ISTANBUL TECHNICAL UNIVERSITY
FACULTY OF COMPUTER AND INFORMATICS

Digital Twin and Predictive Maintenance of Cellular Traffic

Graduation Project Interim Report

Berdan Çağlar Aydın
150170068

Department: Computer Engineering
Division: Computer Engineering

Advisor: Prof. Dr. Berk Canberk
February 2022

Statement of Authenticity

I/we hereby declare that in this study

1. all the content influenced from external references are cited clearly and in detail,
2. and all the remaining sections, especially the theoretical studies and implemented software/hardware that constitute the fundamental essence of this study is originated by my/our individual authenticity.

İstanbul, Şubat 2022

Berdan Çağlar Aydın

Acknowledgments

Thank you to my supervisor, Prof. Dr. Berk Canberk, for providing guidance and feedback throughout this project.

Digital Twin and Predictive Maintenance of Cellular Traffic

(SUMMARY)

The use of cellular communication technologies has made an incredible rise in the last two decades. Billions of people benefit from mobile communications. This traffic density requires good planning and maintenance of these systems. Coverage planning is one of the subjects studied in this area.

When deploying a cellular network, coverage planning is a fundamental issue. The aim is to achieve maximum quality coverage while keeping the total power usage and cost to a minimum. For this purpose, the base station location is a necessary planning task. The issue is determining how many, where, and at what height base stations should be placed to achieve the requirements.

While determining the base station locations, the distance of the users to the cell, the number of users in a cell, and the traffic generated by each user should be taken into consideration.

For the best performance, these plans must be made in detail. Predictive maintenance techniques using machine learning provide promising possibilities to make data-driven decisions and take required maintenance. One of the promising solutions is generating data from a system simulation. The simulation can even be tuned to a real system operation as a digital twin.

A digital twin uses real data about a real-life object or system as input. It then generates predictions or simulations of how the real object or system will react, based on these inputs. In its simplest form, it is a computer program that can simulate.

The digital twin can be fed with users location and traffic information and help cellular planning using machine learning. User locations and user generated traffic can be modeled as weighted data samples. This data can be clustered with various machine learning algorithms and a base station can be assigned to each cluster. With this approach, the number, position, and radius of cells can be optimized in order to provide the best coverage to a given user population.

The method will be converted into a simulation. The simulation can be controlled with a screen. This enables one to control and monitor the results using different parameters on algorithm and tying different population data.

In this study, it will be examined whether machine learning and digital twins can be effective in base station location planning. In the future, resulting simulation can be fed with real user data and play role in real-life coverage planning.

Dijital İkiz ve Hücresel Trafiğin Öngörülü Bakımı

(ÖZET)

Hücresel iletişim teknolojilerinin kullanımı son 20 yılda büyük artış gösterdi. Bu artış sonucunda oluşan trafik yoğunluğu bu sistemlerin iyi bir şekilde planlanmasını gerektiriyor. Bu alandaki optimizasyon çalışmaları kapsama planlaması,güç optimizasyonu ve kanal ataması gibi konuları içermektedir.

Hücresel bir ağ kurulurken kapsama planlaması temel bir konudur. Amaç toplam güç kullanımını ve maliyeti minimumda tutarken maksimum kalite kapsamayı elde etmektir.Bu amaçla baz istasyonu lokasyonu gerekli bir planlama görevidir. Bunun sonucunda ortaya çıkan bir problem, kapsama ve kapasite gereksinimlerini karşılamak için baz istasyonlarının kaç tane, nerede,ve hangi yükseklikte bulunması gerektiğini belirlemektir.

Baz istasyonlarının konumları belirlenirken kullanıcıların hücreye olan uzaklığı, bir hücredeki kullanıcı sayısı ve her bir kullanıcının oluşturduğu trafik dikkate alınmalıdır.

En iyi performans için bu planlar detaylı bir şekilde yapılmalıdır. Makine öğrenimini kullanan kestirimci planlama teknikleri, veriye dayalı kararlar almak ve gerekli bakımı yapmak için umut verici olanaklar sağlar. Gelecek vadeden çözümlerden biri, bir sistem simülasyonundan veri üretmektir.

Simülasyon, dijital ikiz olarak gerçek bir sistem çalışmasına bile ayarlanabilir. Dijital ikiz, girdi olarak gerçek hayattaki bir nesne veya sistem hakkındaki gerçek verileri kullanır. Ardından, bu girdilere dayanarak gerçek nesnenin veya sistemin nasıl tepki vereceğine dair tahminler veya simülasyonlar üretir. En basit haliyle simüle edebilen bir bilgisayar programıdır.

Dijital ikiz, kullanıcıların konun ve trafik bilgileri ile beslenebilir ve makine öğrenimini kullanarak hücresel planlamaya yardımcı olabilir. Kullanıcı konumları ve kullanıcı tarafından oluşturulan trafik, ağırlıklı veri örnekleri olarak modellenebilir. Bu veriler çeşitli makine öğrenmesi algoritmaları ile kümelenebilir ve her kümeye bir baz istasyonu atanabilir. Bu yaklaşımla belirli bir kullanıcı popülasyonuna en iyi kapsamı sağlamak için hücrelerinin sayısı, konumu, ve yarıçapı optimize edilebilir.

Yöntem bir simülasyona dönüştürülecektir. Simülasyon bir ekran ile kontrol edilebilir. Bu algoritma üzerinde farklı parametreler kullanarak ve farklı popülasyon verilerini bağlayarak sonuçların kontrol edilmesini ve izlenmesini sağlar.

Bu çalışmada baz istasyonu yerleşim planlamasında makine öğrenmesi ve dijital ikizlerin etkili olup olmayacağı incelenecektir. Gelecekte ortaya çıkan simülasyon, gerçek kullanıcı verileri ile beslenebilir ve gerçek yaşamda kapsama planlamasında rol oynayabilir.

Contents

1 Introduction and Problem Definition

The aim of this study is maximizing the coverage while maintaining the quality of service by using minimum number of cells. In order to do this, users are properly divided into clusters and a base station is placed in the center of each cluster.

Clustering is the grouping of data showing similar characteristics in a data set. Within the same cluster, the similarities are high, and the similarities between the clusters are low. While clustering the users, unsupervised clustering algorithms will be used since the labels of data samples are not predetermined.

To be able to implement clustering, the users should be converted into data samples. A location vector (x, y) for each user and a weight attribute representing the traffic generated by the user should be defined. As a result, each user will be converted into a data sample with three features: x, y and traffic.

K-Means algorithm is an unsupervised learning and clustering algorithm. According to the working mechanism of the K-means algorithm, K objects are randomly selected to represent the center point or mean of each cluster. The remaining objects are included in the clusters with which they are most similar, taking into account their distance from the mean values of the clusters. Then, by calculating the average value of each cluster, new cluster centers are determined and the distances of the objects to the center are examined again. The algorithm continues to repeat until there is no change.

K-means algorithm is used while clustering. Using this algorithm, k is given as input by iterating within a certain interval. While iterating, it is checked whether the created clusters meet certain criteria.

At each iteration, the following is calculated for each cluster:

- The distance of the furthest sample from the cluster center to the center (Mean Distance).
- The average of the distances from the center of all the samples of the cluster (Max Distance).
- Sum of squared distances of samples to their closest cluster center, weighted by the sample weights. (Inertia)
- The number of samples each cluster contains. (Number of Samples)
- Total traffic within each cluster. (Total Traffic)

When these metrics reach the desired range, the iteration stops and the optimal number of clusters is obtained. After, the radius of cells should be determined with reference to the cluster centers. Various approaches should be tried when finding cell radius. The distance of the sample farthest from the cluster center or divisions of this distance can be used.

Different examples of the work in its current state can be seen below (max distance from cluster center taken as radius):

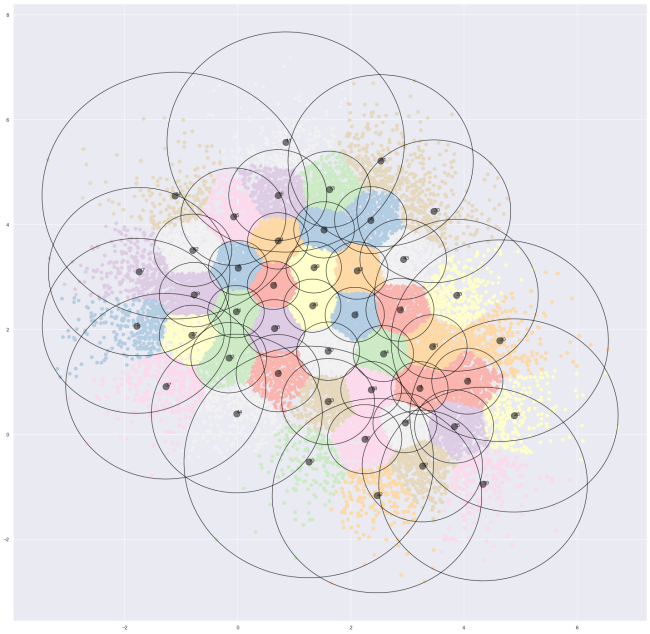

Figure 1.1: n_samples = 25000, mean_distance $<$ 1, max_distance $<$ 2.5, inertia $<$ 5000, k = 46

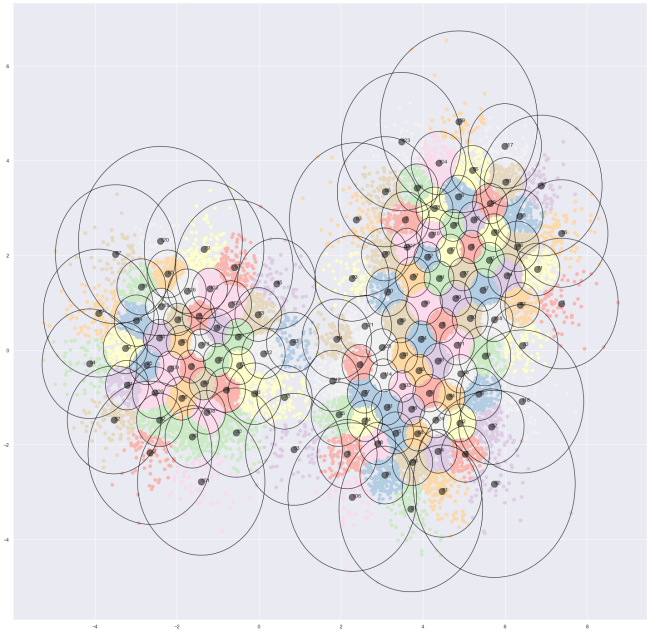

Figure 1.2: n_samples = 27000, mean_distance $<$ 1.5, max_distance $<$ 2.7, inertia $<$ 5000, k = 126

Comparing the Figure 1.1 and Figure 2.2, it can be seen that decreasing the maximum distance from cell center is resulting a huge increase in number of cells.

2 Literature Survey

Cellular scheduling is a general problem but is actually a complex and non-scalable optimization problem that has been shown to be NP-hard [1]. Even just detecting the location of cells has been shown to be NP-hard. [1]

Several studies has been made to come up with a unsupervised method in order to minimize the number of the cells while considering coverage, capacity and power constraints. Some research, for example, used a predetermined number of cells and then looked at their placements and users' associations [2] [3]. The problem is separated into two parts in [4]. First, given a certain number of cells, their placements are adjusted to reduce the total amount of power needed. Second, the proposed technique discovers and removes unnecessary cells. In [5] the goal is to install a large number of cells at random, then use an iterative method to remove the duplicate ones.

In 2021, it is explained that it is possible to use digital twins to build a corresponding virtual network for cellular networks. The network can collect the traffic information and the data can be processed with appropriate methods. [6]

3 Novel Aspects and Technological Contributions

In this study, we are going to build a simulation which will enable users to do cellular network planning and see the resulting metrics according to the their desire. In other studies, an easy-to-use simulation with good visual presentation has not been created. AnyLogic simulation modeling software tool will be used to achieve this.

4 System Requirements

Non-functional requirements can be written as below.

- Clustering should occur within an acceptable time frame.
- Backend server should serve to the simulation reliably.
- Simulation screen should be user-friendly.

4.1 Use Cases / User Stories

User stories are used to represent functional requirements below

- As a user, I want to be able to run the simulation using different datasets.
- As a user, I want to be able to run the simulation with the parameters I have given myself.
- As a user, I want the simulation to recommend me the optimal planning.
- As a user, I want the simulation to show the results in a clear and understandable way.

5 Project Plan

Project is planned to be implemented in a serial fashion. The plan will generally focus on choosing the best and logical parameters for the algorithm to run.

5.1 Resource Requirements

While developing, a laptop with 8GB of RAM and a 1.8Ghz processor will be used. Python scikit-learn library will be used for implementing the k-means and other machine learning related algorithms. Flask and AWS will be used to host an API in order to send and receive JSON data. Additionally, AnyLogic simulation software is required to build a simulation and a control screen.

5.2 Work Breakdown and Work Assignment

In Figure 5.1 you can see the Work Breakdown Structure diagram of the project.

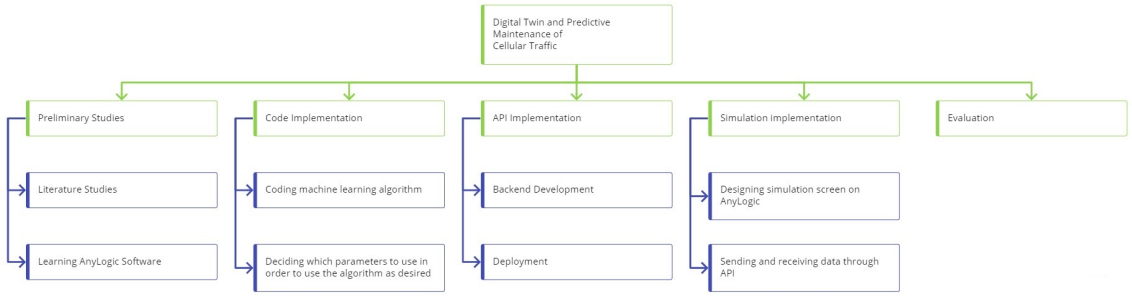

Figure 5.1: Work Breakdown Structure

5.3 Time Plan

In Figure 5.2 you can see the Gantt diagram of the project.

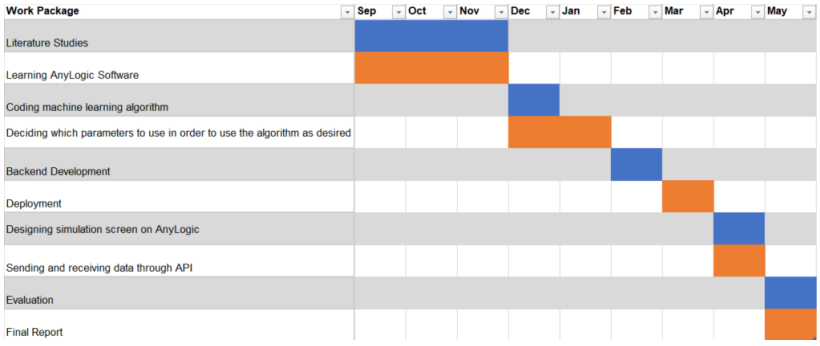

Figure 5.2: The Gantt diagram of the project

6 Goals and Evaluation Criteria

The aim of the project is to enable users to make cellular planning according to the parameters they want and observe the results. In addition, it is to provide the user with optimal cellular planning in a convenient way. This simulation can also be used as a digital twin in the future.

1. The algorithm should be able to work with any given parameter
2. API latency should be less than 10 seconds
3. Zero bugs or crushes in simulation

References

[1] E. Amaldi, A. Capone, and F. Malucelli, "Planning umts base station location: optimization models with power control and algorithms," *IEEE Transactions on Wireless Communications*, vol. 2, no. 5, pp. 939–952, 2003.

[2] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Drone small cells in the clouds: Design, deployment and performance analysis," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.

[3] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-d placement of an aerial base station in next generation cellular networks," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–5.

[4] A. Abdel Khalek, L. Al-Kanj, Z. Dawy, and G. Turkiyyah, "Optimization models and algorithms for joint uplink/downlink umts radio network planning with sir-based power control," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1612–1625, 2011.

[5] W. El-Beaino, A. M. El-Hajj, and Z. Dawy, "A proactive approach for lte radio network planning with green considerations," in *2012 19th International Conference on Telecommunications (ICT)*, 2012, pp. 1–5.

[6] Y. Wu, K. Zhang, and Y. Zhang, "Digital twin networks: A survey," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13 789–13 804, 2021.