Survey towards Automatic Performance Optimization of Networks Using Machine Learning

Kaustubh Chaudhuri
Electrical and Computer Engineering Department
California State University, Chico
Chico, United States of America
kchaudhuri@mail.csuchico.edu

***Abstract*—This is a survey that discusses how Machine Learning techniques can be used to optimize the performance of the network over the existing prior models. Due to the continuous change in the network, Machine Learning can enable the network devices to configure with the most optimal settings automatically. By finding the relationship between the network topologies, configuration parameters and protocols, the authors are able to maximize the performance of the network. It is observed that the network performance is increased by at least 50% for a particular application based on TCP (Transmission Control Protocol) [1].**

Keywords—TCP; Machine Learning; Causal Bayesian Network; Support Vector Regression

I. MOTIVATION

With the evolution of Artificial Intelligence, Machine Learning(ML) is a fascinating and potential area for next generation solutions. Machine learning was born with only the ability to recognize patterns but now it is capable of achieving much greater goals and can be implemented in almost every domain. Machine learning is simply the science of enabling a system to learn from a given set of data. With this feature there are new possibilities of research in numerous fields which deals with information. Incorporating ML in networks can forward communication networks to an unprecedented level of growth.

II. INTRODUCTION

Every day, the size of the networks are increasing tremendously with large amounts of data being transmitted and received, thus affecting the performance of the networks. A method for optimizing the performance of the network is by setting the perfect configuration for a particular network topology, this influences the way network protocols work. An efficient way to handle this situation is to use structured protocols and models which are optimized mathematically. But this approach fails when it comes to scalability, and predicting changes in the network topology. There are several ways to counter this problem by using different models, but these have not been able to consider all possible configuration parameters due to ever changing networks hence being redundant for many kinds of topologies.

To overcome these disadvantages, the best alternative is to use Machine Learning. ML with the help of statistical data can accurately predict the performance of the networks. Firstly I will discuss about the technique specified in *'Towards Automatic Performance Optimization of Networks Using Machine Learning'*, by the authors; Fabien Geyer *et al* [1]. They use statistical network information along with a Machine Learning algorithm, Causal Bayesian Network to generate accurate predictions. Which are then used in an optimization algorithm; the random-restart hill climbing, to optimize the network topologies. For testing this method the authors implement an application by placing virtual machines in a given topology for maximizing the average bandwidth of TCP flows. By doing so they obtain an overall average increase in the performance of TCP flows by at least 50%. The challenge addressed is faced by all data centers and cloud architectures.

In the second part of the dissertation I will discuss about another state-of-the-art Machine Learning technique which is the Support Vector Regression, based on *"A Machine Learning Approach to TCP Throughput Prediction"* [2] [3], to predict the throughput for small and large transfers. It is a resourceful, easy to implement and a light weight model for predicting TCP throughput. It is 1.6 times better than the prior non ML methods. It handles level shifts in the network with greater agility and with lesser traffic comparatively. A tool; *PathPerf* is used for testing various wide area paths. This ML technique is extremely fast and can make predictions in 0.33 seconds of measurements, which prove to be very useful for wireless applications.

III. COMMUNICATION NETWORK BACKGROUND

A. Network Protocols

In short network protocols are a set of rules defined for various kinds of networks. They are formatting standards of how data should be packaged into a message and fetched from a message. They may include handshaking or acknowledgement messages as well. One of the most common protocol is the packet switching technique in which a data is divided into small data packets before transmission and joined together at the receiving end. There are various upper level protocols such as TCP, HTTP and UDP protocols. Out of which TCP based network has been used for the applications being surveyed.

B. Configuration parameters

These are specific parameters which are unique to different kinds of network protocols. Since the authors use statistical data for Machine Learning algorithms, our data mainly

consists of configuration parameters based on TCP. With the help of these parameters the authors can build a feature set for classification and training purposes in Machine Learning domain. They are able to relate and classify protocol performance with the configuration parameters.

C. Directed Acyclic Graph

Acyclic graphs are directed graphs with no cyclic sub graphs. Not necessarily be a fully connected graph.

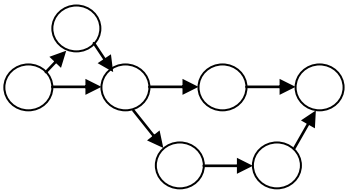

Figure 1: Directed Acyclic Graph

IV. Prior Prediction Models

A. Analytical Models

Analytical models are models used to predict certain kinds of data from the pre-existing data sets. These models are used to find the best configuration for particular network topology to increase its performance. These can be either statistical or probabilistic methods. Theoretically these show promising results.

B. Mathematical models (formula based)

These models are used for prediction of behaviour by using mathematical formulae based on properties such as round-trip time (RTT), packet loss rate, and receive window size. Different measuring tools are used to gather the data.

C. History based models

In these models, the standard time series forecasting based data is used for prediction of different parameters of networking such as TCP throughput.

D. Redundancy of these models.

All of these models show promising results but fail at a certain point due to which they have been proved to become redundant when compared to the changing needs of the network.

While analytical models are quite efficient for predicting certain configuration parameters for networks topologies, these are not effective when in real-world use-cases. It is due to continuous change in the network and inclusion of new protocols and parameters. Along with a changing networks the analytical models fail in practicality as these generate approximate results and predictions which may not cover all the configuration parameters. Scalability is another aspect in which these models fail. These models also have to be explicitly changed when the type of data change.

For formula based models the problem is, limited instrumentation access complicate the basic task of gathering path information at regular intervals of time and with the ever changing network parameters and features, the formula based models have to be changed also, which becomes inconvenient.

The problem with history based approach is that they rely on large amounts of file transfer measurements, hence to predict the throughput of a particular transfer size, it is necessary that there should be previous transfers of the same transfer size. Even though this has occurred previously the history based approach averages the output of similar transfers, hence it reaches its limit when it come to small light weight transfers.

Machine Learning is a much more effective technique to address these problems.

V. Machine Learning

Machine Learning(ML) is a branch of Artificial Intelligence that allows a system to learn without being explicitly programmed. The system is provided a set of data which helps it make decisions depending on the kind of information provided. Machine Learning is fairly a recent field of Computer Science and is evolving at an exceptional rate. ML has mainly two kinds of classification techniques; supervised and unsupervised classification. Out of which supervised classification technique is used by the authors. In this technique the classifier is provided two sets of data, one of which is a training set that trains the classifier and the other is the test set, in which the classifier tests its predictions.

A. Machine learning over analytical models

Machine Learning(ML) algorithms solely work on the data it is provided with. Hence the more data an ML algorithm is provided, the better result it generates. Thus with ever-changing networks these algorithms have numerous kinds of data(training set) to learn from. This results with more accurate predictions than the prior models which provide approximate predictions. Since Machine Learning is not about explicit programming and modeling, scalability is not a weakness but a strength.

B. Causal Bayesian Network

This is a kind of network which can be represented by an Acyclic Directed Graph. The nodes are dependent on their parent nodes by parameters. Each vertex acts as a variable or a node and an edge defining a causal reason for the variables to depend on each other. With no edge exist no dependency.

C. Support Vector Regression

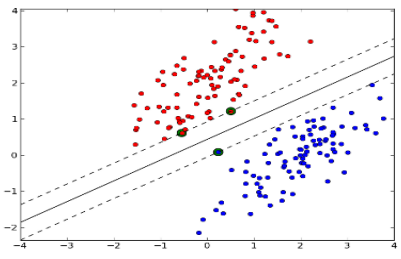

Figure 2: Support Vector Machine

Support Vector Machine (SVM) is an approach of Machine Learning which is used to classify a given set of data by a hyper-plane. The hyper-plane divides the data into different categories and is placed nearest to the differing

categories of the data points. The name *'Support Vector'* arises from the concept of the vectors supporting the hyper-plane are the nearest data points.

Support Vector Regression is a version of SVM in which the classification depends on a sub part of the training data set and ignores the neighboring data. The term 'Regression' is used in Machine Learning for 'Continuous' which may be misleading. So this SVM is used to classify and predict the TCP throughput.

VI. Optimization using Causal Bayesian Network (CBN)

In this section, I discuss the network optimization using CBN. Fabien Geyer *et al* [1] assigned N Virtual Machines (VMs) to the physical hosts. The goal was to get the maximum bandwidth in the networks, by changing the physical hosts for the VMs. The transmission of data was taking place via TCP flows. They denoted bandwidth of each TCP flow *i* by ρ and the number of flows by M , the objective function of optimization is denoted as :

$$maximize \sum_{i}^{M} \rho_i \qquad (1)$$

Taking into consideration that the position of the physical host remain constant while changing the position of the VMs, the best combination of VMs was found, such that the objective function was maximized. The whole process is divided in two major steps; firstly Machine Learning (ML) is used for predicting the performance of the protocol against various topologies, configurations, topology sizes and architectures. Secondly optimizing the topology by using the prediction.

Since this method is based on data and no prior knowledge is required for the network topology, this method can be extended to other network protocols as well. The target function can also be changed depending on the requirement.

A. Feature Selection

For an effective way of measuring the changes in the network, there should be a way of calculating the changes without giving rise to those changes instead. For selecting the features that affect the performance of the network i.e. the bandwidth. Two kinds of analytical algorithms are used for feature selection. First one is square root formula; which is a single TCP flow based on round-trip time (RTT) and drop probability, and the second model shows us the interaction between each TCP flow in a link *l*, bandwidth C_l and N_l flows [1]:

$$\rho = \frac{constant}{RTT\sqrt{p}} \qquad (2)$$

$$\sum_{i \le N_l} \rho_i = \sum_{i \le N_l} \frac{constant_i}{RTT_i\sqrt{p_i}} \le C_l \qquad (3)$$

Lastly the features selected for predicting the bandwidth are number of hops by one flow on RTT, maximum number of flows interfering in the same hop other than the measured flow and the same excluding the ACK packets, and maximum of equation 3 for each link *l*.

B. Learning Process

For predicting and learning the performance of the network topologies the authors use Causal Bayesian Network. They start by building an undirected fully connected graph while continuously removing the edges for every independent vertices(variables) or when a cut set is in the graph. Two tests namely pair-wise correlation test and partial correlation test is used for testing variable independence. Then the edges are given direction based on the causal relationship between the variables. Using copula; a multivariate probability distribution, a function is obtained for the CBN.

C. Topology optimization

Now, this part describes the algorithm used for optimizing the network topology. This method is specially made for cases in which prior knowledge of the protocol is not needed. Hence random-restart hill climbing algorithm is used, whose approach is similar to traveling salesman problem [1].

Initially random positions of the VMs are selected and various permutations of the topologies are measured for performance by *predictPerformance* function and stored if the performance measured is maximum. In case there is no change in the performance for a few iterations then the program terminates.

VII. Optimization using Support Vector Regression

A. Using Support Vector Regression

The authors M. Mirza *et al.* started by experimenting the correlation of TCP throughput and path properties of a network which are packet loss, bandwidth available and queuing delays as the features. Hence there was no need of the long feature selection process as carried out for the CBN model. Here they started with bulk transfers and used the SVR technique to predict the TCP throughput and compared it to the actual TCP throughput. It was observed that during heavy traffic in the network this technique showed 3 times more accurate results than history based models, 87% of predictions were accurate within 10% of actual throughput as compared to 32% within 10% of actual for the history based models, even though these were passive measurements. But with active measurements the accuracy of SVR model dropped significantly to 49% but still maintained a lead over history based model (32%).

It was also observed that available bandwidth had a very small influence over the TCP throughput, hence made this model very light weight. In the history based models the prediction can only be made by large transfers and formula based models are not accurate in practical life when it comes to a variety of file sizes. But here the SVR model can predict accurately even with a huge range of file sizes and without a bulk transfer. By just adding three file size in the training features an accurate TCP throughput can be generated.

VIII. Prediction Accuracy

This shows the accuracy of machine learning used for optimizing the network topology with respect to individual flows and topology wide features. The results of both the Machine Learning methods are compared; from paper [1]

Causal Bayesian Network, and from paper [2] Support Vector Regression (SVR). These two Machine Learning techniques with analytical model (AM)in the table below. The accuracy is measured using the three most commonly used metrics:

- *Mean Absolute Error (MAE or MAR)*
- *Explained Variance Regression Score*
- *Coefficient of Determination (R^2)*

Perf. prediction	Method	MAE	Expl. var.	R^2 score
Individual flow	CBN	4.1024	0.8749	0.8747
	SVR	4.3012	0.8676	0.8624
	AM	1.4208	0.8910	0.8879
Topology-wide	CBN	2.5405	0.8427	0.8427
	SVR	4.3464	0.5374	0.4687
	AM	2.0045	0.8575	0.8030

Table I: Performance prediction of CBN vs. SVR vs. AM [1]

As observed above of individual flows CBN and SVR are as efficient as each other but when it comes topology-wide CBN is much better than SVR technique for R^2 score but, SVR is better in than CBN for MAE metric. But in both the flows the ML techniques give higher results than analytical models.

Finally the results are compared of the Machine Learning technique using CBN for the most optimized placement of Virtual Machines (VMs) allocation to the initial allocation of the VMs, and shown below:

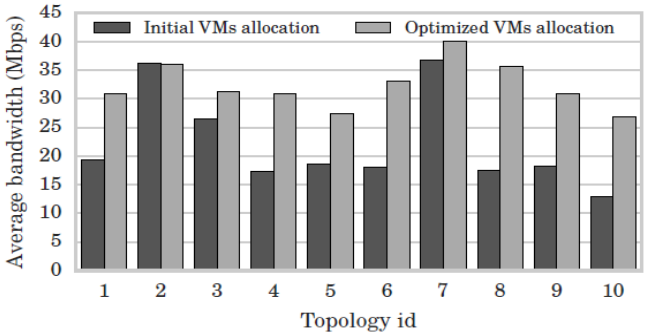

Figure 3: Initial allocation and Optimized allocation of VMs [1]

IX. Conclusion

In this survey it is observed that implementing Machine Learning approach to predict the TCP throughput is a better and feasible alternative to the existing non ML models. When using Causal Bayesian Networks for performance prediction of the bandwidth for TCP flows, the results show an increase in the average performance by more than 50%. Alternatively Support Vector Regression model prediction show an improvement by a factor of 1.6 than the best method available in the prior models.

X. Future work

A. Data Centers

Data centers face numerous challenges which are:

1) *Network Planning:* Data centers usually face with the problem of figuring out unused switches and failures in the network but with the help of Machine Learning the network nodes can plan the network topology without delay and optimize the flow before hand.

2) *Performance Management:* With a lot of VMs running in the host and different kinds of traffic in the network, the data centers can make use of Machine Learning (ML) techniques to optimize the performance of the network. Since nowadays most of the data center networks are Software Defined Networks they can easily collect statistical data for Machine Learning techniques.

3) *Scalability:* Since an everlasting challenge for data centers is scalability, ML techniques are effective ways of not explicitly programming the nodes in the network with the network topology every time it changes.

4) *Anomaly Detection:* SVM can easily classify any anomalies or any intrusion in the network. One of the main advantages of using this method is that high feature dimensionality is never a problem with this ML technique [3].

5) *Node localization in wireless networks:* This is similar to the first application, that is network planning, the physical location of a wireless node can be easily determined. SVMs can be used as a cheaper and effective way to achieve this goal [3].

B. Other Protocols

As discussed before, Machine Learning has a lot of advantages one of which being, is scalability and does not require explicit information about the network topology or changes in the network when used in the nodes for performance enhancement. Thus this kind of technique can be used for other kinds of protocols as well. In the survey only TCP protocol is used due to it being the most commonly used protocol.

XI. References

[1] Fabien Geyer, Georg Carle, *"Towards Automatic Performance Optimization of Networks Using Machine Learning,"* IEEE Networks, 2016.

[2] M. Mirza, J. Sommers, P. Barford, and X. Zhu, *"A Machine Learning Approach to TCP Throughput Prediction,"* IEEE/ACM Trans. Netw., 2010.

[3] Mariyam Mirza, *"A Machine Learning Approach to Problems in Computer Networks Performance Analysis"*, Thesis Paper, University of Wisconsi