824	0.676	0.148	0.0240	0.007	0.001	0.011	0.001
	VAE	same size	0.794	0.669	0.005	0.0500	0.002	0.000	0.017	0.000
		1.5 times	0.816	0.670	0.031	0.0190	0.005	0.000	0.008	0.000
		3 times	0.816	0.670	0.120	0.0260	0.007	0.001	0.011	0.001

- **Graph-based Financial Data Synthesis**
  - **Explanation of Basic Ideas (Current Research Topic)**

금융데이터의 특징 (기존 모델들의 한계점)

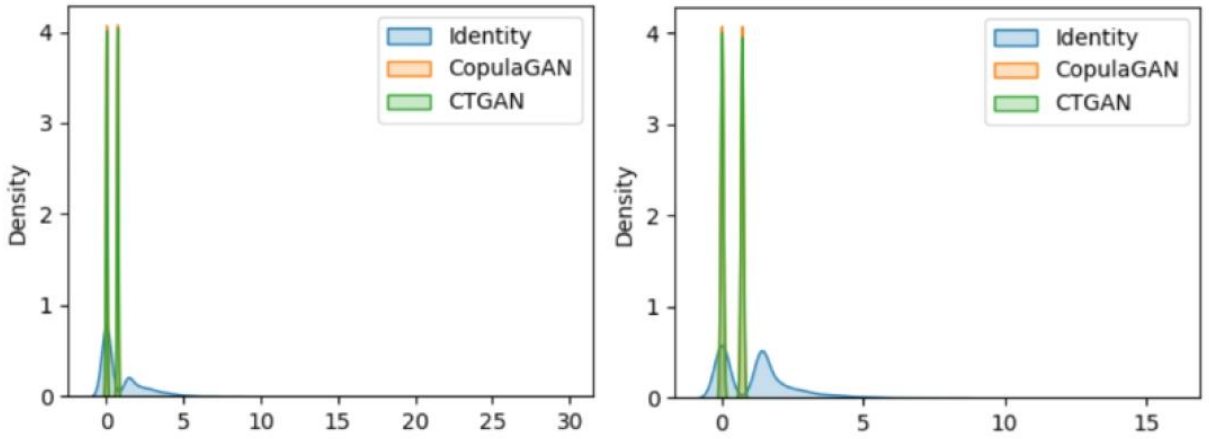
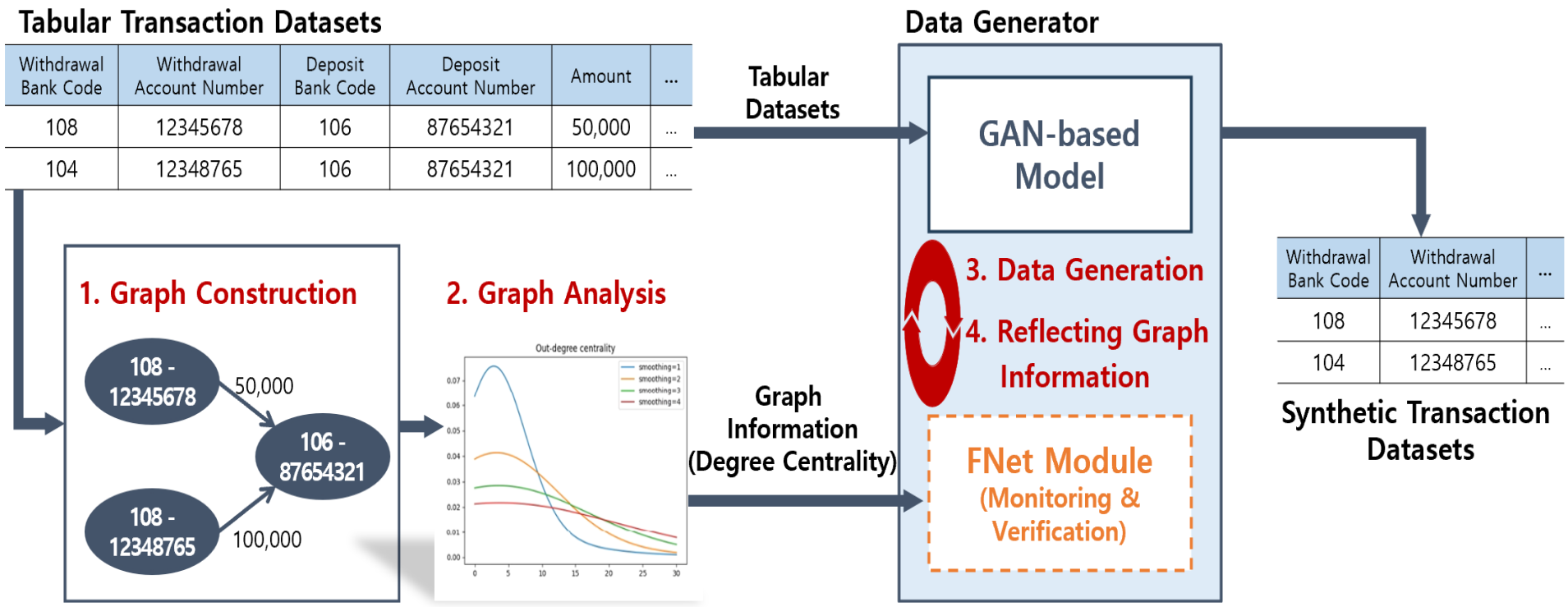


Figure 1: KDE plots of degree centrality measures for account graph.

Proposal Method & Future Work



감사합니다.