Machine Learning Techniques for Diminution of Network Assaults to Improve the Performance Overwhelmed Network Traffic

D. Swetha, MD. Jaffar Sadiq, Munipalle Madhavi Latha,
Mahesh K & Ch. Lavanya Kumari

Assistant Professor, CSE Dept, Sphoorthy Engineering College
Associate Professor, IT Dept, Sreenidhi Institute of Science and Technology
Assistant Professor, CSE Department
Assistant Professor, CSE Dept, CMR Technical Campus
Assistant Professor, CSE Dept, Guru Nanak Institutions Technical Campus

Abstract: The developing multifaceted nature and size of Internet frameworks combined with the improved capacities of Machine Learning (ML) techniques lately have persuaded specialists and professionals to progressively depend on these techniques for data driven plan and investigation of wide scope of issues in organize frameworks in networking issues like identifying system assaults, performing asset the executives, or improving nature of administration. In this study, we audit an enormous number of conspicuous exploration papers that apply ML techniques to structure, investigation or assessment of arrange frameworks. To this end, we partition these examinations into the following six gatherings dependent on the region of system frameworks that they target: 1) Domain name framework, 2) Application ID, 3) QoS, 4) Cloud administrations, 5) Network security, and 6) Traffic expectation. Inside each gathering, we inspect the sort of ML strategies just as information datasets that are utilized by person considers, portray how they address different difficulties, and sum up their key discoveries. Specifically, we investigate how space information in systems administration can educate various perspectives of applying ML procedures, for example, issue plan, include building, highlight determination, and organization. We sum up delegate normal practices in the systems administration network alongside various residual difficulties and examination hole.

Keywords: Machine learning, networks, security, supervised learning, unsupervised learning.

I. INTRODUCTION

During the previous twenty years, we have seen a critical increment in the scale and heterogeneity of system substances, hidden applications, and conventions. Both plan furthermore, investigation of these conventions progressively request catching also, understanding examples in enormous scope multi-dimensional datasets. The size of information and a more significant level of the theoretical in the stream level datasets contrast with port-and payload-based strategies set up for ML techniques. As of late, the commonness of AI (ML) strategies with an appropriate fit for the referenced difficulties settles on them a sensible decision and prompted developing organizations of these techniques in structure and assessment of system frameworks.

This study analyzes an assortment of ongoing investigations that influence different methods specifically ML strategies for the plan also, portrayal of system frameworks. Principally, we gathering these investigations by their objective issue in systems administration that serves as a typical setting and foundation for them. Inside each gathering, we further classify dependent on more explicit subjects whenever the situation allows. For each group of studies, we talk about the objective of each investigation and related difficulties, its information dataset, the projecting of their objective issue into ML strategy and related difficulties and issues, and their discoveries. Here, the emphasis is generally

on the plan of the methods for information examination and related difficulties, openings specifically how space information from systems administration has been utilized to alter any strategy. We select the generally welcomed, profoundly refered to, and peer-investigated papers in top levels organizing settings for example, IMC, SIGCOMM, TON, and InfoComm in the last ten a long time.

Rather than concentrating on subtleties of execution, we audit the exploration questions, challenges that are tended to, kind of the ML strategies and how they have been utilized (counting highlights and approaches), lastly, the primary discoveries. That will prompt an attention on how ML strategies are utilized to reply a particular inquiry while overlooking less significant subtleties.

II.LITERATURE SURVEY

Novel machine learning techniques such as convolution neural networks or deep learning allow to tackle real world problems. At the same time local computer networks and the Internet face threats unseen before. With emerging machine learning paradigms we can detect and counteract them. We can also approach identifying software defects and hardware faults in networks as well as seasonality of the traffic. The Session will provide an opportunity for researchers working with machine learning to showcase their developments in the field computer network traffic analysis and will be a forum to bring these two areas of research together and elicit fruitful discussions.

Over the past decades, Artificial intelligence (AI) stream has become the broad and exciting field in computer science as it prepare the machines to perform the tasks that human being may do. and it aims to train the computers to solve real world problems with the maximum success rate. As perceiving scientific growth and advancement in technology AI systems are now capable to learn and improve through past experiences without explicitly assistance code if they exposed to new data. Eventually it leads to technology of Machine learning (ML) which uses learning algorithms to learn from the data available [1]. Machine Learning uses data mining techniques to extract the information from the huge size datasets. ML and Data
Mining techniques explore data from end to end to find the hidden patterns inside dataset [2]. Machine Learning and data mining algorithms has been deployed in various fields such as Computer networking, travel and tourism industry, finance forecasting, telecommunication industry and electric load forecasting and so on.

Methods used in Machine Learning

Over past years an enormous number of ML algorithms was introduced. Only some of them were able to solve the problem so they replaced by another one [3]. There are three ML algorithms for example unsupervised learning and reinforcement learning, supervised learning, which are

Supervised learning: It consists of a given set of input variables (training data) which are pre labeled and target data [5]. Using the input variables it generate a mapping function to map inputs to required outputs. Parameter adjustment procedure continues until the system acquired a suitable accuracy extent regarding the teaching data.

Unsupervised learning : In this algorithm we only have training data rather a outcome data. That input data is not previously labeled. It is used in classifiers by recognizing existing patterns or cluster in the input datasets [4].

Reinforcement learning : Applying this algorithm machine is trained to map action to a specific decision hence the reward or feedback Signals are generated. The machine trained itself to find the most rewarding actions by reward and punishment using past experience
.
Authors Vahidov et al. (2008) compares the methods of predicting demand in the last of a supply chain, the naive forecasting and linear regression and trend moving average with advanced machine

learning methods such as neural networks and support vector machines, recurrent neural networks finds that recurrent neural networks and support vector machines show the best performance.

Wang (2007) describes the machine learning method with genetic algorithm (GA)-SVR with real value GAs, . The experimental findings investigates this , SVR outshines the ARIMA models and BPNN regarding the base the normalized mean square error and mean absolute percentage error .

Turksen et al. (2012) presents next-day stock price prediction model which is based on a four layer fuzzy multi agent system (FMAS) structure. This artificial intelligence model used the coordination of intelligent agents for this task. Authors investigates that FMAS is a suitable tool for stock price prediction problems as it outperforms all previous methods.

Shahrabi et al. (2013) proposed a method for estimating tourism demand which is a new combined intelligent model i.e. Modular Genetic-Fuzzy Forecasting System using a genetic fuzzy expert systems and finds that accuracy of predicting power of MGFFS is better than approaches like Classical Time Series models , so it is suitable estimating tool in tourism demand prediction problems.

Authors Fouilloy et al. (2018) suggested a statistical method employing machine learning model and to analyze and applied it to solar irradiation prediction working hourly. This methodology used the high, low and medium meteorological variability like Ajacio, Odeillo , Tilos . They compared model with auto regressive moving average and multi-layer preceptor .

Makridakis et al. (2018) presents Machine Learning methods to statistical time series forecasting and compared the correctness of those methods with the correctness of conventional statistical methods and found that the first one is better and out top using the both measures of accuracy. They provide the reason for the accuracy of learning models is less that of statistical models and suggested some other achievable ways.

III.RELATED WORK

With the prosperous development of the Internet, networking research has attracted a lot of attention in the past several decades both in academia and industry. Researchers and network operators can face various types of networks (e.g., wired or wireless) and applications (e.g., network security and live streaming [1]). Each network application also has its own features and performance requirements, which may change dynamically with time and space. Because of the diversity and complexity of networks, specific algorithms are often built for different network scenarios based on the network characteristics and user demands. Developing efficient algorithms and systems to deal with complex problems in different network scenarios is a challenging task.

Recently, machine learning (ML) techniques have made breakthroughs in a variety of application areas, such as bioinformatics, speech recognition and computer vision. Machine learning tries to construct algorithms and models that can learn to make decisions directly from data without following pre-defined rules. Existing machine learning algorithms generally fall into three categories:

supervised learning (SL), unsupervised learning (USL) and reinforcement learning (RL). More specifically, SL algorithms learn to conduct classification or regression tasks from labeled data, while USL algorithms focus on classifying the sample sets into different groups (i.e., clusters) with unlabeled data. In RL algorithms, agents learn to find the best action series to maximize the cumulated reward (i.e., objective function) by interacting with the environment. The latest breakthroughs, including deep learning (DL), transfer learning and generative adversarial networks (GAN), also provide potential research and application directions in an unimaginable fashion.

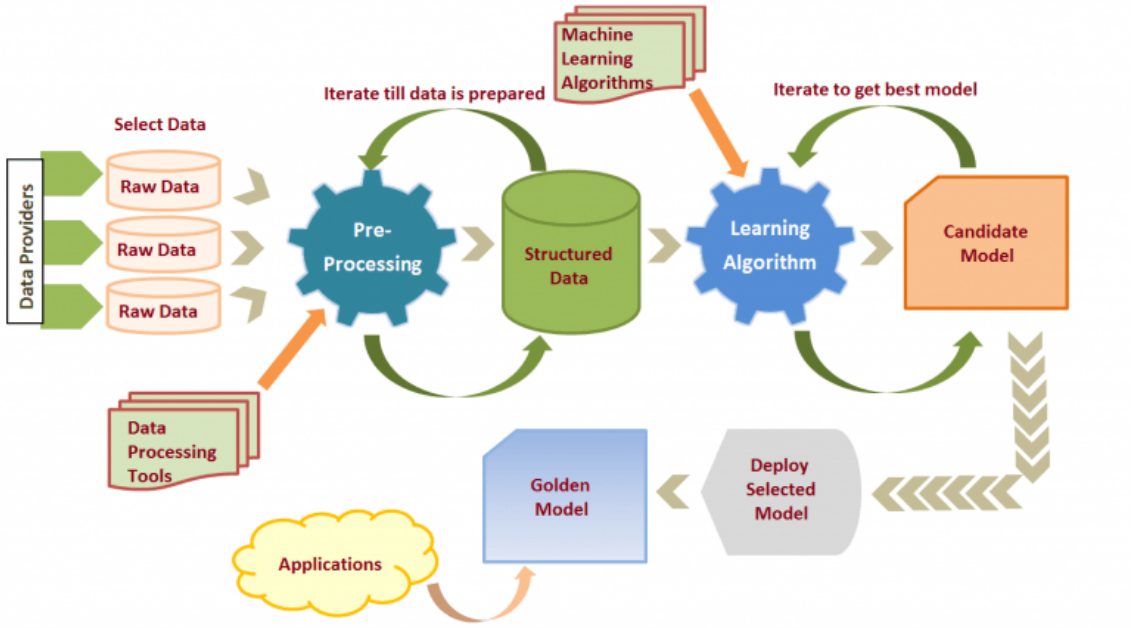

Fig.1. Process of machine learning

Specifically, machine learning for networking (MLN) is suitable and efficient for the following reasons. First, as the best known capabilities of ML, classification and prediction play basic but important roles in network problems such as intrusion detection and performance prediction [1]. In addition, machine learning can also help decision making, which will facilitate network scheduling [2] and parameter adaptation [3, 4], according to the current states of the environment. Second, many network problems need to interact with complicated system environments. It is not easy to build accurate or analytic models to represent complex system behaviors such as load changing patterns of CDN [5] and throughput characteristics [1].

Machine learning can provide an estimated model of these systems with acceptable accuracy. Finally, each network scenario may have different characteristic (e.g., traffic patterns and network states) and researchers often need to solve the problem for each scenario independently. Machine learning may provide new possibilities to construct the generalized model via a uniform training method [3, 4]. Among efforts in MLN, deep learning has also been investigated and applied to provide end-to-end solutions. The latest work in [6] conducts a comprehensive survey on previous efforts that apply deep learning technology in network related areas.

In this article, we investigate how machine learning technology can benefit network design and optimization. Specifically, we summarize the typical workflows and requirements for applying machine learning techniques in the network domain, which could provide a basic but practical guideline for researchers to have a quick start in the area of MLN. Then we provide a selective survey of the important networking advances with the support of machine learning technology, most of which have been published in the last three years. We group these advances into several typical networking fields and explain how these prior efforts perform at each step of the MLN workflow. Then we discuss the opportunities of this emerging inter-discipline area. We hope our studies can serve as a guide for potential future research directions.

We categorize the prior studies into six sub-domains as follows:
1) Domain Name System (DNS): We review studies that examine the ML-based solutions for challenges in performance and security of DNS. Namely detection of one-time-used domains to be

excluded from caching [1] fluctuation of open DNS resolvers [2], and detection of malicious domains using decision tree models [3], and statistical techniques [4].

2) Network Monitoring and Controlling: We grouped studies in this section into two main sub-sections: protocol and application identification, and network controlling. Studies in this section, tackle various problems, such as the accuracy of labeled data [5], the role of different elements when applying ML techniques such as feature set or hyper-parameters [6], considering protocols as a language [7]. Detection of the zero-day (unknown) applications [8], real-time protocol identification [8] and considering only the packet length feature, as well as the combined use of clustering and classification methods for traffic labeling, are among the novel studies reviewed in this section.

3) Quality of Experience (QoE): The studies in this section mainly used the off-the-shelf ML techniques (mostly decision trees) leveraging the network traffic features for prediction and improving the user engagement and satisfaction in a range of applications such as video streaming Skype and web. However, not all studies benefit directly from the ML techniques. Some studies get inspired by the ML techniques , some compare the results with ML techniques as a baseline , and some use them as an enabler to capture validated dataset.

4) Security: While most of the studies we review in all other sections have security implications, we group a number of them in this section that are solely addressing security
concerns. User profiling, anomaly detection, and security or IoT devices are the common themes of these studies. Multiple ML methods including SVM, MLP, CNN, and DT are applied on user's (encrypted) web browsing data to group them with other similar users. Such grouping (user profiling) can be used for marketing and user segmentation purposes [1]. Anomaly detection methods have a wide range of flavors, such as detection of anomaly without prior context [2], statistical methods to detect anomalies [3] scaleable anomaly detection in mobile networks [4], detection focusing on temporary related traffic [5], anomaly detection based on server profiling [6], anomaly detection in the scale of WAN [7].

IoT related privacy challenges are summarized by Wang et al. [8]. This subsection follows by traffic analysis of IoT devices and related privacy and security concerns , and detection DDoS attacks in IoT devices and behavior analysis of residential areas based on IoT traffics.

5) Cloud and Cluster Computing: In this section we review HUYGENS system that explore the software clock synchronization in the network by predicting oneway propagation time, Ernest framework that predicts the performance of large scale parallel analytic for efficient scheduling and resource management, and finally Gaia framework that addresses the efficient usage of scarce WAN bandwidth in geo-distributed ML systems.

6) Network Traffic Prediction: In this section, we review studies in two groups: we review studies that leverage time-series analysis on network traffic data to forecast the traffic in the next time slot(s) along with studies with new ideas to obtain improved results. In the first group, studies focus on traffic prediction on mobile networks given its limited resources and facing increasing demand using Neural network methods and unsupervised learning . In addition, capturing spatio-temporal dynamics, using network traffic matrix , considering variations of LSTM , and predicting TCP output are also included in this part. In the second group, improving LSTM by employing random connectivity trick , decomposing time-series , analyzing the prediction uncertainty factors , transforming the 1D traffic to 2D matrices and then applying the CNN , encoding traffic to image , and utilizing statistical methods are covered.

IV.NETWORK FLEXIBILITY USING RESOURSE MANAGMENT

Efficient resource management and network adaption are the keys to improving network system performance. Some example issues to address are traffic scheduling, routing [2], and TCP congestion control [3, 4]. All these issues can be formulated as a decision-making problem . However, it is challenging to solve these problems with a rulebased heuristic algorithm due to the complexity of diverse system environments, noisy inputs and difficulty in optimizing the tail performance . Specifically, arbitrary parameter assignments based on experiences and action taken following predetermined rules often result in a scheduling algorithm that is understood by people but far from optimal.

Deep learning is a promising solution due to its ability to characterize the inherent relationships between the inputs and outputs of network systems without human involvement. In order to meet the requirements of changing network environments, previous efforts in [2] design a traffic control system with the support of deep learning techniques. Reconsidering backbone router architectures and strategies, it takes the traffic pattern in each router as input and outputs the next nodes in the routing path with Deep Belief Architectures. These advancements unleash the potential of the DL-based strategy in network routing and scheduling. Harnessing the powerful representational ability of deep neural networks, deep reinforcement learning achieves great results in many AI problems.

VI.THROUGHPUT OF NETWORK

Performance prediction can guide decision making. Some example applications are video QoE prediction, CDN location selection, best wireless channel selection, and performance extrapolation under different configurations. Machine learning is a natural approach to predict system states for better decision making. Typically, there are two general prediction scenarios. First, the system owner has the ability to get various and enough historical data, but it is non-trivial to build a complex prediction model and update it in real time, which requires a new approach exploiting domain-specific knowledge to simplify the problem (e.g., CFA for video QoE optimization). In prior work, CS2P [1] wants to improve video bitrate selection with accurate prediction. It finds that sessions with similar key features may have more related throughput behavior from data analysis. CS2P learns to cluster similar sessions offline and trains different Hidden-Markov Models for each cluster to predict the corresponding throughput given the current session information.

VII.OPEN DATA SETS FOR NETWORK DOMAINS

Collecting a large amount of high quality data that contain both network profiles and performance metrics is one of the most critical issues for MLN. However, acquiring enough labeled data is still expensive and labor intensive even in today's machine learning community. For many reasons, it is not easy for researchers to acquire enough real trace data even if there are many existing open datasets in the networking domain. This reality drives us to learn from the machine learning community to put much more effort into constructing open datasets like ImageNet. With unified open datasets, performance benchmarks are an inevitable outcome to provide a standard platform for researchers to compare their new algorithms or architectures with state-of-theart ones. This can reduce the unrepresentative repeated experiments and have a positive effect on academic loyalty. In addition, it has been proved in the machine learning domain that learning with a simulator rather than in a real environment is more effective and with lower cost in RL scenarios .

VIII.SELF DESIGNED NEWTWORK PROTOCOLS AND ARCHITECTURE DESIGN

With a deeper understanding of the network, researchers gradually find that the existing network has many limitations. The network system is totally created by human beings. The current network components are likely to be added based on people's understanding at a time instant rather than a

paragon of engineering. There is still enough room for us to improve network performance and efficiency by redesigning the network protocol and architecture. It is still quite difficult to design a protocol or architecture automatically today. However, the machine learning community has made some of the simplest attempts in this direction and has achieved some amazing results, such as letting agents communicate with others to finish a task cooperatively. Other new achievements, e.g., GAN, have also shown that the machine learning model has the ability to generate elements existing in the real world and create strategies people do not discover (e.g., AlphaGo). However, these generated results are still far from the possibility of protocol design. There is great potential and the possibility to create new feasible network components without human involvement, which may refresh human's understanding of network systems and propose some currently unacceptable destructive-reconstruction frameworks.

IX.MONITORING OF NETWORK SYSTEMS

Network behaviour is quite complex due to the end-to-end network design principle, which generates various protocols that have simple actions in the end system but causes nontrivial in-network behaviour. From this perspective, it is not easy to figure out what factors can directly affect a certain network metric and can be simplified during an algorithm design process even in a mature network research domain like TCP congestion control. However, with the help of machine learning methods, people can analyze the output of learning algorithms through a posterior approach to find useful insights for us to understand how the network behaves and how to design a high performance algorithm. For a detailed explanation, DeepRM, a resource management framework, is a good example. To understand why DeepRM performs better, the authors find that DeepRM is not work-conserving but decides to reserve room for those yet-to-arrive small jobs, which eventually contributes to reducing job waiting time. For other evidence, refer to CFA and Remy and their following works, which provide insights for key influence factors in video QoE optimization and TCP congestion control, respectively.

XI.NEED OF MACHINE LEARNING TECHNIQUES

When applying machine learning into networking fields, due to specific requirements of network systems and practical implementation problems, some inherent limitations and other emerging problems of machine learning can be pushed forward to a new understanding stage with the joint efforts of two research communities. Typically, there are several problems that are expected to be resolved. First, the robustness of machine learning algorithms is a key challenge for applications (e.g., self-driving cars and network operation) in real-world environments where learning errors could lead to high costs. The networking situation often requires hard constraints on the algorithm output and the worst performance guarantee. Second, a model with high generalization ability that can adapt in the high-variance and dynamic traffic circumstances is needed, since it is unacceptable to retrain the model every time the characteristics of network traffic change.

XII.CONCLUSION

Due to the heterogeneity of networking systems, it is imperative to embrace machine learning techniques in the networking domain for potential breakthroughs. However, it is not easy for networking researchers to take it into practice due to the lack of machine learning related experiences and insufficient directions. In this article, we present a basic workflow to provide researchers with a practical guideline to explore new machine learning paradigms for future networking research. For a deeper comprehension, we summarize the latest advances in machine learning for networking, which covers multiple important network techniques, including measurement, prediction and scheduling. Moreover, numerous issues are still open and we shed light on the opportunities that need further research effort from both the networking and machine learning perspectives.

REFERENCES

1.Machine Learning for Networking: Workflow, Advances and Opportunities, Mowei Wang, Yong Cui, Xin Wang, Shihan Xiao, and Junchen Jiang

2.The Applications of Machine Learning Techniques in Networking ,Area Exam, Soheil Jamshidi, Feb. 2019

3.Machine Learning for Computer Networks at the 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA 2017) Gdynia, Poland, July 3-5, 2017, Conference website: http://inista.org/

4.Ravindra Changala ,"Development of Predictive Models for Medical Domains to Predict Chronic Diseases (Diabetes) using Machine Learning Algorithms and classification Techniques" ,ARPN Journal of Engineering and Applied Sciences, Volume 14, Issue 6, Pages 1202-1212 in 2019.Asian Research Publishing Network (ARPN)

5.S. Hao and H. Wang, “Exploring domain name based features on the effectiveness of dns caching,” ACM SIGCOMM Computer Communication Review, vol. 47, no. 1, pp. 36–42, 2017.

6. B. Mao et al., “Routing or Computing? The Paradigm Shift Towards Intelligent Computer Network Packet Transmission Based on Deep Learning,” IEEE Trans. Computers, 2017.

7.Z. Fadlullah et al., “State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow’s Intelligent Network Traffic Control Systems,” IEEE Commun. Surveys & Tutorials, 2017.

8.Ravindra Changala, "A Survey1 on Clustering Techniques to Improve Energy Efficient Routing in Wireless Sensor Networks" International Journal of Applied Engineering Research, Volume 10, Issue 58, Pages 1-5.