Large-Scale IP Usage Identification via Deep Ensemble Learning (Student Abstract)

Zhiyuan Wang[1], Fan Zhou[1*], Kunpeng Zhang [2], Yong Wang [3]

[1] University of Electronic Science and Technology of China
[2] University of Maryland, College park
[3] Zhengzhou Aiwen Computer Technology Co., Ltd.
zhy.wangcs@gmail.com, fan.zhou@uestc.edu.cn, kpzhang@umd,edu, wangyong@ipplus360.com

Abstract

Understanding users' behavior via IP addresses is essential towards numerous practical IP-based applications such as online content delivery, fraud prevention, and many others. Among which profiling IP address has been extensively studied, such as IP geolocation and anomaly detection. However, less is known about the scenario of an IP address, e.g., dedicated enterprise network or home broadband. In this work, we initiate the first attempt to address a large-scale IP scenario prediction problem. Specifically, we collect IP scenario data from four regions and propose a novel deep ensemble learning-based model to learn IP assignment rules and complex feature interactions. Extensive experiments support that our method can make accurate IP scenario identification and generalize from data in one region to another.

Introduction

An IP address is a unique identity assigned to each device connected to the Internet, which serves as personally identifiable information and location addressing. Actively probing into the details of IP addresses and distilling the knowledge behind through analyzing risk behaviors of IPs in a variety of dimensions can be broadly considered as IP address profiling (IAP). Among which IP geolocation is an important dimension, which refers to mapping the IP addresses into geographical locations and is an essential step towards a range of downstream applications (Wang et al. 2011).

In this work, we study a new problem in IAP – IP usage scenario identification (IPusi) – which aims to predict the role of IP address owners (e.g., data centers, home broadband) by analyzing the network attributes and behaviors associated with the IPs. This problem becomes important in many network-based applications. For example, accurate IPusi can verify legality and authenticity while helping firms to effectively intercept fraud risks and network attacks that could result in losses. Another case is by analyzing the application scenario of IP addresses, advertising companies and demand-side platforms can filter out bot IPs to reduce the online advertising cost, improve advertisement delivery effectiveness, and increase the return on investment.

To the best of our knowledge, this is among the first work towards proposing an effective and efficient method to address the IPusi problem. Our approach considers IP blocks each of which consists of many IP addresses. Thus, more statistic features can be leveraged and analyzed for IPusi. Besides, we propose a novel neural network to capture various explicit/implicit features and interactions in rich network measurements. Specifically, we employ differentiable boosted decision trees to learn interpretable feature transformations and provide model differentiability in feature splitting and decision tree routing. Meanwhile, our method stacks multiple layers of ensemble trees in a continuous way, enabling the learning of deep decision rules.

We conduct extensive evaluations on our collected real-world datasets. Experimental results demonstrate that our method achieves the best performance compared to baselines on IPusi problem.

Data and Problem Definition

Data. After scrutinizing the IP address assignment, we find that continuous IPs are usually used in the same scenario. Therefore, we consider IP blocks rather than individual IP when predicting their usage scenarios. Besides, features of an IP block are more stable and intuitively understood with good interpretability compared to those of a single IP address. Precisely, in our processing, segmenting IPs into blocks is based on the smallest IP subnetwork division in the WHOIS database – the registration information of the global IP allocation agency, and the BGP information – the routing table information of the worldwide network autonomous system domain data exchange. If the number of IPs in a subnet is more than 256, remaining IPs will be grouped into a new block.

Problem Definition. In this work, we study the IP usage scenario identification (IPusi) problem, which aims at learning a data-driven model that can classify each IP block into one of the four typical IP usage scenarios, i.e., Home Broadband (HB), Dedicated Enterprises Network (DEN), Cellular Network (CN), and Data Center (DC).

Proposed Approach

Model. In our IPusi problem, there exists various numerical and categorical features. The most effective methods for

[*] Corresponding author

Region	IP Block	IP address	Area (km^2)	Population (M)
Sichuan	30,029	6,999,780	481,400	83.41
Shandong	67,443	12,731,730	153,800	100.47
Chongqing	18,719	3,304,308	82,300	31.02

Table 1: Descriptive statistics of datasets.

tabular data feature learning and classification are tree-based boosting models which also follow a rules-based decision-making process that may be consistent with real IP scenario assignments and provide accurate and explainable predictions.

In this work, we improve oblivious decision tree (ODT) by making it differentiable via stochastic routing theory (Kontschieder et al. 2015) and treat it as the basic structure for its strong generalizability. Different from the decision nodes in traditional decision trees simply routing by a binary number, the routing directions in our design are decided by a random variable, which provides feasibility for global optimization, whereby we can update the parameters of decision and leaf nodes by gradient descent.

Further, we design a deep continuous architecture in term of differentiable ODT for more expressive and higher-order representation in a parameter-efficient way. There are M ODTs in each layer whose outputs are composed of the concatenation of the prediction of all trees, i.e., $H^k = \{H_1^k, ..., H_M^k\}$, where H_m^k denotes prediction of the m-th tree in the k-th layer. For the straightforward propagation between ensemble tree layers, we introduce neural ODE (Chen et al. 2018) to build continuous transformation instead of the discrete one (Popov, Morozov, and Babenko 2020), i.e.:

$$H^k = H^{k-1} + \int_{k-1}^{k} \text{ODTSR}(k', H^{k'} + x; \Theta)dk', \quad (1)$$

where ODTSR is our differentiable ODT and Θ denotes learned parameters. We employ efficient RK-4 to solve this integral since it has higher precision than simple Euler method. The final prediction of the deep model is the average of all layers.

Optimization. We use an efficient optimizer (Ma and Yarats 2019) and employ cross-entropy as optimization objective in our IPusi problem. The loss function is defined as:

$$\mathcal{L} = -\ln \sum Q(x) \circ y, \quad (2)$$

where $\circ$ denotes the Hadamard product, B denotes the set of a mini-batch.

Experiments

Dataset. We evaluate our proposed method using the IP data collected from four regions: Shandong province, Sichuan province, Chongqing city from China, and state of Illinois from the U.S. Table 1 records the statistics of them.

Baselines. We evaluate our model against the following baselines that can be grouped into three categories:

- Machine learning methods: Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA).
- Ensemble learning methods: CatBoost and XgBoost.
- Deep learning methods: TabNet (Arik and Pfister 2021) and AutoInt (Arik and Pfister 2021).

Model	**Sichuan**	**Shandong**	**Chongqing**
SVM	0.9705	0.9916	0.9819
LDA	0.9553	0.9810	0.9717
CatBoost	0.9226	0.9805	0.9464
XgBoost	0.9732	0.9947	0.9913
TabNet	0.9623	0.9878	0.9757
AutoInt	0.9591	0.9926	0.9820
Ours	**0.9822**	**0.9954**	**0.9922**

Table 2: AUC comparison on the IPusi problem.

Evaluation Metrics. We use the area under the ROC curve (AUC), which is computed based on the relative ranking of IP block's prediction probability and not impacted by any simple scaling of prediction.

Performance Comparison. The overall evaluations of all methods are reported in Table 2, from which we can observe that our model achieves the best performance in IPusi problem across all datasets. Compared with baselines, our model based on a differentiable divide-and-conquer strategy is similar to real IP scenario assignments and learns expressive information flow in deep neural layers, and, as a result, rich feature interactions for IP classification.

Conclusion

In this work, we initiated the first attempt to study the problem of IP usage scenario identification, a new paradigm of IP address profiling that can benefit various downstream applications. We proposed a novel data processing method and deep ensemble learning approach based on continuous differentiable tree layers for this problem, which was demonstrated effective by extensive experiments.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62072077 and No. 62176043).

References

Arik, S. Ö.; and Pfister, T. 2021. TabNet: Attentive Interpretable Tabular Learning. In *AAAI*, 6679–6687.

Chen, R. T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural Ordinary Differential Equations. In *NeurIPS*.

Kontschieder, P.; Fiterau, M.; Criminisi, A.; and Bulo, S. R. 2015. Deep Neural Decision Forests. In *ICCV*.

Ma, J.; and Yarats, D. 2019. Quasi-hyperbolic momentum and adam for deep learning. In *ICLR*.

Popov, S.; Morozov, S.; and Babenko, A. 2020. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. In *ICLR*.

Wang, Y.; Burgener, D.; Flores, M.; Kuzmanovic, A.; and Huang, C. 2011. Towards Street-Level Client-Independent IP Geolocation. In *NSDI*, volume 11, 27–27.