연구 주제 소개

컴퓨터과학과 증강지능 연구실 황승현

2023-09-19

목차

1. 이전 연구 현황
    1. KoGES 데이터
    2. Tabular Data Augmentation
2. Imbalanced Classification (불균형 분류)
3. Anomaly Detection (이상치 탐지)
4. 간단한 이상치 탐지 알고리즘 적용한 결과
5. 결론 및 향후 계획

이전 연구 현황

2023년 1학기에 한 연구

Deep Neural Network and Tabular Data

Tabular Data Augmentation

이전 연구

- KoGES 데이터를 이용한 질병 예측 모델 개발
- 식품영양학과
- 고혈압 – 완료

  Kim, H., Hwang, S., Lee, S., & Kim, Y. (2022). Classification and Prediction on Hypertension with Blood Pressure Determinants in a Deep Learning Algorithm. *International Journal of Environmental Research and Public Health*, *19*(22), 15301.
- 골다공증 – 진행중

KoGES

- 한국인유전체역학조사사업
  (Korean Genome and Epidemiology Study; KoGES)
    - 건강 및 생활습관 관련 설문조사 / 검진
    - 혈액, DNA 등 수집
- 표 형태 데이터 (Tabular Data)
    - 딥러닝으로 할 수 있지 않을까?

DNN and Tabular Data

심층신경망을 이용하여 표 형태 데이터를 분
석하는 모델 survey 논문

IEEE Transactions on Neural
Networks and Learning Systems

Borisov, V., Leemann, T., Seßler, K.,
Haug, J., Pawelczyk, M., & Kasneci, G.

Deep Neural Networks and Tabular Data: A Survey

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug,
Martin Pawelczyk and Gjergji Kasneci

***Abstract*—Heterogeneous tabular data are the most commonly used form of data and are essential for numerous critical and computationally demanding applications. On homogeneous data sets, deep neural networks have repeatedly shown excellent performance and have therefore been widely adopted. However, their adaptation to tabular data for inference or data generation tasks remains highly challenging. To facilitate further progress in the field, this work provides an overview of state-of-the-art deep learning methods for tabular data. We categorize these methods into three groups: data transformations, specialized architectures, and regularization models. For each of these groups, our work offers a comprehensive overview of the main approaches. Moreover, we discuss deep learning approaches for generating tabular data, and we also provide an overview over strategies for explaining deep models on tabular data. Thus, our first contribution is to address the main research streams and existing methodologies in the mentioned areas, while highlighting relevant challenges and open research questions. Our second contribution is to provide an empirical comparison of traditional machine learning methods with eleven deep learning approaches across five popular real-world tabular data sets of different sizes and with different learning objectives. Our results, which we have made publicly available as competitive benchmarks, indicate that algorithms based on gradient-boosted tree ensembles still mostly outperform deep learning models on supervised learning tasks, suggesting that the research progress on competitive deep learning models for tabular data is stagnating. To the best of our knowledge, this is the first in-depth overview of deep learning approaches for tabular data; as such, this work can serve as a valuable starting point to guide researchers and practitioners interested in deep learning with tabular data.**

***Index Terms*—Deep neural networks, Tabular data, Heterogeneous data, Discrete data, Tabular data generation, Probabilistic modeling, Interpretability, Benchmark, Survey**

I. Introduction

Ever-increasing computational resources and the availability of large, labelled data sets have accelerated the success of deep neural networks [1], [2]. In particular, architectures based on convolutions, recurrent mechanisms [3], or transformers [4] have led to unprecedented performance in a multitude of domains. Although deep learning methods perform outstandingly well for classification or data generation tasks on homogeneous data (e.g., image, audio, and text data), tabular data still pose a challenge to deep learning models [5]–[8]. Tabular data – in contrast to image or language data – are heterogeneous, leading to dense numerical and sparse categorical features. Furthermore, the correlation among the features is weaker than the one introduced through spatial or semantic relationships in image or speech data. Hence, it is necessary to discover and exploit relations without relying on spatial information [9]. Therefore, Kadra et al. called tabular data sets the last *"unconquered castle"* for deep neural network models [10].

Heterogeneous data are the most commonly used form of data [7], and it is ubiquitous in many crucial applications, such as medical diagnosis based on patient history [11]–[13], predictive analytics for financial applications (e.g., risk analysis, estimation of creditworthiness, the recommendation of investment strategies, and portfolio management) [14], click-through rate (CTR) prediction [15], user recommendation systems [16], customer churn prediction [17], [18], cybersecurity [19], fraud detection [20], identity protection [21], psychology [22], delay estimations [23], anomaly detection [24], and so forth. In all these applications, a boost in predictive performance and robustness may have considerable benefits for both end users and companies that provide such solutions. Simultaneously, this requires handling many data-related pitfalls, such as noise, impreciseness, different attribute types and value ranges, or the missing value problem and privacy issues.

Meanwhile, deep neural networks offer multiple advantages over traditional machine learning methods. First, these methods are highly flexible [25], allow for efficient and iterative training, and are particularly valuable for AutoML [26]–[31]. Second, tabular data generation is possible using deep neural networks and can, for instance, help mitigate class imbalance problems [32]. Third, neural networks can be deployed for multimodal learning problems where tabular data can be one of many input modalities [28], [33]–[36], for tabular data distillation [37], [38], for federated learning [39], and in many more scenarios.

Successful deployments of data-driven applications require solving several tasks, among which we identified three *core challenges*: (1) *inference* (2) *data generation*, and (3) *interpretability*. The most crucial task is inference which is concerned with making predictions based on past observations. While a powerful predictive model is critical for all the applications mentioned in the previous paragraph, the interplay between tabular data and deep neural networks goes beyond simple inference tasks. Before a predictive model can even be trained, the training data usually needs to be preprocessed. This is where data generation plays a crucial role, as one of the standard deployment steps involves the imputation of missing values [40]–[42] and the rebalancing of the data set [43], [44] (i.e., equalizing sample sizes for different classes). Furthermore, it might be simply impossible to use

All authors are with the Data Science and Analytics Research (DSAR) group at the University of Tübingen, 72070 Tübingen, Germany. Gjergji Kasneci is also affiliated with Schufa Holding AG, 65201 Wiesbaden, Germany.
Corresponding authors:
vadim.borisov@uni-tuebingen.de
tobias.leemann@uni-tuebingen.de

DNN and Tabular Data

- 결론
  - GBDT are still state-of-the-art
  - Fundamental reorientation
  - Deep Learning is not suitable for tabular data.
- 딥러닝이 표에 약한 이유
  - Low-Quality
  - Missing or Complex
  - Single Features
  - Preprocessing

DNN and Tabular Data

Survey 논문의 비관적인 결론에도 불구하고…

여러 딥러닝 기반 Tabular data 분석 모델

<u>SAINT</u>에 관심을 가지게 됨

	HELOC		Adult		HIGGS		Covertype		Cal. Housing
	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	MSE ↓
Linear Model	73.0±0.0	80.1±0.1	82.5±0.2	85.4±0.2	64.1±0.0	68.4±0.0	72.4±0.0	92.8±0.0	0.528±0.008
KNN [65]	72.2±0.0	79.0±0.1	83.2±0.2	87.5±0.2	62.3±0.1	67.1±0.0	70.2±0.1	90.1±0.2	0.421±0.009
Decision Tree [197]	80.3±0.0	89.3±0.1	85.3±0.2	89.8±0.1	71.3±0.0	78.7±0.0	79.1±0.0	95.0±0.0	0.404±0.007
Random Forest [198]	82.1±0.2	90.0±0.2	86.1±0.2	91.7±0.2	71.9±0.0	79.7±0.0	78.1±0.1	96.1±0.0	0.272±0.006
XGBoost [53]	<u>83.5±0.2</u>	92.2±0.0	<u>87.3±0.2</u>	<u>92.8±0.1</u>	<u>77.6±0.0</u>	<u>85.9±0.0</u>	**97.3±0.0**	**99.9±0.0**	0.206±0.005
LightGBM [78]	<u>83.5±0.1</u>	<u>92.3±0.0</u>	**87.4±0.2**	**92.9±0.1**	77.1±0.0	85.5±0.0	93.5±0.0	99.7±0.0	**0.195±0.005**
CatBoost [79]	**83.6±0.3**	**92.4±0.1**	87.2±0.2	<u>92.8±0.1</u>	77.5±0.0	85.8±0.0	<u>96.4±0.0</u>	<u>99.8±0.0</u>	<u>0.196±0.004</u>
Model Trees [199]	82.6±0.2	91.5±0.0	85.0±0.2	90.4±0.1	69.8±0.0	76.7±0.0	-	-	0.385±0.019
MLP [200]	73.2±0.3	80.3±0.1	84.8±0.1	90.3±0.2	77.1±0.0	85.6±0.0	91.0±0.4	76.1±3.0	0.263±0.008
DeepFM [15]	73.6±0.2	80.4±0.1	86.1±0.2	91.7±0.1	76.9±0.0	83.4±0.0	-	-	0.260±0.006
DeepGBM [70]	78.0±0.4	84.1±0.1	84.6±0.3	90.8±0.1	74.5±0.0	83.0±0.0	-	-	0.856±0.065
RLN [72]	73.2±0.4	80.1±0.4	81.0±1.6	75.9±8.2	71.8±0.2	79.4±0.2	77.2±1.5	92.0±0.9	0.348±0.013
TabNet [5]	81.0±0.1	90.0±0.1	85.4±0.2	91.1±0.1	76.5±1.3	84.9±1.4	93.1±0.2	99.4±0.0	0.346±0.007
VIME [88]	72.7±0.0	79.2±0.0	84.8±0.2	90.5±0.2	76.9±0.2	85.5±0.1	90.9±0.1	82.9±0.7	0.275±0.007
TabTransformer [98]	73.3±0.1	80.1±0.2	85.2±0.2	90.6±0.2	73.8±0.0	81.9±0.0	76.5±0.3	72.9±2.3	0.451±0.014
NODE [6]	79.8±0.2	87.5±0.2	85.6±0.3	91.1±0.2	76.9±0.1	85.4±0.1	89.9±0.1	98.7±0.0	0.276±0.005
Net-DNF [57]	82.6±0.4	91.5±0.2	85.7±0.2	91.3±0.1	76.6±0.1	85.1±0.1	94.2±0.1	99.1±0.0	-
STG [201]	73.1±0.1	80.0±0.1	85.4±0.1	90.9±0.1	73.9±0.1	81.9±0.1	81.8±0.3	96.2±0.0	0.285±0.006
NAM [202]	73.3±0.1	80.7±0.3	83.4±0.1	86.6±0.1	53.9±0.6	55.0±1.2	-	-	0.725±0.022
SAINT [9]	82.1±0.3	90.7±0.2	86.1±0.3	91.6±0.2	**79.8±0.0**	**88.3±0.0**	96.3±0.1	<u>99.8±0.0</u>	0.226±0.004

TABLE V: Open performance benchmark results based on (stratified) 5-fold cross-validation. We use the same fold splitting strategy for every data set. The top results for each dataset are in **bold**, we also <u>underline</u> the second-best results. The mean and standard deviation values are reported for each baseline model. Missing results indicate that the corresponding model could not be applied to the task type (regression or multi-class classification).

SAINT

SAINT: Self-Attention and Intersample Attention Transformer

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T.

SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training

Gowthami Somepalli
Department of Computer Science
University of Maryland, College Park
gowthami@umd.edu

Micah Goldblum
Department of Computer Science
University of Maryland, College Park
goldblum@umd.edu

Avi Schwarzschild
Department of Mathematics
University of Maryland, College Park
avi1@umd.edu

C. Bayan Bruss
Capital One
Center for Machine Learning
bayan.bruss@capitalone.com

Tom Goldstein
Department of Computer Science
University of Maryland, College Park
tomg@umd.edu

Abstract

Tabular data underpins numerous high-impact applications of machine learning from fraud detection to genomics and healthcare. Classical approaches to solving tabular problems, such as gradient boosting and random forests, are widely used by practitioners. However, recent deep learning methods have achieved a degree of performance competitive with popular techniques. We devise a hybrid deep learning approach to solving tabular data problems. Our method, SAINT, performs attention over both rows and columns, and it includes an enhanced embedding method. We also study a new contrastive self-supervised pre-training method for use when labels are scarce. SAINT consistently improves performance over previous deep learning methods, and it even outperforms gradient boosting methods, including XGBoost, CatBoost, and LightGBM, on average over a variety of benchmark tasks.

1 Introduction

While machine learning for image and language processing has seen major advances over the last decade, many critical industries, including financial services, health care, and logistics, rely heavily on data in structured table format. Tabular data is unique in several ways that have prevented it from benefiting from the impressive success of deep learning in vision and language. First, tabular data often contain heterogeneous features that represent a mixture of continuous, categorical, and ordinal values, and these values can be independent or correlated. Second, there is no inherent positional information in tabular data, meaning that the order of columns is arbitrary. This differs from text, where tokens are always discrete, and ordering impacts semantic meaning. It also differs from images, where pixels are typically continuous, and nearby pixels are correlated. Tabular models must handle features from multiple discrete and continuous distributions, and they must discover correlations without relying on the positional information. Sufficiently powerful deep learning systems for tabular data have the potential to improve performance beyond what is achieved by classical methods, like linear classifiers and random forests. Furthermore, without performant deep learning models for

Preprint. Under review.

SAINT

같은 열에 있는 데이터 비교

같은 행에 있는 데이터 비교

트랜스포머

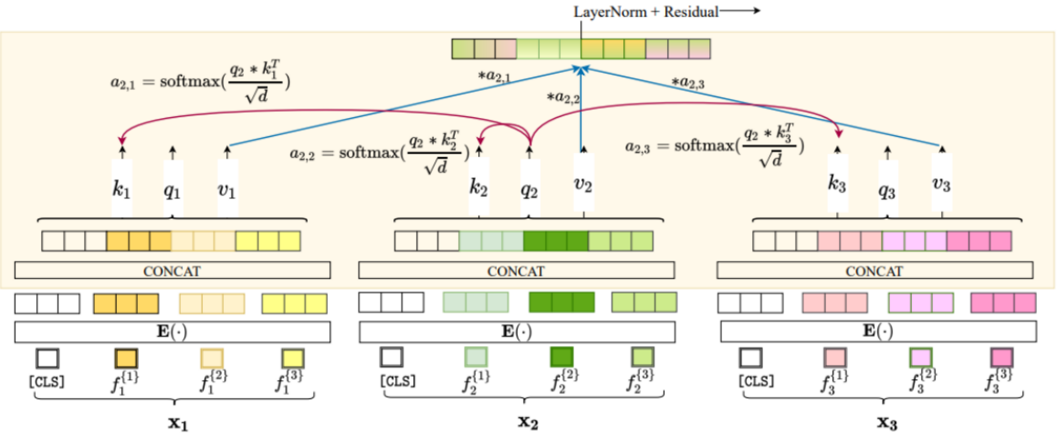
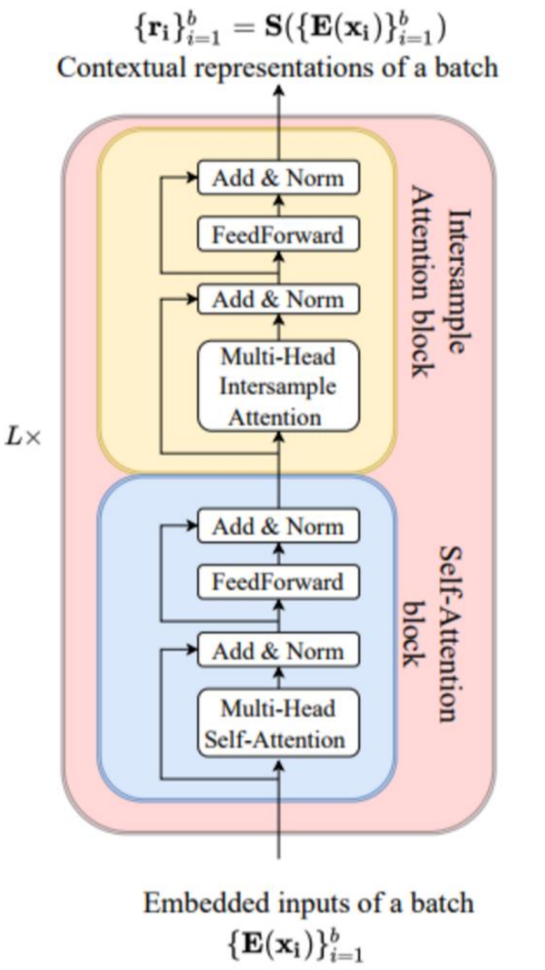


Figure 2: Intersample attention on a batch of 3 points. In this plot, d is the size of value vectors v_i. See Section 3.2 for details.

Algorithm 1 PyTorch-style pseudo-code for intersample attention. For simplicity, we describe just one head and assume the value vector dimension is same as the input embedding dimension.

```
# b: batch size, n: number of features, d: embedding dimension
# W_q, W_k, W_v are weight matrices of dimension dxd
# mm: matrix-matrix multiplication
def self_attention(x):
    # x is bxnxd
    q, k, v = mm(W_q,x), mm(W_k,x), mm(W_v,x) #q,k,v are bxnxd
    attn = softmax(mm(q,np.transpose(k, (0, 2, 1)))/sqrt(d)) # bxnxn
    out = mm(attn, v) #out is bxnxd
    return out

def intersample_attention(x):
    # x is bxnxd
    b,n,d = x.shape # as mentioned above
    x = reshape(x, (1,b,n*d)) # reshape x to 1xbx(n*d)
    x = self_attention(x) # the output x is 1xbx(n*d)
    out = reshape(x,(b,n,d)) # out is bxnxd
    return out
```



(a)

SAINT – Augmentation...?

- 이미지의 증강 기법을 표 형태 데이터 증강에 적용
- CutMix:
  이미지의 특정 픽셀 다른 이미지로 대체
- Mixup:
  다른 특징과 레이블 결합, 신경망이 특징과 레이블에 과적합 되지 않도록 함

$$\mathbf{x_i'} = \mathbf{x_i} \odot \mathbf{m} + \mathbf{x_a} \odot (\mathbf{1} - \mathbf{m}) \quad \text{CutMix in raw data space}$$

$$\mathbf{p_i'} = \alpha * \mathbf{E}(\mathbf{x_i'}) + (1 - \alpha) * \mathbf{E}(\mathbf{x_b'}) \quad \text{mixup in embedding space}$$

Imbalanced Classification

Imbalanced Classification (불균형 분류) 의 정의와 기존 접근법 소개

Imbalanced classification

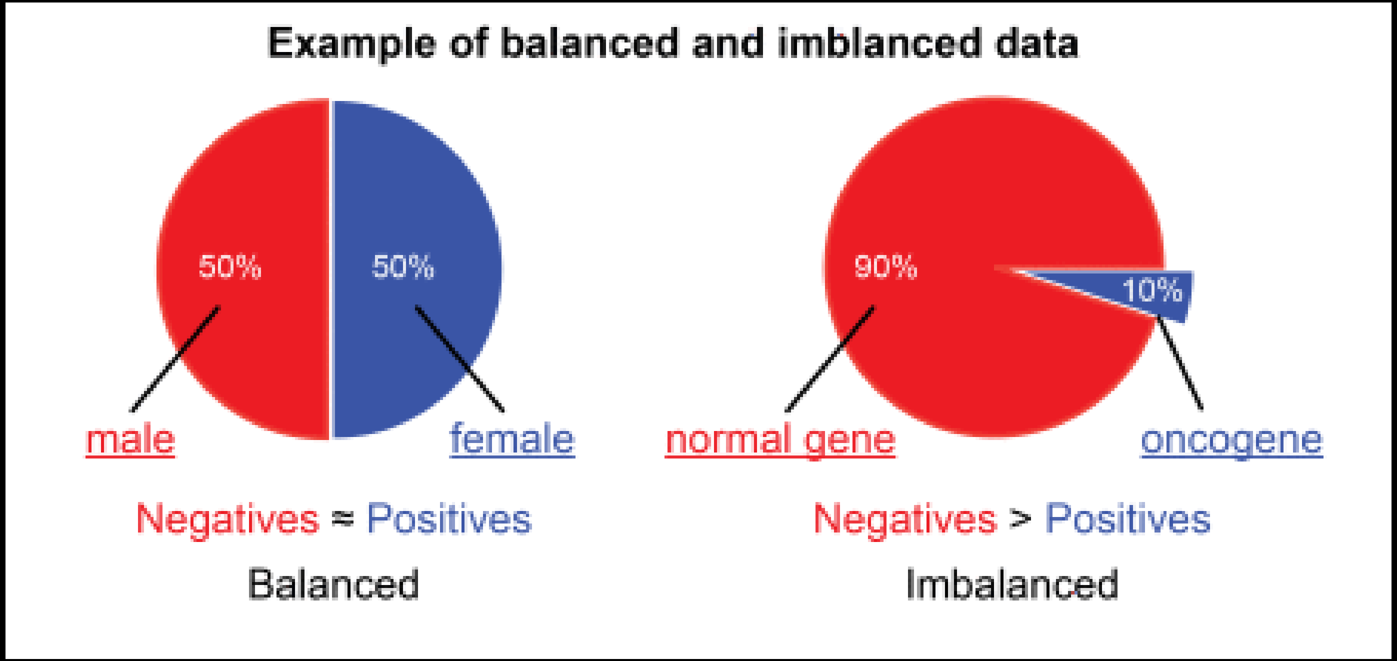
불균형 분류

각 클래스 비율 크게 차이

고혈압 분류: 정상 90명 비정상 10명

Rare class : 소수 (환자)

Abundant class : 다수 (정상인)



Imbalanced classification 문제점

- 대충 만들어도 Accuracy 가 높게 나온다
  - 모델이 모두 정상이라고 판단하면 된다.
  - 정상인 90명 환자 10명을 분류할 때, 모두 정상인이라고 분류하면?
  - 정확도 90%

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

접근법 : Resampling

Oversampling

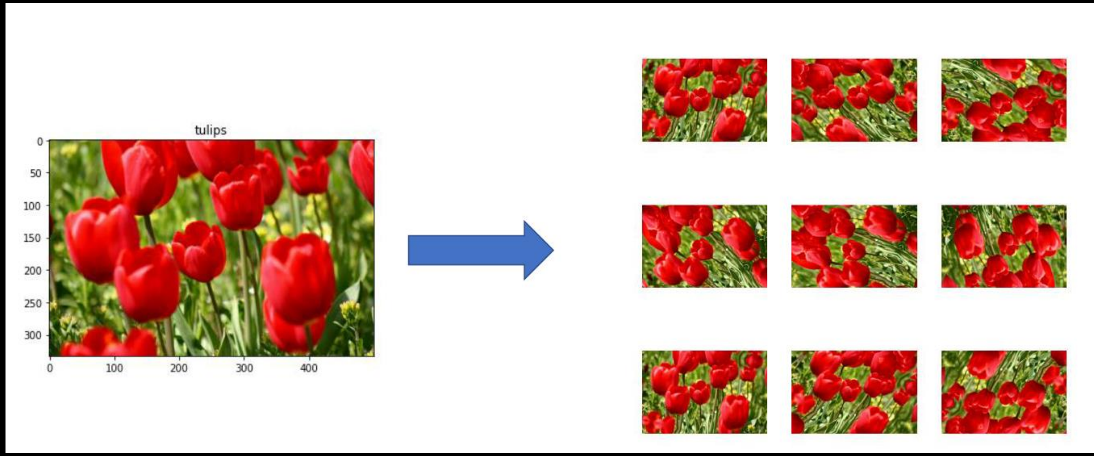
- Rare class 늘리기 (증강)
- 소수를 복제하여 다수에 맞춤
- SMOTE, ADASYN 등

Undersampling

- Abundant class 줄이기
- 다수를 소수에 맞게 자름
- NCL, OSS 등

데이터 증강

- Data Augmentation
- 양이 적은 데이터를 여러 알고리즘으로 양을 늘리는 것
- SMOTE
  - Synthetic Minority Oversampling TEchnique.

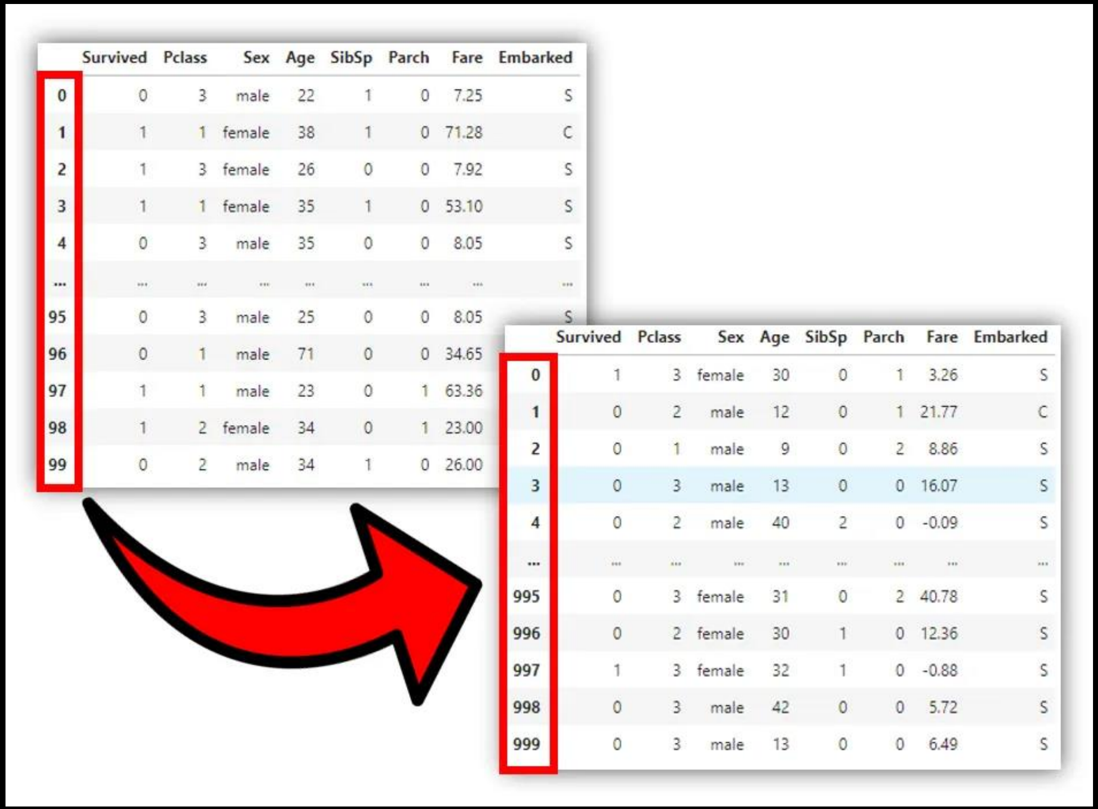


Tabular Data Augmentation

표 형태 데이터 증강

Tabular Data Augmentation

- 선행 연구 고찰
  - Fed-TDA, SDAT, VIME
  - 그 외... 별로 없음.
- 성능이 특출나지는 않음..
- 오히려 좋아?



	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22	1	0	7.25	S
1	1	1	female	38	1	0	71.28	C
2	1	3	female	26	0	0	7.92	S
3	1	1	female	35	1	0	53.10	S
4	0	3	male	35	0	0	8.05	S
...	...	...	...	...	...	...	...	...
95	0	3	male	25	0	0	8.05	S
96	0	1	male	71	0	0	34.65	
97	1	1	male	23	0	1	63.36	
98	1	2	female	34	0	1	23.00	
99	0	2	male	34	1	0	26.00	

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	3	female	30	0	1	3.26	S
1	0	2	male	12	0	1	21.77	C
2	0	1	male	9	0	2	8.86	S
3	0	3	male	13	0	0	16.07	S
4	0	2	male	40	2	0	-0.09	S
...	...	...	...	...	...	...	...	...
995	0	3	female	31	0	2	40.78	S
996	0	2	female	30	1	0	12.36	S
997	1	3	female	32	1	0	-0.88	S
998	0	3	male	42	0	0	5.72	S
999	0	3	male	13	0	0	6.49	S

Tabular … tion

학과	컴퓨터과학과
학번	2023210125
이름	황승현

과목명	담당교수	담당확인
기술문서 작성및발표	김민기	

Tabular data Augmentation for Deep Learning. 연구계획서

2023. 6. 15.

경상국립대학교
Gyeongsang National University

Fed-TDA: Federated Tabular Data Augmentation o

Shaoming Duan[1,2], Chuanyi Liu[1,2,3*], Peiyi Han[1,2,3], Tianyu He[1,2], Yifeng
Zhang[1], Xinyu zha[1]

[1]School of Computer Science, Harbin Institute of Technology (Shenzhen), Sh
[2]Insititute of Data Security, Harbin Institute of Technology (Shenzhen), She
[3]Department of New Networks, Peng Cheng Laboratory, Shenzhen 5

arXiv:2211.13116v2 [cs.LG] 12 Jan 2023

Abstract

Non-independent and identically distributed (non-IID) data is a key challenge in federated learning (FL), which usually hampers the optimization convergence and the performance of FL. Existing data augmentation methods based on federated generative models or raw data sharing strategies still suffer from low performance, privacy protection concerns, and high communication overhead in decentralized tabular data. To tackle these challenges, we propose a federated tabular data augmentation (Fed-TDA) method that synthesizes fake tables using some low-dimensional statistics (e.g., distributions of each column and global covariance). Specifically, we propose the multimodal distribution transformation and inverse cumulative distribution mapping to synthesize continuous and discrete columns in tabular data according to the pre-learned statistics, respectively. Furthermore, we theoretically analyze that our Fed-TDA not only preserves data privacy but also maintains the distribution of the original data and the correlation between columns. Through extensive experiments on five real-world tabular datasets, we demonstrate the superiority of Fed-TDA over the state-of-the-art methods in test performance, statistical similarity, and communication efficiency. Code is available at https://github.com/smduan/Fed-TDA.git.

Introduction

In real-world applications, tabular data is the most common data type (Shwartz-Ziv and Armon 2022), which has been widely used in many relational database-based applications, such as medicine, finance, manufacturing, climate science, etc. Numerous organizations are using these data and machine learning (ML) to optimize their processes and performance. The wealth of data provides huge opportunities for ML applications. However, most of these tabular data are highly sensitive and typically distributed across different organizations. Due to privacy and regulatory concerns, these organizations are reluctant to share their private data.

In response to these concerns, federated learning (FL) (McMahan et al. 2017; Li et al. 2020a) has been extensively studied in recent years, where multiple participants jointly train a shared deep learning model under the coordination of a central server without transmitting their local data. Since

*The corresponding author: liuchuanyi@hit.edu.cn.

FL provides a feasible s
ity, and data ownership,
lar data applications, e.
2022), drug discovery
prediction (Qi et al. 20
FL is that the data of d
and identically distribu
degrade the performanc
et al. 2022).

Existing studies for s
roughly divided into tw
ods and *data-based met*
igate the weight drift b
signing appropriate los
and Song 2021) or per
gies (Zhu et al. 2021;
our experimental result
show that these method
vanilla FedAvg (McMa
ods perform data augm
et al. 2021; Jeong et al.
(Wen et al. 2020; Auge
which convert non-IID
eliminate data distributi
based methods, numer
that data-based methods
mance of FL (Yoon et a
2022).

Unfortunately, exist
methods directly applie
following challenges. F
high performance on t
data usually consists of
columns. Synthesizing
challenge for GAN-bas
column distributions an
et al. 2019; Lee et al. 20
further degrade the per
et al. 2021a). Secondly,
tains sensitive informat
erators with other partie
private data. Finally, ex
tween the server and c
tional communication o

Semi-Supervised Learning with Data Augmentation for Tabular Data

Junpeng Fang
junpeng.fjp@antgroup.com
Ant Group
Hangzhou, China

Caizhi Tang
Qing Cui
caizhi.tcz@antgroup.com
cuiqing.cq@antgroup.com
Ant Group
Hangzhou, China

Feng Zhu
Longfei Li
zhufeng.zhu@antgroup.com
longyao.llf@antgroup.com
Ant Group
Hangzhou, China

Jun Zhou*
jun.zhoujun@antgroup.com
Ant Group
Hangzhou, China

Wei Zhu
wei.wz@antgroup.com
Ant Group
Hangzhou, China

ABSTRACT

Data augmentation-based semi-supervised learning (SSL) methods have made great progress in computer vision and natural language processing areas. One of the most important factors is that the semantic structure invariance of these data allows the augmentation procedure (e.g., rotating images or masking words) to thoroughly utilize the enormous amount of unlabeled data. However, the tabular data does not possess an obvious invariant structure, and therefore similar data augmentation methods do not apply to it. To fill this gap, we present a simple yet efficient data augmentation method particular designed for tabular data and apply it to the SSL algorithm: SDAT (*Semi-supervised learning with Data Augmentation for Tabular data*). We adopt a multi-task learning framework that consists of two components: the data augmentation procedure and the consistency training procedure. The data augmentation procedure which perturbs in latent space employs a variational auto-encoder (VAE) to generate the reconstructed samples as augmented samples. The consistency training procedure constrains the predictions to be invariant between the augmented samples and the corresponding original samples. By sharing a representation network (encoder), we jointly train the two components to improve effectiveness and efficiency. Extensive experimental studies validate the effectiveness of the proposed method on the tabular datasets.

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning.**

KEYWORDS

semi-supervised learning, tabular data, data augmentation

*Corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557699

ACM Reference Format:
Junpeng Fang, Caizhi Tang, Qing Cui, Feng Zhu, Longfei Li, Jun Zhou, and Wei Zhu. 2022. Semi-Supervised Learning with Data Augmentation for Tabular Data. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557699

1 INTRODUCTION

Deep learning has made progress in many fields such as computer vision [1, 17], search engine [6, 9], and recommendation [7, 14], etc., while it largely depends on massive labeled data. The collection of labeled data often needs manual annotation, which is expensive, and even inaccessible (some information involves user privacy). Semi-supervised learning (SSL) is a good solution for those scenarios with limited labeled data but a large number of unlabeled data.

The recent studies in SSL [2, 3, 18, 19] are mainly based on consistency regularization, especially through data augmentation. Generally, consistency regularization enforces that an unlabeled example x to be predicted the same as its augmentation, like

$$\min D(p(y|x;\theta), p(y|\hat{x};\theta)) \quad (1)$$

where D is a distance metric, θ is the trainable parameters, and $\hat{x}$ is an augmentation of x. In a nutshell, Eq. 1 simply regularizes model predictions to be invariant to augmentation (from x to $\hat{x}$).

There are simple augmentation methods for image and text data such as rotating images or masking words. However, tabular data does not possess an invariant structure, and existing data augmentation methods used in CV and NLP areas are not applicable for tabular data. VIME [19] filled this gap for tabular data, but it directly perturbs the sample space which could change the semantics and thus be inefficient. To address the above issues, we propose SDAT. Our method consists of two components: (1) data augmentation and (2) consistency training. Specifically, we employ a generative model which is a variational auto-encoder (VAE) [10] to learn the latent semantically invariant knowledge[1] and add noise in the latent space to obtain an augmented sample. It is an intuitive assumption that the reconstruction of an instance by the generative model will probably keep its semantics (label).

[1]We design this method specifically for tabular data. For image and text data, simple rotation and synonyms substitution are more suitable.

Tabular Data Augmentation

- 선행 연구 고찰
  - Fed-TDA, SDAT, VIME
  - 그 외... 별로 없음.
- 성능이 특출나지는 않음..
- 오히려 좋아?

- 구현 어려움.
- 선행 연구 많이 없음.
  - 참고 자료 없음
  - 이 방향이 맞나 의문..

Tabular Data Augmentation

뭔가 논?문을 쓰기는 함

딥러닝 기반 표 형태 데이터 증강의 문제와 한계: GBDT 모델과의 비교 연구

황승현, 이수원

경상국립대학교 컴퓨터과학과
{ghkd3531, leesuwon}@gnu.ac.kr

Challenges and Limitations of Deep Learning-Based Tabular Data Augmentation: A Comparative Study with GBDT Models

Seunghyeon Hwang, Suwon Lee
Department of Computer Science, Gyeongsang National University, Jinju, Korea

요 약

이 논문은 딥러닝을 사용하여 Tabular 데이터를 분석하는 것과 GBDT(Gradient Boosted Decision Tree) 모델과의 정확도 차이를 조사하였다. 표 형태 데이터는 결측값, 이상치, 데이터 불균형 등의 문제로 인해 전처리와 의존성 처리가 어렵다는 한계가 있다. 이러한 한계를 극복하고 데이터 분석의 성능을 개선하기 위해 표 데이터를 증강하는 모델을 개발하는 것이 중요하다. 관련 연구를 살펴보면, 딥러닝을 활용한 데이터 증강 방법은 성능 향상에 한계가 있음을 확인할 수 있다. 실험 결과에서도 GBDT 기반 모델인 XGBoost의 성능이 딥러닝 기반 모델보다 우수한 것으로 나타났다. 이러한 결과는 딥러닝 기반 표 형태 데이터 증강의 한계를 강조하고, GBDT 모델의 우수성을 확인하였으며, 다른 연구자들에게 해당 주제로의 연구를 권장하지 않는다는 결론을 내렸다.

1. 서 론

Tabular Data는 우리 현실의 많은 분야에서 쓰이고 있다. 스프레드시트, 데이터베이스 등 우리가 흔히 모으고, 만들고, 활용하는 데이터는 모두 표 형태이다. 딥러닝은 최근 이미지, 자연어 분야에서 엄청난 활약을 펼치고 있다. 그러나, 여러 벤치마크[1]에서 알 수 있듯이 딥러닝으로 Tabular 데이터를 분석하는 것은 GBDT(Gradient Boosted Decision Tree)보다 정확도가 떨어진다. 표 형태 데이터의 정확도가 낮은 이유는 다음과 같다. 전처리에 민감하며, 이기종 데이터로 분석이 어렵고, 변수 간 관계가 복잡하며, 데이터 수가 적다.

표 형태 데이터는 다양한 분야에서 매우 많이 사용되는 형태이다. 그러나 이러한 데이터는 수량이 제한적이라, 데이터 분석 모델의 정확도를 향상하는 것은 한계가 있다. 따라서 표 데이터를 증강하는 모델을 개발하여 데이터 분석의 성능을 개선하는 것은 매우 중요하다.

표 데이터를 증강하는 모델은 다양한 분야에서 활용할 수 있다. 예를 들어, 의료 분야에서는 환자의 건강 정보가 담긴 표 데이터를 증강하여 더 정확한 진단 및 치료 방법을 제공할 수 있다. 기업에서는 고객 데이터나 금융 데이터를 분석하여 새로운 비즈니스 모델을 개발할 수 있다.

2. 관련 연구

표 데이터를 증강하는 모델 연구는 많이 진행되고 있다 [2][3]. [2]은 데이터 증강 네트워크와 일관성 훈련 네트워크 2개를 구축하는 아이디어를 제시하였다. 데이터 증강 네트워크는 영상 분야에서 많이 쓰는 Variational auto-encoder (VAE)[4]를 이용하였다. 일관성 네트워크는 데이터 증강 네트워크에서 증강된 샘플과 해당 원본 샘플 간에 예측이 불변하도록 제한한다. 이때 두 샘플의 차이를 KL-divergence 함수를 이용하여 구한다.

[3]은 기존 표 형태 데이터에 랜덤하게 마스킹하여 데이터를 훼손하고 복원하는 것을 훈련하면서 모델을 학습한다. 이 과정은 특징 벡터와 마스크 벡터를 추정하면서 이루어진다. 특징 벡터 추정은 손상된 특징의 값을 예측하고, 손상된 변형을 복구한다. 마스크 벡터 추정은 어떤 특징이 마스킹 되었는지 예측한다.

3. 문제 정의

딥러닝 기반 모델은 영상, 자연어 처리 분야에서 다른 알고리즘을 압도한다. 그렇다면 딥러닝 기반으로 표 형태 데이터를 증강하였을 때, 기존 모델보다 성능이 좋아지는지 실험하였다.

표 1 데이터셋의 특성

Dataset	Adult Income	BlogFeedback	MNIST
Instances	32561	60021	70000
Features	123	281	784
Classes	2	2	10
Task	Binary	Regression	Multi-class

다양한 특성의 데이터셋을 가져와 각각의 모델을 평가하였다. 이 실험에 사용한 데이터셋은 Adult Income, Blog Feedback, MNIST이다. 데이터셋은 UC Irvine Machine Learning Repository등에서 찾을 수 있다. 연속형 데이터는 같은 분포로 정규화하고, 범주형 데이터는 원핫인코딩으로 벡터화하였다.

4. 실험

각 모델의 하이퍼파라미터는 최적의 성능을 낼 수 있도록 적절하게 튜닝하였다. 모델은 5겹으로 교차검증하였다.

표 2 데이터 증강 모델과 기존 모델의 성능 비교

Model / Dataset	XGBoost	DNN	VIME	SDAT
Adult Income	**0.862±0.002**	0.773±0.011	0.788±0.014	0.808±0.012
BlogFeedback	**0.798±0.003**	0.684±0.012	0.705±0.013	0.719±0.016
MNIST	**0.941±0.002**	0.821±0.010	0.843±0.017	0.859±0.014

5. 결론

[2], [3] 모두 딥러닝 기반으로 표 형태 데이터를 증강하였을 때, 기존 모델보다 성능이 뛰어나게 좋아지지 않았다. 표 2에서 볼 수 있듯, 딥러닝 기반 최신 표 형태 데이터 증강 모델 2가지와 GBDT 기반 모델 XGBoost, Deep Neural Network를 비교하였을 때, 모든 데이터셋에서 XGBoost의 성능이 뛰어났다.

본 논문에서는 기존의 연구된 딥러닝 기반 표 형태 데이터 증강 알고리즘의 한계를 보여주고, GBDT 기반 모델이 아직 뛰어난 성능을 보여준다는 것을 밝혔다. 다른 연구자에게 이 주제로 연구를 하는 것은 추천하지 않는다.

참고문헌

[1] Borisov, Vadim, Tobias Leemann, Seßler, Haug, Pawelczyk, and Gjergji Kasneci. "Deep neural networks and tabular data: A survey." IEEE Transactions on Neural Networks and Learning Systems. 2022.

[2] Fang, Junpeng, Caizhi Tang, Qing Cui, Feng Zhu, Longfei Li, Jun Zhou, and Wei Zhu. "Semi-Supervised Learning with Data Augmentation for Tabular Data." In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 3928-3932. 2022.

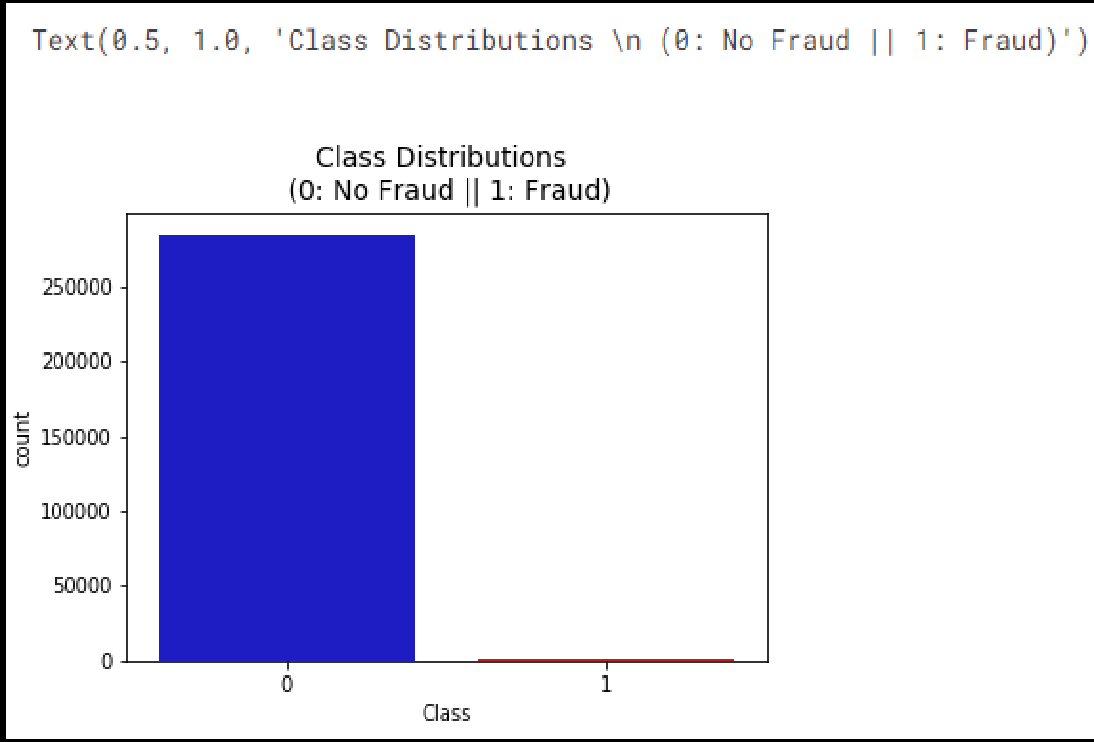
[3] Yoon, Zhang, Jordon, and Mihaela van der Schaar. "Vime: Extending the success of self-and semi-supervised learning to tabular domain." Advances in Neural Information Processing Systems

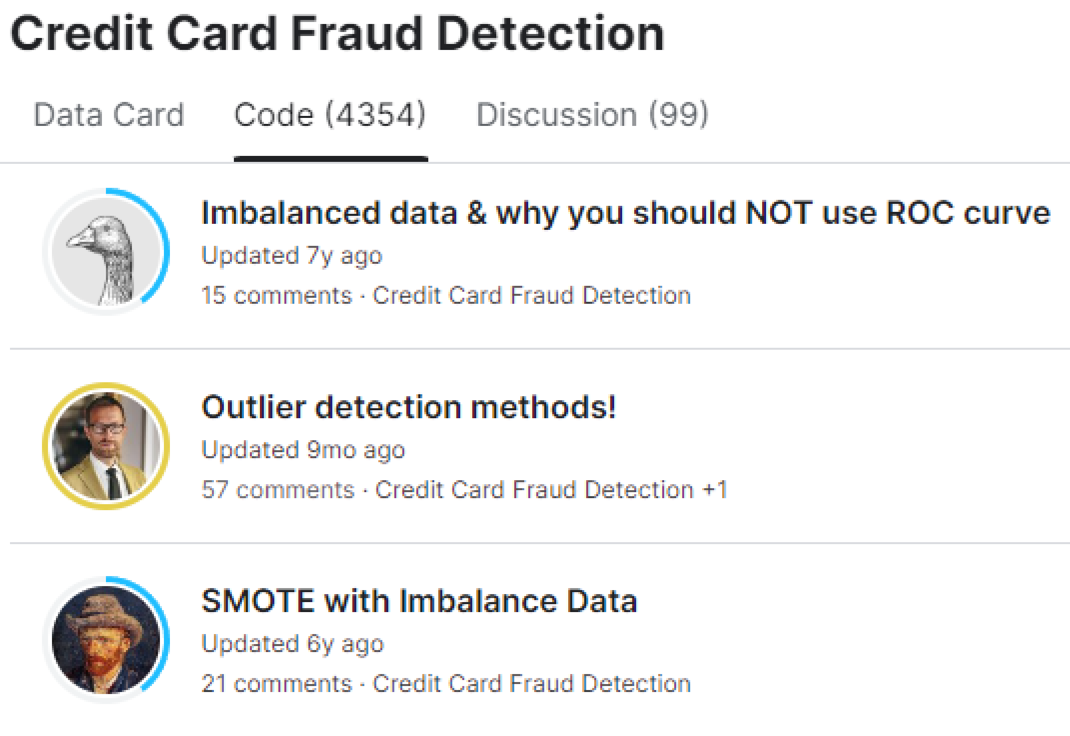
결론

- 지금까지 나온 여러 모델
- 성능 좋은 지 모르겠다
- 사람들이 연구를 안 하는 이유는 다 있다.
- Tabular Data Augmentation을 연구하는 것은 썩 합리적이지 않다.

Tabular Data는 이제 끝?

Anomaly detection





Anomaly detection

- Tabular Data의 Imbalanced Classification 다루는 새로운 방법
- 이상치 탐지
  - 정상 또는 이상치 구별
  - 이상치: 비정상적인 패턴, 데이터 포인트
- 이상치 탐지와 질병 예측…?
  - 정상: 다수(건강한 사람)
  - 이상치: 소수(환자)

연구 주제

- 이상치를 판단 요소로 추가
  - Improving Imbalanced Classification by Anomaly Detection
  - Outlier score
  - Type of data
- 이상치 탐지 모델을 활용한 질병 탐지
  - 고전적 이상치 탐지 모델
  - 딥러닝을 활용한 질병 탐지

논문 소개

PPSN 2020: Parallel Problem Solving from Nature

Jiawen Kong 1

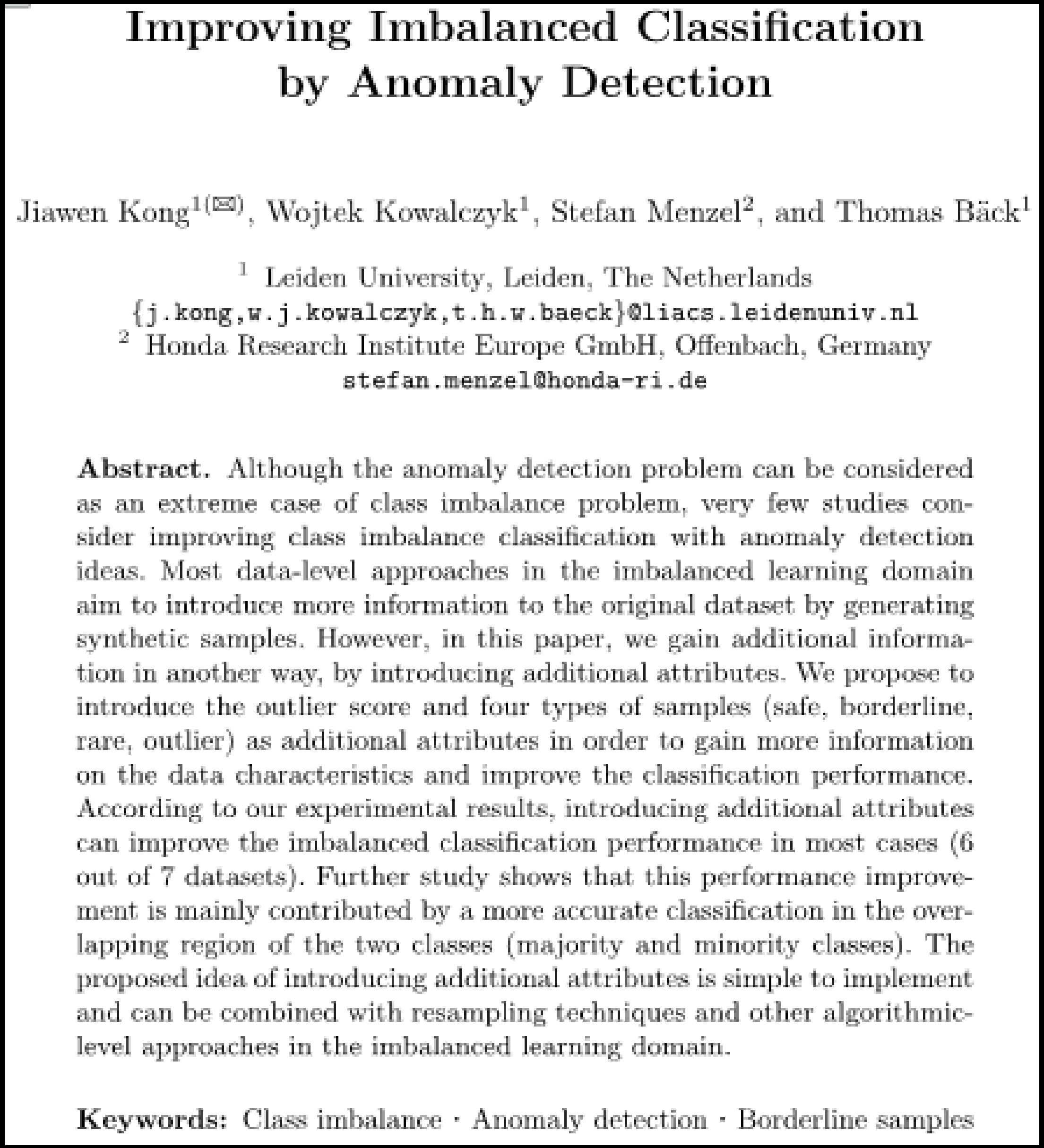
Wojtek Kowalczyk 1

Stefan Menzel 2

Thomas Bäck 1

[1] Leiden University, Leiden, The Netherlands

[2] Honda Research Institute Europe GmbH, Offenbach, Germany



Improving Imbalanced Classification by Anomaly Detection

Jiawen Kong[1(✉)], Wojtek Kowalczyk[1], Stefan Menzel[2], and Thomas Bäck[1]

[1] Leiden University, Leiden, The Netherlands
{j.kong,w.j.kowalczyk,t.h.w.baeck}@liacs.leidenuniv.nl
[2] Honda Research Institute Europe GmbH, Offenbach, Germany
stefan.menzel@honda-ri.de

Abstract. Although the anomaly detection problem can be considered as an extreme case of class imbalance problem, very few studies consider improving class imbalance classification with anomaly detection ideas. Most data-level approaches in the imbalanced learning domain aim to introduce more information to the original dataset by generating synthetic samples. However, in this paper, we gain additional information in another way, by introducing additional attributes. We propose to introduce the outlier score and four types of samples (safe, borderline, rare, outlier) as additional attributes in order to gain more information on the data characteristics and improve the classification performance. According to our experimental results, introducing additional attributes can improve the imbalanced classification performance in most cases (6 out of 7 datasets). Further study shows that this performance improvement is mainly contributed by a more accurate classification in the overlapping region of the two classes (majority and minority classes). The proposed idea of introducing additional attributes is simple to implement and can be combined with resampling techniques and other algorithmic-level approaches in the imbalanced learning domain.

Keywords: Class imbalance · Anomaly detection · Borderline samples

결과

Feature Importance

추가된 속성이 분류할
때 유의미하게 쓰임

2D chess dataset (2 original & 2 added attributes)

Score \ Attr / Add	org1	org2	add1	add2	—	—	—	—	—	—	—	—	—
NO	0.4636	0.5364	—	—	—	—	—	—	—	—	—	—	—
YES	0.0101	0.0097	0.8152	0.1650	—	—	—	—	—	—	—	—	—

glass1 dataset (9 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	org8	org9	add1	add2	—	—
NO	0.2063	0.0213	0.2354	0.1291	0.0302	0.0418	0.2634	0.0000	0.0726	—	—	—	—
YES	0.1770	0.0056	0.1527	0.1099	0.0000	0.0110	0.1892	0.0000	0.0056	0.2413	0.1077	—	—

ecoli4 dataset (7 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	add1	add2	—	—	—	—
NO	0.1093	0.0587	0.0000	0.0000	0.6591	0.1729	0.0000	—	—	—	—	—	—
YES	0.0000	0.0337	0.0000	0.0000	0.6119	0.0000	0.0808	0.1742	0.0994	—	—	—	—

vehicle1 dataset (18 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	org8	org9	org10	org11	org12	org13
NO	0.1304	0.0654	0.0892	0.0403	0.0563	0.0233	0.0028	0.0707	0.0000	0.0635	0.0172	0.0416	0.0438
YES	0.0248	0.0216	0.1317	0.0426	0.0227	0.0205	0.0179	0.0024	0.0000	0.0338	0.0260	0.1291	0.0828
Score \ Attr / Add	org14	org15	org16	org17	org18	add1	add2	—	—	—	—	—	—
NO	0.0862	0.0414	0.0498	0.0516	0.1265	—	—	—	—	—	—	—	—
YES	0.0146	0.0310	0.0325	0.0291	0.0471	0.2413	0.0485	—	—	—	—	—	—

yeast4 dataset (8 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	org8	add1	add2	—	—	—
NO	0.3301	0.2446	0.1839	0.0720	0.0106	0.0000	0.1233	0.0355	—	—	—	—	—
YES	0.0385	0.0297	0.0483	0.0116	0.0000	0.0000	0.0248	0.0053	0.7771	0.0646	—	—	—

wine quality dataset (11 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	org8	org9	org10	org11	add1	add2
NO	0.0466	0.1402	0.1215	0.1194	0.0806	0.0635	0.0428	0.1287	0.0483	0.0841	0.1244	—	—
YES	0.0000	0.0098	0.0000	0.0000	0.0000	0.0000	0.0263	0.0000	0.0000	0.0000	0.0000	0.9639	0.0000

page block dataset (10 original & 2 added attributes)

Score \ Attr / Add	org1	org2	org3	org4	org5	org6	org7	org8	org9	org10	add1	add2	—
NO	0.5452	0.0096	0.0117	0.1899	0.0530	0.0285	0.0983	0.0382	0.0122	0.0134	—	—	—
YES	0.5282	0.0006	0.0036	0.1745	0.0205	0.0223	0.0833	0.0129	0.0007	0.0082	0.1288	0.0164	—

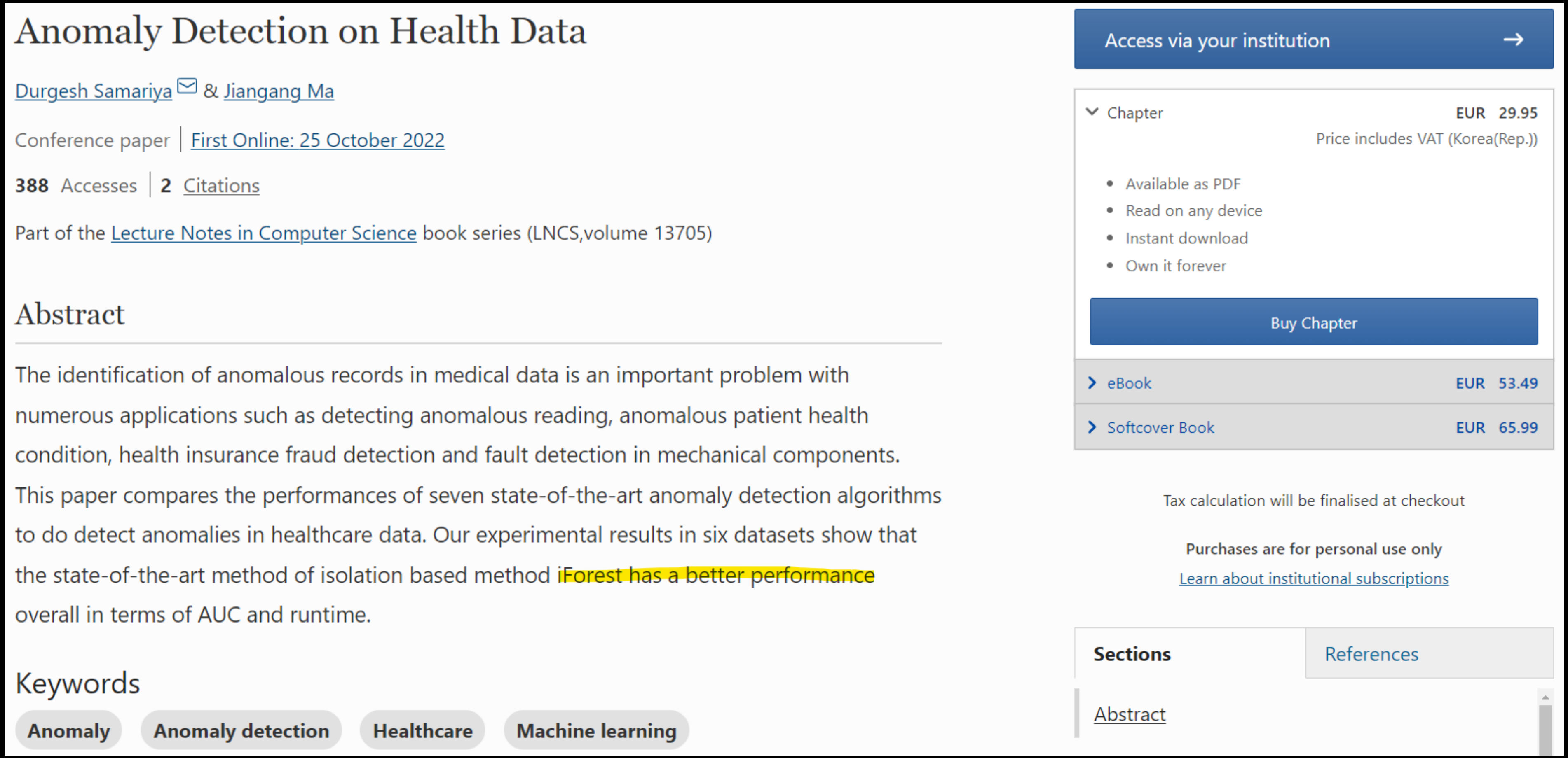
전체 데이터를 기준으로 했을 때의 문제점

- 과적합
  - 학습한 모델에 새로운 데이터가 들어오면 잘 판단할 수 있을까?
- 의미있는 실험인가?
  - 애초에 Outliar score, type을 제공하고 학습
  - 정답을 보여주는 것과 다름 없지 않나?

이상치 탐지 모델을 활용한 질병 탐지

- 전통적인 Outlier detection method
  - Tukey's IQR method
  - Standard deviation method
  - Z-score method
  - Isolation Forest
  - DBSCAN
- 딥러닝 Outlier detection method
  - Anomaly transformer 등등

향후 계획 - iForest



Anomaly Detection on Health Data

Durgesh Samariya & Jiangang Ma

Conference paper | First Online: 25 October 2022

Part of the Lecture Notes in Computer Science book series (LNCS,volume 13705)

Access via your institution →

Chapter EUR 29.95

Price includes VAT (Korea(Rep.))

- Available as PDF
- Read on any device
- Instant download
- Own it forever

Buy Chapter

eBook EUR 53.49

Softcover Book EUR 65.99

Tax calculation will be finalised at checkout

Purchases are for personal use only

Learn about institutional subscriptions

Abstract

The identification of anomalous records in medical data is an important problem with numerous applications such as detecting anomalous reading, anomalous patient health condition, health insurance fraud detection and fault detection in mechanical components. This paper compares the performances of seven state-of-the-art anomaly detection algorithms to do detect anomalies in healthcare data. Our experimental results in six datasets show that the state-of-the-art method of isolation based method iForest has a better performance overall in terms of AUC and runtime.

Keywords

Anomaly | Anomaly detection | Healthcare | Machine learning

Sections | References

Abstract