JES
Journal of Engineering Sciences

Vol 13, Issue 02, FEB /2022
ISSN NO:0377-9254

FLEXIBLE MACHINE LEARNING-BASED CYBERATTACK DETECTION USING SPATIOTEMPORAL PATTERNS FOR DISTRIBUTION SYSTEMS

MANINDRANATH KOMMINENI[1], POOJITHA YATA[2]

#1,#2, SRM Institute of Science and Technology

Mail id : manichowdary710@gmail.com

Mail id : poojithareddy558@gmail.com

Abstract:

In the development of a machine learning strategy for detecting cyberattacks on distribution systems, spatial and temporal patterns are considered. System-wide data are used to determine regional and temporal trends in Laplacian's graph. Bayes classifier (BC) has been used to train spatiotemporal patterns that could be violated by cyber assaults. Cyberattacks can be detected using flexible BCs that are available online. Test feeds for IEEE 13 and 123 nodes are utilised to demonstrate the approach's applicability.

I. Introduction:

Many distribution-level technologies and assets have had a profound impact on how power distribution networks have been built and operated during the past few decades. Data-driven observability and gridedge data analytics rely on a growing number of sensors (such as micro-PMUs) deployed on the distribution system in conjunction with conventional SCADA systems, advanced measurement infrastructure (AMI), as well as other field devices. When it comes to making judgments and exchanging information, it is imperative that data management systems be utilised (DMS). Distribution systems' current cybersecurity measures are still vulnerable to attack. Future energy delivery systems must be able to detect, dynamically adapt, successfully withstand, and reject a cyberattack using cyber-resilient DMS functions and cybersecurity technology. In contrast to common cyberattack detection methods like naive Bayes classifiers, flexible Bayes classifiers can capture the continuous property of spatiotemporal patterns in system data (BCs). This is all about trends in space and time.

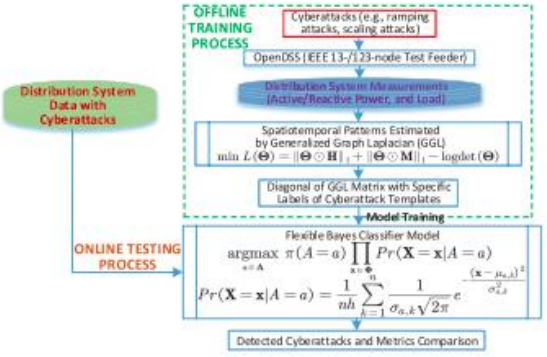


Fig. 1. Flowchart for the newly-developed approach of identifying cyberattacks.

When hacks occur, the integrity of measurement data will be compromised. Based on this premise, the purpose of this letter is to address two major issues in the detection of cyberattacks on distribution infrastructures. I Can the spatiotemporal patterns of cyberattacks and typical settings be quantified? It is possible for operators to strengthen standard cyberattack detection methods by deploying flexible BCs. It is the intention of the authors of this letter to leverage system metrics to construct an adaptable BC for cyberattack detection by using geographical and temporal patterns. Graph Laplacian (GGL) matrices are used to capture spatiotemporal patterns in system measurements. Flexible BC training involves using them as input variables and

then using cyberattack labels as output variables. To create the cyberattack detection findings for testing, the proposed flexible BC makes use of the online spatiotemporal patterns gathered by GGL.

It's not unusual to see S MART metres popping up all over the place in the last few years. Since the end of 2016, 2.9 million smart metres were placed in the UK, with 70 million smart metres in the U.S., and 96 million smart metres in China. An important function of AMI is to record load profiles and facilitate bi-directional information flow [4], which is enabled by the widespread use of smart metres and the associated communication network and data management system. Smart metres are no longer limited to billing.

Customers' electricity consumption habits and lifestyles are revealed by the high-resolution data provided by smart metres. Many countries throughout the world are currently deregulating the electricity industry, notably in the delivery sector. These countries are now focusing their efforts on reforming the retail power sector. Retailers, customers, and aggregators are all taking part in the retail industry, making it more diverse and inclusive than ever before. Worldwide, the use of huge smart metre data for demand-side optimization has emerged as a major topic of discussion. It's been an exciting time for the power business as big data analytics has become increasingly prevalent in the production, transmission, equipment, and consumption processes. Additionally, there has been an increase in the number of smart metre data analytics initiatives in the works. National Science Foundation (NSF) funding encourages interdisciplinary study of smart grid big data analytics. An innovation centre for smart metre data analytics has been established in Denmark by the CITIES Innovation Center. Programs that employ smart metre data to save customers money and improve forecasting [7] are available. Researchers at the Bits to Energy Lab, an interdisciplinary research collaboration between ETH Zurich and the universities of St. Gallen and Bamberg [8], have launched numerous projects based on data gathered from smart metres. [7] Developing intelligent grid data analytics is made possible thanks to funding from the Siebel Energy Institute [9], a global alliance for creative and collaborative energy research. An increasing number of smart grid data analytics projects have been sanctioned by China's National Science Foundation and National Key R&D Program, including the 863 Program for China's National High Technology Research and Development on Key Technologies for Intelligent Distribution and Utilization. An ESSnet Large Data project seeks to examine the use of large data in applications like smart metres [10]. Access, management, and deployment of analytics for smart metres are the focus of this workpackage in the ESSnet Big Data project. Countries including Austria, Denmark, Sweden, Italy, and Portugal are working together on projects.

It is now possible to monitor distribution systems in real time using distribution-level phasor measurement units (PMU). Using GPS coordinates, three-phase voltage and current phasor measurements can be obtained, as well as related magnitude and angle data. [1] Voltage and current measurements in distribution systems can be improved by synchronising voltage and current measurements with increased resolution and precision. PMU data anomaly detection can be used in a wide range of distribution system applications, including metering. According to Farajollahi et al. [1], problems with the power distribution network can be traced back to distribution-level PMUs. While analysing PMUs, Jamei et al. [3] discovered unexpected

activity in the perimeter control. Although the CUSUM algorithm [3] relies heavily on a predefined threshold that can only be determined by putting data into a semi-supervised behaviour detector, this letter attempts to capture affinity similarities and patterns in PMU data by employing a generalised graph Laplacian (GGL) matrix and the three-sigma rule. To model affinity similarity between PMU anomalies, weighted graphs can be utilised to represent the PMU data. As a result, when abnormal events occur, the PMU data's similarity is disrupted. The purpose of this letter is to provide answers to two key issues about PMU-based anomaly detection. (1) Can the degree to which anomalies and regular PMU data are comparable be described numerically? Users can find system-wide abnormalities by comparing the spatial and temporal similarity. For spatiotemporal analysis and visualisation, a GGL-based anomaly detection approach was developed in this letter Data from high-resolution PMUs was used to investigate the spatiotemporal characteristics of abnormalities for the first time using graph learning algorithms.

Anomalies in load forecasting data are the subject of current detection approaches due to cyberattacks. It is possible to detect cyberattacks on power systems that are affected by renewable energy or system reconfigurations using Mohammadpourfard et al. [8]. A real-time anomaly detection system built by Moghaddass and Wang [9] was used to detect anomalous occurrences and abnormal situations at both the lateral and consumer levels. The foundation for effective data injection detection is short-term state predictions that take into account temporal correlations. A method by Chen et al. [11] to identify and restore incorrect measurements and abnormal disturbances in order to improve the precision of load forecasting was presented in this paper.

II. Literature survey:

Mingjian Cui, Student Member, IEEE, Jie Zhang, Senior Member, IEEE, Anthony R. Florita, Member, IEEE, Bri-Mathias Hodge, Member, IEEE, Deping Ke, and Yuanzhang Sun, Senior Member, IEEE" An Optimized Swinging Door Algorithm for Identifying Wind Ramping Events"

It's possible that more frequent WPREs will have a negative impact on grid economy and dependability. An upgraded swinging door method (OpSDA) that we believe will help identify WPREs is presented here. A piecewise linear approximation is utilised to separate wind power data using the swinging door approach (SDA). Segments are optimised utilising a dynamic programming technique before wind power bumps and post-processing insignificant-ramp periods are taken into consideration. The suggested OpSDA is evaluated using data from two different wind farm scenarios. In terms of performance, OpSDA exceeds SDA and is even better than the L1-Ramp Detect with Sliding Window technique.

Mingjian Cui , Senior Member, IEEE, Jianhui Wang , Senior Member, IEEE, and Meng Yue , Member, IEEE" Machine Learning-Based Anomaly Detection for Load Forecasting Under Cyberattacks"

Power system operators can benefit economically and reliably from accurate load forecasts. Cyberattacks on load forecasts could lead to operators making suboptimal operational decisions for power delivery. These cyberattacks can be effectively and reliably detected using a machine learning-based anomaly detection (MLAD) technique. To recreate the benchmark and scaling data from neural network predictions, k-means clustering is employed. This is followed by an estimation of the cyberattack template using the cumulative distribution function and statistical elements of scaling data. It's

imperative that a cyberattack on the system that generates load projections be estimated with dynamic programming. The MLAD approach is compared to a commonly used symbolic aggregation approximation method. For load forecasting data, numerical simulations suggest that the MLAD approach may effectively and reliably detect cyberattacks. According to Monte Carlo simulations, MLAD is able to withstand thousands of attacks.

Hilmi E. Egilmez, Student Member, IEEE, Eduardo Pavez, Student Member, IEEE, and Antonio Ortega, Fellow, IEEE" Graph Learning From Data Under Laplacian and Structural Constraints"

graphs are essential mathematical structures used in numerous domains to represent data, signals, and processes in diverse ways. Graphs can be learned from data using a unique framework proposed in this paper. As part of this framework, we offer a number of different graph-learning issues, as well as their probabilistic interpretations and corresponding methods. Graph Laplacian matrices can be estimated from observed data under specific structural constraints in graph learning issues (e.g., graph connectivity and sparsity level). A graph Laplacian matrix's precision (inverse covariance) is a Gaussian–Markov random field model's precision (a posteriori parameter estimate), hence these problems are probabilistic in nature. Graph Laplacian and structural restrictions are used to construct customised algorithms for the specified graph learning issues. Experiments have shown that the proposed algorithms are more accurate and efficient than the current best practises.

III. Methodology:

It is shown in Fig. 1 and can be summed up like this: how the new cyberattack detection method works in practise To begin, an unsupervised machine learning method known as GGL is used to explain the spatiotemporal patterns of system measurements. The flexible BC approach is utilised in the second stage to train the GGL matrix's spatiotemporal patterns. When it comes to evaluating detection technologies, true positive rates (TPRs) and contingency tables are used. In the following part, we'll go through each step in depth.

Patterns in Space and Time The Laplacian Method for Graphs

As an unsupervised machine learning method, graph learning may quantitatively express spatiotemporal patterns [2]. Consequently, the GGL is able to retain all positive weights and practically incorporate more connections thanks to negative weights. The GGL matrix can be estimated using Lagrangian optimization:

$$\min \; L(\Theta) = \|\Theta \odot \mathbf{H}\|_1 + \|\Theta \odot \mathbf{M}\|_1 - \log\det(\Theta) \qquad (1)$$

$$\mathcal{L}(\mathbf{A}) = \left\{ \Theta \in \mathcal{L} \middle| \begin{matrix} (\Theta)_{ij} \le 0 & \text{if } (\mathbf{A})_{ij} = 1 \\ (\Theta)_{ij} = 0 & \text{if } (\mathbf{A})_{ij} = 0 \end{matrix} \right\}_{\forall i,j \; i \neq j} \qquad (2)$$

H is the regularisation matrix in this example, and h = (I II). Identifiers are connected together to form me. The regularisation parameter is and the estimated GGL matrix are both in Matrix II. L is a collection of graph Laplacians that we utilise to compare them. A, for instance, is a matrix of similarity. The term "multiplying two matrices by themselves" is used to describe the process. The total of all the components' absolute values is 1. (1-norm). The logarithm of an indicator's acronym is log det (). The Lagrange multiplier matrix is a mathematical formula (M).

BC Detection: Adaptable Cybersecurity

Spatiotemporal patterns are generally treated as discrete and assumed to follow a Gaussian distribution by conventional naive BCs. However, this numerical attribute-based assumption does not applicable to all domains (or classes). Because it relies on nonparametric kernel

estimates rather than a normalcy assumption, flexible BC outperforms nave BC in the majority of applications. The adaptable BC can store any attribute value that is seen over the course of training. Given that the probability density function (f(x)) for one spatiotemporal pattern of measurements is likely to have been compromised by cyberattacks, an approximation of the probability density function for an indefinitely large number of samples could yield an estimate of f(x). The ideal kernel density estimation function fn(x) is considered to be Fn(x), and it can be used to fit the ideal function f to a data set (x). Furthermore, fn(x) is highly pointwise consistent when all samples of the spatiotemporal pattern are guaranteed to be consistent. The aforementioned assumption can be mathematically expressed as follows:

$$Pr\left(\lim_{n\to\infty}\left|\widehat{f}_n(x)-f(x)\right|<\epsilon\right)=1,\quad \forall\epsilon:\epsilon>0 \qquad (3$$

In this case, the fitting error is probably a modest positive integer. Individual instances' templates are A and the observed spatiotemporal patterns are X, respectively. An attack template and an observed spatiotemporal pattern vector are referred to herein as "a" and "x." If A = a, then Pr(A = a|X = x) should be the conditional distribution of the cyberattack template in reality. The flexible Bayes estimation method can be used to acquire an exact value for the probability density function Pr(A = a|X = x). It is possible to use the Bayes method to estimate the likelihood of any cyberattack template based on observable data patterns. Aims of Flexible BC include the following:

$$\arg\max_{a\in\Lambda} Pr\left(A=a|\underbrace{\Phi_1^{\mathbb{S}},\ldots,\Phi_i^{\mathbb{S}},\ldots,\Phi_{N_{\mathbb{S}}}^{\mathbb{S}}}_{\text{Spatial Patterns}},\underbrace{\Phi_1^{\mathbb{T}},\ldots,\Phi_j^{\mathbb{T}},\ldots,\Phi_{N_{\mathbb{T}}}^{\mathbb{T}}}_{\text{Temporal Patterns}}\right)$$
$$=\frac{\pi(A=a)\prod_{\mathbf{x}\in\Phi}Pr(\mathbf{X}=\mathbf{x}|A=a)}{\sum_{a\in\Lambda}\pi(A=a)\prod_{\mathbf{x}\in\Phi}Pr(\mathbf{X}=\mathbf{x}|A=a)} \qquad (4)$$

$$\Longrightarrow \arg\max_{a\in\Lambda}\ \pi(A=a)\prod_{\mathbf{x}\in\Phi}Pr(\mathbf{X}=\mathbf{x}|A=a) \qquad (5)$$

GGL-estimated patterns of space and time are derived from NS spatial data. The total number of time frames in the simulation is represented by this value. Assault template a has a high probability of being employed (A = a). The ramping, random, and smooth-curve forms of cyberattacks may all be seen in this diagram. The limited flexibility of the BC is a result of

$$Pr(\mathbf{X}=\mathbf{x}|A=a)=\frac{1}{nh}\sum_{k=1}^{n}G(\mathbf{x};\mu_{a,k},\sigma_{a,k})$$
$$=\frac{1}{nh}\sum_{k=1}^{n}\frac{1}{\sigma_{a,k}\sqrt{2\pi}}e^{-\frac{(\mathbf{x}-\mu_{a,k})^2}{\sigma_{a,k}^2}} \qquad (6a)$$

$$\Lambda=\{Scaling, Ramping, Random, Smooth\} \qquad (6b)$$

$$\Phi=\left\{\text{diag}\left(\Theta^{\mathbb{S}}\right),\ \text{diag}\left(\Theta^{\mathbb{T}}\right)\right\}$$
$$=\left\{\Phi_1^{\mathbb{S}},\ldots,\Phi_i^{\mathbb{S}},\ldots,\Phi_{N_{\mathbb{S}}}^{\mathbb{S}},\right.$$
$$\left.\Phi_1^{\mathbb{T}},\ldots,\Phi_j^{\mathbb{T}},\ldots,\Phi_{N_{\mathbb{T}}}^{\mathbb{T}}\right\} \qquad (6c)$$

$$\Phi_i^{\mathbb{S}}\in\text{diag}\left(\Theta^{\mathbb{S}}\right),\ \Phi_j^{\mathbb{T}}\in\text{diag}\left(\Theta^{\mathbb{T}}\right),\ \mathbf{x}\in\Phi \qquad (6d)$$

In cyberattacks, k is the range of training points for attribute X, A. G() is the kernel function of the Gaussian distribution. As can be seen in Equation (6a), kernel smoothing density functions are used to estimate the continuous attribute's value. Based on [4], Eq. (6b) represents a set of four cyberattack templates. Equivalently, the diagonal of GGL matrix yields patterns in space and time, as shown in Eqs. (6c) and (6d). In the MISE function, h is the bandwidth that can be selected, which is provided by:

$$MISE(h)=\mathbf{E}\left[\int\left(\widehat{Pr}_h(\mathbf{X}|A)-Pr(\mathbf{X}|A)\right)^2\mathbf{dx}\right]. \qquad (7)$$

IV.Results:

Sensors now monitor electricity consumption in smart cities, and this information is sent to distributed energy systems, which then modify energy supply in response to sensor data. If a sensor detects a rise in energy usage at a

particular home, the distributed system will automatically store more energy for that home. It's possible for hackers to hijack sensors and subsequently alter sensor values, which can then be sent to distributed systems, where they can influence the decisions they make. Different sorts of attacks can be created by altering sensors' data, such as SCALING, RAMPING, RANDOM, and SMOOTH CURVE attacks, by attackers. Sensor data containing location, date, and time are referred to as Spatiotemporal Patterns in this context. The term "Spatiotemporal Patterns" refers to any data that contains the location and date of an event.

The author is utilising GGL (General Graph Laplacian) and Flexible Bayes classifier to tackle the problem of existing machine learning algorithms not being able to detect the above threats coming from online sensors.

Steps 1 and 2 of a concept work proposal

As part of this process, the GGL algorithm will be used on sensor data from previous iterations in order to build up a similarity matrix (W) and subsequently a degree matrix (D). It is possible to obtain the laplacian matrix by subtracting D and W.

$$W = \text{Similarity Matrix}$$

$$D = \text{Degree Matrix}$$

$$L = D - W$$

Laplacian matrix rows with negative values indicate aberrant sensor data, which will be marked as a CYBER attack. If the row does not contain negative values, the row is considered normal and will not be flagged as suspicious.

Data from GGL will be fed into a flexible Bayes classifier, which will then be trained on the GGL data to build a classifier model. Sensor data will be fed into a classifier model, which will then be used to forecast if the data contains attack or normal signatures. The Nave Bayes algorithm is analogous to the human brain in that it makes decisions based on satisfying conditions, just like the human brain does.

Humans, for example, will assess the following conditions before allowing a child to play outside:

If it's not raining and the temperature isn't too high,

After that, let the child have some fun by letting him or her play.

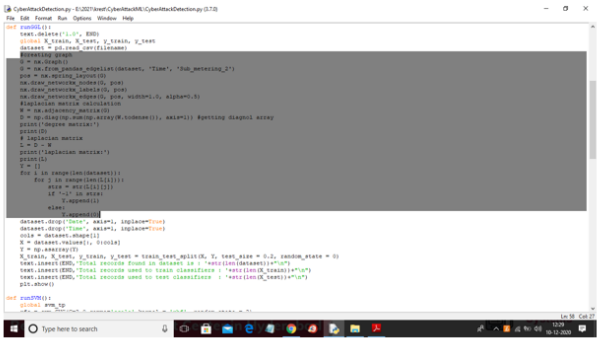
If not,

This is not acceptable.

}

We used the GRID energy distributed system's load dataset to implement this project, and this dataset was saved in the dataset folder. In this research, the author compares the performance of SVM with that of the existing Nave Bayes Classifier and then proposes a new classifier called the Flexible Bayes. Using GGL data, the author came up with the moniker "Bayes Classifier" for the proposed Bayes classifier.

The code screenshots below demonstrate how we arrived at our GGL values.



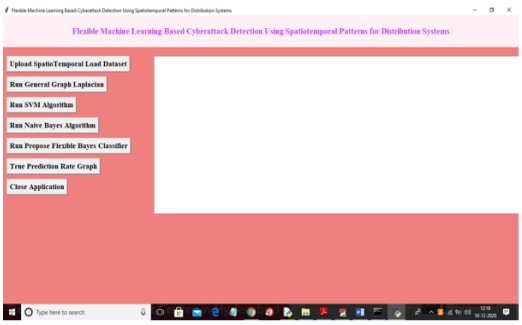
In above screen in selected text read comments to understand laplacian calculation

To run code install below package

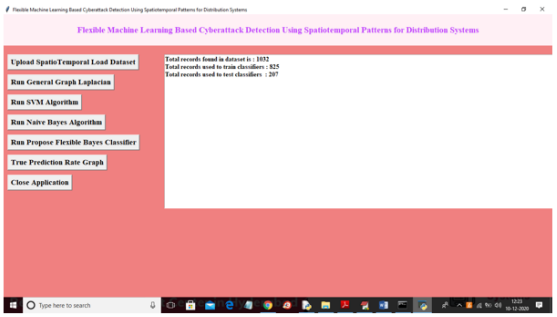
Pip install networkx==2.4

SCREEN SHOTS

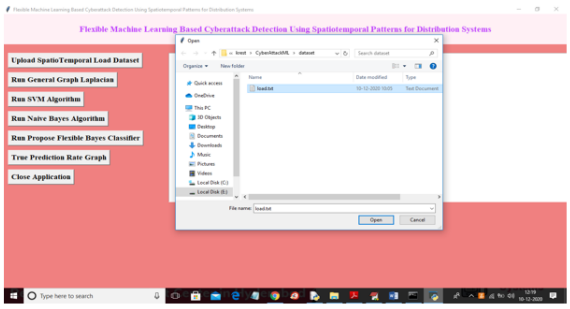
To run project double click on 'run.bat' file to get below screen



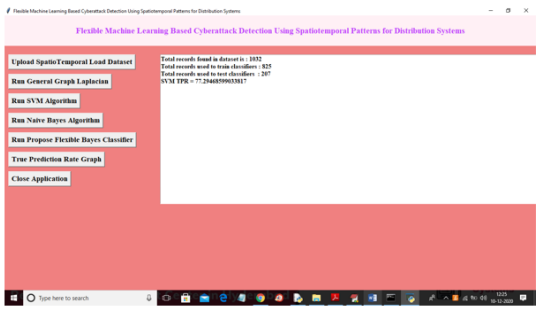
In above screen click on 'Upload SpatioTemporal Load Dataset' button to load dataset



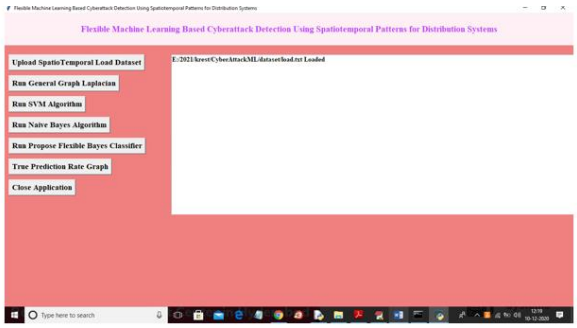
In above screen we can see application got total 1032 GGL records and application using 825 records to train Bayes classifier and 267 records to test or calculate Bayes classifier true prediction rate. Now input data is ready and now click on 'Run SVM Algorithm' button to get its TPR value



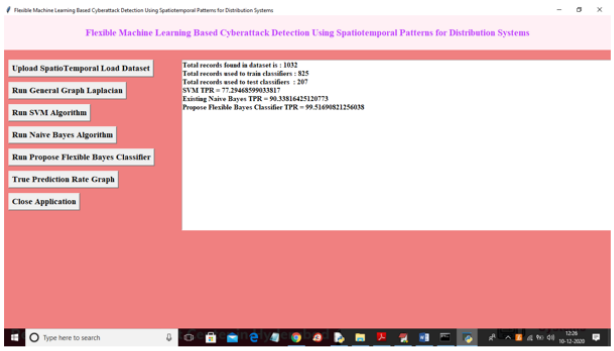
In above screen selecting and uploading 'load.txt' file and then click on 'Open' button to load dataset and to get below screen



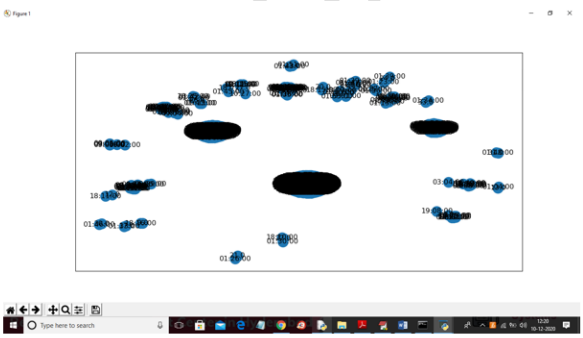
In above screen SVM True Prediction Rate is 77% and now click on 'Naïve Bayes and Propose Bayes Classifier' buttons to get its TPR



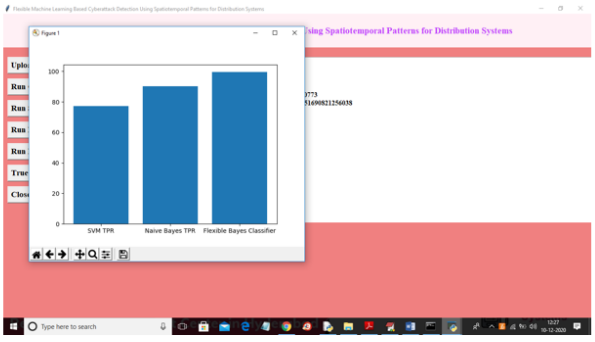
In above screen dataset is loaded and now click on 'Run General Graph Laplacian' button to calculate laplacian matrix and to get below graph



In above screen we can see existing Naïve Bayes TPR is 90% and Bayes classifier TPR is 99% and now click on 'True Prediction Rate Graph' button to get below graph



Above graph generated based on electricity data consumption details send by sensors and in above graph we can see time details (here time is node) connected as edges to energy consumption. From above graph we will calculate laplacian values which will be input to Bayes classifier and below screen we can see input data generated from GGL



In above graph x-axis represents algorithm name and y-axis represents TPR values of those algorithms and from above graph we can conclude that Flexible Bayes classifier giving better prediction rate compare to existing SVM and Naive Bayes algorithms

V. conclusion:

Using generalised graph Laplacian (GGL) and flexible Bayes classifiers, we devise a machine learning-based technique for detecting cyberattacks (BCs). Quantitatively characterising spatial and temporal patterns is done using GGL, which is vulnerable to cyberattacks. Training system measurements and spotting online cyberattacks can be done with the help of the flexible BCs. Based on the outcomes of case studies, the machine learning-based cyberattack detection method has been proven to be effective.

References:

[1] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," IEEE Trans. Smart Grid, vol. 10, no. 3, pp. 3125–3148, May 2019.

[2] M. Cui, J. Wang, A. R. Florita, and Y. Zhang, "Generalized graph Laplacian based anomaly detection for spatiotemporal microPMU data," IEEE Trans. Power Syst., vol. 34, no. 5, pp. 3960–3963, Sep. 2019.

[3] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," IEEE J. Sel. Top. Signal Process., vol. 11, no. 6, pp. 825–841, Sep. 2017.

[4] M. Cui, J. Wang, and M. Yue, "Machine learning based anomaly detection for load forecasting under cyberattacks," IEEE Trans. Smart Grid, vol. 10, no. 5, pp. 5724–5734, Sep. 2019.

[5] M. Cui, J. Zhang, A. R. Florida, B.-M. Hodge, D. Ke, and Y. Sun, "An optimized swinging door algorithm for identifying wind ramping events," IEEE Trans. Sustain. Energy, vol. 7, no. 1, pp. 150–162, Jan. 2016.

[6] Pecan Street Data. Accessed: 2020. [Online]. Available: https://www. pecanstreet.org/category/data port/

[7] R. C. Dugan, Reference Guide: The Open Distribution System Simulator (Opens), vol. 7, Elect. Power Res. Inst. Inc., Palo Alto, CA, USA, p. 29, 2012.

[8] C. Wang, Z. Wang, J. Wang, and D. Zhao, "SVM-based parameter identification for composite ZIP and electronic load modelling," IEEE Trans. Power Syst., vol. 34, no. 1, pp. 182–193, Jan. 2019.

[9] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," IEEE Trans. Smart Grid, vol. 5, no. 2, pp. 580–591, Mar. 2014.