Inspiring the
Leading
Teachnologies

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

ElysiumPRO
Final Year Projects

Data Mining

Titles & Abstract
2023-2024

elysiumpro.in

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_001

PhishCatcher: Client-Side Defense Against Web Spoofing Attacks Using Machine Learning

Cyber security confronts a tremendous challenge of maintaining the confidentiality and integrity of user's private information such as password and PIN code. Billions of users are exposed daily to fake login pages requesting secret information. There are many ways to trick a user to visit a web page such as, phishing mails, tempting advertisements, click-jacking, malware, SQL injection, session hijacking, man-in-the-middle, denial of service and cross-site scripting attacks. Web spoofing or phishing is an electronic trick in which the attacker constructs a malicious copy of a legitimate web page and request users' private information such as password. To counter such exploits, researchers have proposed several security strategies but they face latency and accuracy issues. To overcome such issues, we propose and develop client-side defence mechanism based on machine learning techniques to detect spoofed web pages and protect users from phishing attacks.

EPRO_DM_002

A Diabetes Monitoring System and Health-Medical Service Composition Model in Cloud Environment

Diabetes is a common chronic illness or absence of sugar in the blood. The early detection of this disease decreases the serious risk factor. Nowadays, Machine Learning based cloud environment acts as a vital role in disease detection. The people who belong to the rural areas are not getting the proper health care treatments. So, this research work proposed an automated eHealth cloud system for detecting diabetes in the earlier stage to decrease the mortality rate and provides health treatment facilities to rural peoples. Extreme Learning Machine (ELM) is a type of Artificial Neural Network (ANN) that has a lot of potential for solving classification challenges. This research work is consisting of several activities like feature normalization, feature selection and classification. We have employed principal component analysis (PCA) for feature selection and extreme learning machine (ELM) for classification. Finally, a cloud computing-based environment with three numbers of virtual machines (vCPU-4, vCPU-8, and vCPU-16), is used for the detection of diabetes.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_003

Design of an Intrusion Detection Model for IoT-Enabled Smart Home

Machine learning (ML) provides effective solutions to develop efficient intrusion detection system (IDS) for various environments. In the present paper, a diversified study of various ensemble machine learning (ML) algorithms has been carried out to propose design of an effective and time-efficient IDS for Internet of Things (IoT) enabled environment. In this paper, data captured from network traffic and real-time sensors of the IoT enabled smart environment has been analyzed to classify and predict various types of network attacks. The performance of Logistic Regression, Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting Machine classifiers have been benchmarked using an open-source largely imbalanced dataset 'DS2OS' that consists of 'normal' and 'anomalous' network traffic. An intrusion detection model "LGB-IDS" has been proposed using the LGBM library of ML after validating its superiority over other algorithms using ensemble techniques and on the basis of majority voting.

EPRO_DM_004

Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection

The recent increase in credit card fraud is rapidly has caused huge monetary losses for individuals and financial institutions. Most credit card frauds are conducted online by illegally obtaining payment credentials through data breaches, phishing, or scamming. Many solutions have been suggested to address the credit card fraud problem for online transactions. However, the high-class imbalance is the major challenge that faces the existing solutions to construct an effective detection model. Most of the existing techniques used for class imbalance overestimate the distribution of the minority class, resulting in highly overlapped or noisy and unrepresentative features, which cause either overfitting or imprecise learning. In this study, a credit card fraud detection model (CCFDM) is proposed based on ensemble learning and a generative adversarial network (GAN) assisted by Ensemble Synthesized Minority Oversampling techniques (ESMOTE-GAN). Multiple subsets were extracted using under-sampling and SMOTE was applied to generate less skewed sets to prevent the GAN from modeling the noise.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_005

Practical Strategies for Extreme Missing Data Imputation in Dementia Diagnosis

Accurate computational models for clinical decision support systems require clean and reliable data but, in clinical practice, data are often incomplete. Hence, missing data could arise not only from training datasets but also test datasets which could consist of a single undiagnosed case, an individual. This work addresses the problem of extreme missingness in both training and test data by evaluating multiple imputation and classification workflows based on both diagnostic classification accuracy and computational cost. Extreme missingness is defined as having 50% of the total data missing in more than half the data features. In particular, we focus on dementia diagnosis due to long time delays, high variability, high attrition rates and lack of practical data imputation strategies in its diagnostic pathway.

EPRO_DM_006

PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification

Recent studies have shown that phishers are using phishing kits to deploy phishing attacks faster, easier and more massive. Detecting phishing kits in deployed websites might help to detect phishing campaigns earlier. To the best of our knowledge, there are no datasets providing a set of phishing kits that are used in websites that were attacked by phishing. In this work, we propose PhiKitA, a novel dataset that contains phishing kits and also phishing websites generated using these kits. We have applied MD5 hashes, fingerprints, and graph representation DOM algorithms to obtain baseline results in PhiKitA in three experiments: familiarity analysis of phishing kit samples, phishing website detection and identifying the source of a phishing website. In the familiarity analysis, we find evidence of different types of phishing kits and a small phishing campaign. In the binary classification problem for phishing detection, the graph representation algorithm achieved an accuracy of 92.50%, showing that the phishing kit data contain useful information to classify phishing.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_007

End-to-End Encryption in Resource-Constrained IoT Device

Internet of Things (IoT) technologies will interconnect with a wide range of network devices, regardless of their local network and resource capacities. Ensuring the security, communication, and privacy protection of end□users is a major concern in IoT development. Secure communication is a significant requirement for various applications, especially when communication devices have limited resources. The emergence of IoT also necessitates the use of low-power devices that interconnect with each other for essential processing. These devices are expected to handle large amounts of monitoring and control data while having limited capabilities and resources. The algorithm used for secure encryption should protect vulnerable devices. Conventional encryption methods such as RSA or AES are computationally expensive and require large amounts of memory, which can adversely affect device performance. Simplistic encryption techniques are easily compromised. To address these challenges, an effective and secure lightweight cryptographic process is proposed for computer devices.

EPRO_DM_008

An Effective Method for Mining Negative Sequential Patterns From Data Streams

Traditional negative sequential patterns(NSPs) mining algorithms are used to mine static dataset which are stored in equipment and can be scanned many times. Nowadays, with the development of technology, many applications produce a large amount of data at a very high speed, which is called as data stream. Unlike static data, data stream is transient and can usually be read only once. So, traditional NSP mining algorithm cannot be directly applied to data stream. Briefly, the key reasons are: (1) inefficient negative sequential candidates generation method, (2) one-time mining, (3) lack of real-time processing. To solve this problem, this paper proposed a new algorithm mining NSP from data stream, called nsp-DS. First, we present a method to generate positive and negative sequential candidates simultaneously, and a new negative containment definition. Second, we use a sliding window to store sample data in current time. The continuous mining of entire data stream is realized through the continuous replacement of old and new data. Finally, a prefix tree structure is introduced to store sequential patterns.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_007

End-to-End Encryption in Resource-Constrained IoT Device

Internet of Things (IoT) technologies will interconnect with a wide range of network devices, regardless of their local network and resource capacities. Ensuring the security, communication, and privacy protection of end□users is a major concern in IoT development. Secure communication is a significant requirement for various applications, especially when communication devices have limited resources. The emergence of IoT also necessitates the use of low-power devices that interconnect with each other for essential processing. These devices are expected to handle large amounts of monitoring and control data while having limited capabilities and resources. The algorithm used for secure encryption should protect vulnerable devices. Conventional encryption methods such as RSA or AES are computationally expensive and require large amounts of memory, which can adversely affect device performance. Simplistic encryption techniques are easily compromised. To address these challenges, an effective and secure lightweight cryptographic process is proposed for computer devices.

EPRO_DM_008

An Effective Method for Mining Negative Sequential Patterns From Data Streams

Traditional negative sequential patterns(NSPs) mining algorithms are used to mine static dataset which are stored in equipment and can be scanned many times. Nowadays, with the development of technology, many applications produce a large amount of data at a very high speed, which is called as data stream. Unlike static data, data stream is transient and can usually be read only once. So, traditional NSP mining algorithm cannot be directly applied to data stream. Briefly, the key reasons are: (1) inefficient negative sequential candidates generation method, (2) one-time mining, (3) lack of real-time processing. To solve this problem, this paper proposed a new algorithm mining NSP from data stream, called nsp-DS. First, we present a method to generate positive and negative sequential candidates simultaneously, and a new negative containment definition. Second, we use a sliding window to store sample data in current time. The continuous mining of entire data stream is realized through the continuous replacement of old and new data. Finally, a prefix tree structure is introduced to store sequential patterns.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_009

Classification and Prediction of Significant Cyber Incidents (SCI) using Data Mining and Machine Learning (DM-ML)

Traditional negative sequential patterns(NSPs) mining algorithms are used to mine static dataset which are stored in equipment and can be scanned many times. Nowadays, with the development of technology, many applications produce a large amount of data at a very high speed, which is called as data stream. Unlike static data, data stream is transient and can usually be read only once. So, traditional NSP mining algorithm cannot be directly applied to data stream. Briefly, the key reasons are: (1) inefficient negative sequential candidates generation method, (2) one-time mining, (3) lack of real-time processing. To solve this problem, this paper proposed a new algorithm mining NSP from data stream, called nsp-DS. First, we present a method to generate positive and negative sequential candidates simultaneously, and a new negative containment definition. Second, we use a sliding window to store sample data in current time. The continuous mining of entire data stream is realized through the continuous replacement of old and new data.

EPRO_DM_010

Analysis of Learning Behavior Characteristics and Prediction of Learning Effect for Improving College Students' Information Literacy Based on Machine Learning

Information literacy is a basic ability for college students to adapt to social needs at present, and it is also a necessary quality for self-learning and lifelong learning. It is an effective way to reveal the information literacy teaching mechanism to use the rich and diverse information literacy learning behavior characteristics to carry out the learning effect prediction analysis. This paper analyzes the characteristics of college students' learning behaviors and explores the predictive learning effect by constructing a predictive model of learning effect based on information literacy learning behavior characteristics. The experiment used 320 college students' information literacy learning data from Chinese university. Pearson algorithm is used to analyze the learning behavior characteristics of college students' information literacy, revealing that there is a significant correlation between the characteristics of information thinking and learning effect. The supervised classification algorithms such as Decision Tree, KNN, Naive Bayes, Neural Net and Random Forest are used to classify and predict the learning effect of college students' information literacy.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_011

Dynamic Replication Policy on HDFS Based on Machine Learning Clustering

Data growth in recent years has been swift, leading to the emergence of big data science. Distributed File Systems (DFS) are commonly used to handle big data, like Google File System (GFS), Hadoop Distributed File System (HDFS), and others. The DFS should provide the availability of data and reliability of the system in case of failure. The DFS replicates the files in different locations to provide availability and reliability. These replications consume storage space and other resources. The importance of these files differs depending on how frequently they are used in the system. So some of these files do not deserve to replicate many times because it is unimportant in the system. This paper introduces a Dynamic Replication Policy using Machine Learning Clustering (DRPMLC) on HDFS, which uses Machine Learning to cluster the files into different groups and apply other replication policies to each group to reduce the storage consumption, improve the read and write operations time and keep the availability and reliability of HDFS as a High-Performance Distributed Computing (HPDC).

EPRO_DM_012

MCAD: A Machine Learning Based Cyberattacks Detector in Software -Defined Networking (SDN) for Healthcare Systems

The healthcare sector deals with sensitive and significant data that must be protected against illegitimate users. Software-defined networks (SDNs) are widely used in healthcare systems to ensure efficient resource utilization, security, optimal network control, and management. Despite such advantages, SDNs suffer from a major issue posed by a wide range of cyberattacks, due to the sensitivity of patients' data. These attacks diminish the overall network performance, and can cause a network failure that might threaten human lives. Therefore, the main goal of our work is to propose a machine learning-based cyberattack detector (MCAD) for healthcare systems, by adapting a layer three (L3) learning switch application to collect normal and abnormal traffic, and then deploy MCAD on the Ryu controller. Our findings are beneficial for enhancing the security of healthcare applications by mitigating the impact of cyberattacks.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_013

Fighting Money Laundering With Statistics and Machine Learning

Money laundering is a profound global problem. Nonetheless, there is little scientific literature on statistical and machine learning methods for anti-money laundering. In this paper, we focus on anti-money laundering in banks and provide an introduction and review of the literature. We propose a unifying terminology with two central elements: (i) client risk profiling and (ii) suspicious behavior flagging. We find that client risk profiling is characterized by diagnostics, i.e., efforts to find and explain risk factors. On the other hand, suspicious behavior flagging is characterized by non-disclosed features and hand-crafted risk indices. Finally, we discuss directions for future research. One major challenge is the need for more public data sets. This may potentially be addressed by synthetic data generation. Other possible research directions include semi-supervised and deep learning, interpretability, and fairness of the results.

EPRO_DM_014

Multi-Modal Deep Learning Diagnosis of Parkinson's Disease—A Systematic Review

Parkinson's Disease (PD) is among the most frequent neurological disorders. Approaches that employ artificial intelligence and notably deep learning, have been extensively embraced with promising outcomes. This study dispenses an exhaustive review between 2016 and January 2023 on deep learning techniques used in the prognosis and evolution of symptoms and characteristics of the disease based on gait, upper limb movement, speech and facial expression-related information as well as the fusion of more than one of the aforementioned modalities. The search resulted in the selection of 87 original research publications, of which we have summarized the relevant information regarding the utilized learning and development process, demographic information, primary outcomes, and sensory equipment related information. Various deep learning algorithms and frameworks have attained state-of-the-art performance in many PD-related tasks by outperforming conventional machine learning approaches, according to the research reviewed.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_015

A Cloud-Based Deep Learning Framework for Early Detection of Pushing at Crowded Event Entrances

Crowding at the entrances of large events may lead to critical and life-threatening situations, particularly when people start pushing each other to reach the event faster. Automatic and timely identification of pushing behavior would help organizers and security forces to intervene early and mitigate dangerous situations. In this paper, we propose a cloud-based deep learning framework for automatic early detection of pushing in crowded event entrances. The proposed framework initially modifies and trains the EfficientNetV2B0 Convolutional Neural Network model. Subsequently, it integrates the adapted model with an accurate and fast pre-trained deep optical flow model with the color wheel method to analyze video streams and identify pushing patches in real□time. Moreover, the framework uses live capturing technology and a cloud-based environment to collect video streams of crowds in real-time and provide early-stage results. A novel dataset is generated based on five real□world experiments and their associated ground truth data to train the adapted EfficientNetV2B0 model.

EPRO_DM_016

Detection of Distributed Denial of Charge (DDoC) Attacks Using Deep Neural Networks With Vector Embedding

To prevent excessive strain on the electrical grid and avoid long waiting times of the electric vehicle (EV) at charging stations, charging coordination mechanisms have been implemented. However, there is a potential vulnerability that enable adversaries to launch distributed denial of charge (DDoC) attacks. In these attacks, fake charging requests are sent to book charging time slots without showing up for charging. Existing mechanisms assume the requests from EVs are valid and do not address the detection of DDoC attacks. This research paper aims to assess the disruptive capabilities of DDoC attacks on charging coordination mechanisms and utilize deep neural networks incorporated with vector embedding to develop detectors that can protect against these attacks. The detection approach relies on identifying abnormal behavior that deviates from the typical patterns of charging demand at the charging station. To train and evaluate the detectors, we utilize real routes of vehicles and technical parameters of EVs released by their manufacturers to create a benign dataset.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_017

Privilege Escalation Attack Detection and Mitigation in Cloud Using Machine Learning

Because of the recent exponential rise in attack frequency and sophistication, the proliferation of smart things has created significant cybersecurity challenges. Even though the tremendous changes cloud computing has brought to the business world, its centralization makes it challenging to use distributed services like security systems. Valuable data breaches might occur due to the high volume of data that moves between businesses and cloud service suppliers, both accidental and malicious. The malicious insider becomes a crucial threat to the organization since they have more access and opportunity to produce significant damage. Unlike outsiders, insiders possess privileged and proper access to information and resources. In this work, a machine learning based system for insider threat detection and classification is proposed and developed a systematic approach to identify various anomalous occurrences that may point to anomalies and security problems associated with privilege escalation.

EPRO_DM_018

Question Answering Versus Named Entity Recognition for Extracting Unknown Datasets

Dataset mention extraction is a difficult problem due to the unstructured nature of text, the sparsity of dataset mentions, and the various ways the same dataset can be mentioned. Extracting unknown dataset mentions which are not part of the training data of the model is even harder.We address this challenge in two ways. First, we consider a two-step approach where a binary classifier filters out positive contexts, i.e., detects sentences with a dataset mention. We consider multiple transformer-based models and strong baselines for this task. Subsequently, the dataset is extracted from the positive context. Second, we consider a one-step approach and directly aim to detect and extract a possible dataset mention. For the extraction of datasets, we consider transformer models in named entity recognition (NER) mode. We contrast NER with the transformers' capabilities for question answering (QA).

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_019

Cloud-Based Intrusion Detection Approach Using Machine Learning Techniques

Cloud computing (CC) is a novel technology that has made it easier to access network and computer resources on demand such as storage and data management services. In addition, it aims to strengthen systems and make them useful. Regardless of these advantages, cloud providers suffer from many security limits. Particularly, the security of resources and services represents a real challenge for cloud technologies. For this reason, a set of solutions have been implemented to improve cloud security by monitoring resources, services, and networks, then detect attacks. Actually, intrusion detection system (IDS) is an enhanced mechanism used to control traffic within networks and detect abnormal activities. This paper presents a cloud-based intrusion detection model based on random forest (RF) and feature engineering. Specifically, the RF classifier is obtained and integrated to enhance accuracy (ACC) of the proposed detection model.

EPRO_DM_020

Fraud Detection in Banking Data by Machine Learning Techniques

As technology advanced and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in an increase in the volume of banking transactions. Furthermore, the significant increase in fraud requires high banking transaction costs. As a result, detecting fraudulent activities has become a fascinating topic. In this study, we consider the use of class weight-tuning hyperparameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization in particular to optimize the hyperparameters while preserving practical issues such as unbalanced data. We propose weight-tuning as a pre□process for unbalanced data, as well as CatBoost and XGBoost to improve the performance of the LightGBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, we use deep learning to fine-tune the hyperparameters, particularly our proposed weight-tuning one.

DATA MINING

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

EPRO_DM_019

Cloud-Based Intrusion Detection Approach Using Machine Learning Techniques

Cloud computing (CC) is a novel technology that has made it easier to access network and computer resources on demand such as storage and data management services. In addition, it aims to strengthen systems and make them useful. Regardless of these advantages, cloud providers suffer from many security limits. Particularly, the security of resources and services represents a real challenge for cloud technologies. For this reason, a set of solutions have been implemented to improve cloud security by monitoring resources, services, and networks, then detect attacks. Actually, intrusion detection system (IDS) is an enhanced mechanism used to control traffic within networks and detect abnormal activities. This paper presents a cloud-based intrusion detection model based on random forest (RF) and feature engineering. Specifically, the RF classifier is obtained and integrated to enhance accuracy (ACC) of the proposed detection model.

EPRO_DM_020

Fraud Detection in Banking Data by Machine Learning Techniques

As technology advanced and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in an increase in the volume of banking transactions. Furthermore, the significant increase in fraud requires high banking transaction costs. As a result, detecting fraudulent activities has become a fascinating topic. In this study, we consider the use of class weight-tuning hyperparameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization in particular to optimize the hyperparameters while preserving practical issues such as unbalanced data. We propose weight-tuning as a pre□process for unbalanced data, as well as CatBoost and XGBoost to improve the performance of the LightGBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, we use deep learning to fine-tune the hyperparameters, particularly our proposed weight-tuning one.

ELYSIUMPRO
INSPIRING THE LEADING EDGE TECHNOLOGIES

50K+
Projects Reached

25+
Years of Experience

24/7
Desk Support

25+ Years of Experience | Automated Services | 24/7 Desk Support
Advanced Technologies and Tools | Legitimate Members of all Journals
Quality Product Training | Industry Exposure

(+91) 99447 93398

#229, First Floor, A Block, Elysium Campus,
Church Rd, Anna Nagar, Madurai,
Tamil Nadu 625020

elysiumpro.in