International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-9, Issue-8; Aug, 2023
Journal Home Page: https://ijaems.com/
Article DOI: https://dx.doi.org/10.22161/ijaems.98.6

Data Mining Framework for Network Intrusion Detection using Efficient Techniques

Inderjit Kaur[1], Dr. Pardeep Saini[2]

[1]Research Scholar, Sunrise University, Alwar, Rajasthan, India
[2]Professor, Sunrise University, Alwar, Rajasthan, India

Received: 06 Jul 2023; Received in revised form: 07 Aug 2023; Accepted: 20 Aug 2023; Available online: 26 Aug 2023

Abstract— *The implementation measures the classification accuracy on benchmark datasets after combining SIS and ANNs. In order to put a number on the gains made by using SIS as a strategic tool in data mining, extensive experiments and analyses are carried out. The predicted results of this investigation will have implications for both theoretical and applied settings. Predictive models in a wide variety of disciplines may benefit from the enhanced classification accuracy enabled by SIS inside ANNs. An invaluable resource for scholars and practitioners in the fields of AI and data mining, this study adds to the continuing conversation about how to maximize the efficacy of machine learning methods.*

Keywords— ***Data Mining, Techniques, SIS, ANNs***

I. INTRODUCTION

Data mining, which "mines" knowledge from data, has recently attracted attention from the information industry and society due to the availability of massive amounts of data and the need to turn it into meaningful information/knowledge. Market research, factory administration, basic research, and even customer retention might all benefit from this data. In order to glean even the most fundamental insights from massive datasets, data mining employs a set of fundamental algorithms. Statistics, machine learning, database systems, and pattern recognition are just few of the areas of study that are included within this multidisciplinary field. System security procedures must be built to prevent Organization access to resources/data, and because data mining enables data analysis applications, this is a need. Protecting applications and networks against intrusion in highly interconnected systems is the job of an intrusion detection system.

Password/biometric user authentication, avoiding programming errors like buffer overflow, and encrypting sensitive data on computers are all initial lines of defense. When systems get complex, intrusion prevention alone isn't enough to keep them safe. As the number of people with access to the internet grows, the number of cyber threats faced by businesses also grows.

II. LITERATURE REVIEW

Kumar et al. (2016) The neural network approach has also been shown for automatic tumor detection in liver CT scans. Because the input may include noise introduced during acquisition, the CT input image is first pre-processed using a Median Filter based on expert opinion. The next step involves using first-order statistics and the gray-level matrix to extract local and textural features. Pixels in the gathered data sets are used to determine whether they are associated with the liver or not using a neural network. Tumor boundary detection using an active contour model of the targeted region 32 is possible, and the study's findings are both effective and timely.

J. Peter Campbell (2020) To provide an introduction to contemporary techniques of machine learning, with a focus on selected machine-learning methodologies, best practice, and deep learning, and their application in medical research. The literature on artificial intelligence techniques in medicine, particularly ophthalmology, was searched extensively in PubMed.a summary of machine learning for those who aren't familiar with the ins and outs of programming. However, there are still several obstacles that must be overcome before AI may be widely used in the medical field. This review article aims to provide an accessible overview of current machine learning

applications in healthcare for readers who are not experts in the field. The goal is to help readers understand the potential and challenges of AI in healthcare.

Jonathan Schmidt (2019) There have been many fascinating new additions to the materials science toolset in recent years, but machine learning ranks among the highest. Previous research has shown that this statistical toolkit may significantly speed up both basic and applied research. Studies focusing on applying machine learning to solid-state electronics have recently proliferated. Here, we survey and evaluate the most recent studies addressing this topic. Here, we lay out the foundations of machine learning by introducing key concepts including algorithms, descriptors, and databases for the study of materials science. We proceed to detail further methods in which machine learning may be used to locate stable materials and predict their crystal structure. Here, we provide findings from studies investigating various strategies for using machine learning to supplant first principles in design, as well as quantitative relationships between structures and their attributes. Using examples from the fields of rational design and related applications, we investigate how active learning and surgical optimization may be used to improve the process. There are always major issues with the interpretability and physical understanding of machine learning models. For this reason, we discuss the different facets of interpretability and their importance in the study of materials. In conclusion, we provide solutions to a variety of computational materials science problems and suggest directions for further study.

Pita Jarupunphol (2022) In order to find the most reliable classification model for predicting dengue illness, this research investigates a wide variety of feature selection and classification combinations. Dengue fever prediction parameters based on association patterns were investigated. In order to get the most effective classification model, several feature selection procedures have been categorized and studied with the use of popular classifiers. Many models' measurements were compared graphically. The three-layer neural network model is the most effective. One hundred ReLu-enabled nodes make up each tier. Accuracy of 64.9%, F-measure of 71.8, accuracy of 65.7%, accuracy of 66.0%, and recall of 79.0% were achieved in the identification of five qualities. In addition to the Naive and information gain combination, the Naive and Relief neural network combination, and the Naive and FCBF combination are all competing machine learning approaches with fairly equivalent efficiency. However, if specific feature selection procedures are investigated, SVM is seen as the weaker model.

Saima Anwar Lashari (2018) In this research, we investigate how medical data is currently being categorized and where future prospects may lie by applying data mining techniques. It explains major modern approaches to classification that have been shown to significantly raise the bar for classification precision. Past research has provided literature on the subject of medical data classification through data mining techniques. Extensive research shows that data mining methods excel at the task of classification. This article evaluated and contrasted the current state of medical data classification. The study's findings suggested that the current system for classifying medical data had room for improvement. However, further research is needed to identify and eliminate the uncertainties associated with classification in order to increase precision.

III. DATA MINING ALGORITHM

Without a priori knowledge of the structure of the data points, clustering labels and distributes them to groups of similar objects. The instances of a cluster are unique, but its members are consistent. Organizationa's clustering techniques include the partition algorithm, the hierarchical algorithm, the grid algorithm, and the density algorithm.

Recursively separating cases, hierarchical methods produce clusters from the top down or the bottom up. The following may be further broken down into:

Clusters are initially items, according to agglomerative hierarchical clustering. Once a suitable cluster architecture has been reached, more clusters are fused.

Distinctive hierarchical clustering - Initially, all data points are assigned to a single cluster. After then, a cluster is divided into even smaller clusters. This process is repeated until the cluster is properly structured.

KDD99 DATASET

Third International Competition for Knowledge Discovery and Data Mining Tools produced the data mining technique known as the KDD99 data detection data set. A data set may be thought of as a collection of inferred characteristics of a network link. When it comes to intrusion detection datasets for data mining, the KDD99 IDS dataset has been widely used. Connection records for each link in the Annie George network are among the 42 primary features that make up the KDD99 dataset benchmark.

The KDD 99 is based on five million logs representing seven weeks of network activity, extracted from four gigabytes (GB) of compressed TCP binary dump data. Two million connection records were gathered from two weeks of test data. Using three servers housing computers belonging to the victims, the network mimicked a military

network, revealing several attacks and routine network activity.

There are a total of 65 features in the training and testing data sets, with 24 types of assaults used in training and 14 in testing. Here are a few examples of names for attributes:

- Duration: continuous,
- protocol_type: symbolic,
- Service: symbolic,
- Flag: symbolic,
- src_bytes: continuous,
- dst_bytes: continuous,
- Land: symbolic,
- wrong_fragment: continuous,
- Urgent: continuous,
- Hot: continuous,
- num_failed_logins:continuous,
- logged_in: symbolic,

I Bayes (Nb)

In the simplest type of Bayesian network, I Bayes (NB), all characteristics are treated as unrelated to the value of the class variable. Conditional autonomy describes this situation. In practice, conditional independence almost never holds. Adding the ability to represent attribute dependency is a simple method for expanding Bayes beyond its naive restrictions.

The class node in an Augmented I Bay expands the original I Bay by pointing out direct nodes with links between attribute nodes. I Bayes classification does this by assuming conditional independence to drastically reduce the number of modeling parameters.

$PX|Y$, from original to just 2n

$$PX|Y^{22nl}$$

In real-world settings, including as text categorization, medical diagnosis, and system performance monitoring, I Bayes has been shown to be useful. It works well when there are interdependencies between features because... The quality of the fit to a probability distribution (the suitability of the independence assumption) is unrelated to the optimality of a zero-one loss (classification error). Certain deterministic or low-entropy dependencies result in strong performance on I Bayes, as shown by the effect of distribution entropy on classification error. As entropy decreases toward zero, the I Bayes error disappears. NB is easy to understand and compute.

$$argmax\left(p(c_i)\prod_{j=1}^{n}p(a_j|c_i)\right)$$

NB classifies I by selecting

Random Tree

The decision-making bodies in the random decision tree classification are selected at random. When classifying a test instance, the posterior probability is calculated as the sum of the weighted probability outputs of the individual trees. Generating a random tree has less memory requirements and reduces training time. There are two primary settings to adjust in this ensemble method:

(i) height h of each random tree, and

(ii) number N of base classifiers.

Database analysis, computer science search methods, and even biological models (evolutionary family trees) all make use of random trees in some capacity. As the number of vertices increases indefinitely, the spectrable distributions of the neighboring matrices of the random trees converge on the line of deterministic probability measures, demonstrating a topology of weak convergence.

The average height and average diameter of a random tree is the subject of a large body of literature. The height/diameter enumeration dilemma holds true for both labelled and unlabeled trees, with the anticipated height of a randomly labelled rooted tree being 1 2n. There is a large but scattered body of work on exact/asymptotic results for various models, and many other random tree models have emerged to meet the needs of certain applications. Deep searching in a particular random treeline pattern is reflected here: the "uniform ordered trees" combinatorial model is the model CBP(n) with a shifted geometric (1/2) offspring distribution. When you build on n nodes, you get a random Tn tree. It is easy to calculate the center of the star graph t with vertex 1.

Neural Network

In mathematics and computers, an Artificial Neural Network (ANN), often known as a "Neural Network" (NN), is a model inspired by biological NNs. Information is processed utilizing a network of artificial neurons and a connectionist approach to computing. During the training phase, an ANN adapts its structure in response to information from the network and the outside world. A NN is a widely dispersed, massively parallel processor with easy access to stored accumulated wisdom. In two aspects, it resembles a brain:

1. One acquires information by way of a networked learning procedure.

2. Synaptic weight information is stored as intensities of connections between neurons.

An algorithm for learning describes the method used to carry out the learning process. Neuro-computers (NNs) are a kind of a distributed parallel processor also known as a neuro-network or a connection network.

Advantages:

- In contrast to linear programs, neural networks are able to.
- When an element of a NN fails, the network as a whole keeps running because to its parallel design.
- It is not necessary to retrain a neural network since it is self-learning.
- It's adaptable enough to use in any scenario.
- Its implementation poses no difficulties.

Disadvantages:

- NN needs training to operate.
- Since NN architecture varies from that of microprocessors, the latter needs to be modeled after the former.
- Significant time for processing is required for rganiz.

IV. INVESTIGATION OF FEATURE SELECTION TECHNIQUES FOR INTRUSION DETECTION SYSTEM

One common method for streamlining businesses is called Feature Selection (FS). It improves learning performance (higher classification accuracy), reduces computational costs, and enhances model interpretability by selecting a small subset of relevant features from the original, based on predefined relevance evaluation criteria. Based on whether or not a training set is labeled, FS algorithms are categorized as supervised, unattended, or semi-supervised. FS is a method for identifying, within a collection of data, the subset of features that is optimal for processing according to a certain set of criteria. The method through which an FS may to find a subset ${}^{A}opt^{a}1,opt^{a}2,opt^{\cdots,a}m,opt_{\text{of A,which guarantees}}$ accomplishment of a processing goal by reducing a defined FS criterion ${}_{JfeatureAfeature\_subset}$. Optimal FS solutions are not need to be unique. The faster computation speed and more accurate predictions are made possible by using fewer characteristics in the learning process. Filters and wrappers are two types of FS procedures. First, there is agnostic classification, which does not include any specific methods of categorization. Instead, the wrappers evaluate the quality of a set of features and, from a statistical and computational standpoint, create an efficient filter based on the performance of a classifier type. The relevance of qualities is analyzed using filter techniques by looking just at the data's fundamental properties. The importance of each item is assigned a value, and those with low scores are omitted. Several FS methods are used in this data gathering process. accuracy values depend on the base rates of different classes, therefore in practice, the percentage of accuracy is not preferred for classification. The accuracy of a predictor may be evaluated by calculating its ROC or F-Measure value. Feature ratings evaluate the importance of an individual trait while disregarding the effects of other traits. The output functions of classifiers or statistical methods provide the basis for many ranking systems.

IDS has the potential to mitigate or prevent attacks in the event of updated signatures or improved attack recognition/response capabilities. Intruder detection systems are now distributed real-time component networks rather than batch-oriented monolithic systems. Monolithic IDS either combines all these features into a single system or splits them out into several procedures and applications.

Feature Selection Techniques For Ids

FS is an essential and popular tool for IDS data pre-processing. It has direct repercussions on IDS because of the decreased functionality and the elimination of irrelevant/redundant/noisy data. Many experts recommend using wrapper, filter, or hybrid methods for feature detection in feature selection. In order to evaluate the features' (or feature set's) quality, the wrapper method employs a learning algorithm. The Filter method relies on the central characteristics of the training data to evaluate the relevance of features and feature sets using objective metrics like distance, correlation, and consistency rather than any machine learning methodology.

Feature Selection Based On Correlation (Cfs)

CFS is an efficient FS method, and it selects, using gene expression data, a set of properties that are important to some class. It often reduces the dimensionality of data by over 60% without sacrificing precision.

On the other hand, CFS is able to establish a link between features and classes, as well as features. CFS is a correlation-based rapid filter used in continuous/discrete circumstances. The CFS algorithm ranks a collection of criteria based on their worth or quality. CFS takes use of the best search by using a correlation measure to evaluate a subset's quality, with each feature's predictive power and inter-feature correlation taken into account.

Analysis Of Independent Components (Ica)

Since many ICA features are predetermined at the primary data processing component analysis (PCA) stage, the ICA approaches do not provide such feature selection opportunities. The only feature selection technique used in

ICA face recognition literature to yet to account for this is the percentage of variance (PoV).

Since the original ICA facial recognition architecture provides local features, we've also been working to determine which of these traits are most useful for identifying specific people. In the ICA method, we know nothing about the mixing matrix or the distribution of sources beyond what is gleaned from the data.

Information Gain (Ig)

Word IG measures the data we learn about a category from the presence/absence of a certain word in a text.

Let me be the class number. The IG of a word t must be defined as

$$IG(t) = -\sum_{i=l}^{m} p(c_i)\, logP(c_i) + P(t)\sum_{i=l}^{m} p(c_i|t) logP(c_i|t) + P(\check{t})\sum_{i=l}^{m} p(c_i|t) logP(c_i|t)$$

The error rate for the test set is substantially higher (13.5 percent) than it is for the training set (for which the stated functions constitute an IG filter). This second discovery suggests that a redundancy reduction approach, such a Markov blanket filter, is necessary for feature selection beyond a simple "relevance check."

This section discussed the feature selection strategies that were put to use in this investigation.

V. CONCLUSION

This research takes a look at how normal/abnormal traffic is currently classified using data mining methods and makes recommendations for improvement. The KDD 99 dataset was mined for UDP data streams, and from there a multi-class dataset was created to emphasize the many threats inherent to UDP data streams. Naive Bayes Algorithm, Random Tree, and NN were all shown to be accurate in classifying the dataset's signatures. The random tree-based methods were 99.88% accurate in their classifications. In this study, we compare PCA to the Fisher Score for dimensionality reduction. PCA is a data-minimization technique for discovering and articulating patterns in order to highlight similarities and differences. Fisher Score is a model-based statistical method that may be used to make distinctions. It's a quick and easy approach to evaluate your ability to distinguish between label and trait.

REFERENCES

[1] Kumar, SN, Lenin Fred, A, Lalitha Kumari, S, Anchalo Bensigerm S M 2016, „Feed Forward Neural Network Based Automatic Detection of Liver in Computer Tomography Images", 2 International Journal of Pharm Tech Research CODEN (USA): IJPRIF, ISSN: 0974-4304, ISSN(Online): 2455-9563, vol. 9, no. 5, pp. 231-239.

[2] J. Peter Campbell 2020," Introduction to Machine Learning, Neural Networks, and Deep Learning," Translational Vision Science & Technology February 2020, Vol.9, 14. doi: https://doi.org/10.1167/tvst.9.2.14

[3] Jonathan Schmidt 2019," Recent advances and applications of machine learning in solid-state materials science," npj Computational Materials volume 5, Article number: 83 (2019)

[4] Pita Jarupunphol 2022," dengue fever; data mining; classification; feature selection; ranking.

[5] Saima Anwar Lashari 2018," Application of Data Mining Techniques for Medical Data Classification: A Review," MATEC Web of Conferences 150, 06003 (2018) https://doi.org/10.1051/matecconf/201815006003 MUCET 2017

[6] Lemnaru C. (2012). Strategies for dealing with Real World Classification Problems, (Unpublished PhD thesis) Faculty of Computer Science and Automation, UniversitateaTechnica, Din Cluj-Napoca. Available at website: http://users.utcluj.ro/~cameliav/documents/TezaFinalLemnaru.pdf

[7] Newsom, I. (2015). Data Analysis II: Logistic Regression. Available at: http://web.pdx.edu/~newsomj/da2/ho_logistic.pdf

[8] Pradeep, K. R. & Naveen, N. C. (2017). A Collective Study of Machine Learning (ML)Algorithms with Big Data Analytics (BDA) for Healthcare Analytics (HcA). International Journal of Computer Trends and Technology (IJCTT) – Volume 47 Number 3, 2017. ISSN: 2231-2803, doi: 10.14445/22312803/IJCTT-V47P121, pp 149 – 155. Available from IJCTT website: http://www.ijcttjournal.org/2017/Volume47/number-3/IJCTT-V47P121.pdf

[9] R. Vijaya Kumar Reddy.et.al 2018," A Review on Classification Techniques in Machine Learning,"InternationalJonroual of Advance Research in Sciences and Engineering Volumn no.7

[10] Anu Sharma.et.al. 2017," Literature Review and Challenges of Data Mining Techniques for Social Network Analysis," Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 1337-1354

[11] Zain Anwar Ali 2018," Role of Machine Learning and Data Mining in Internet Security: Standing State with Future Directions," Review Article | Open Access Volume 2018 |Article ID 6383145 | https://doi.org/10.1155/2018/6383145

[12] AJAY SHRESTHA 2019," Review of Deep Learning Algorithms and Architectures," Department of Computer Science and Engineering, University of Bridgeport, Bridgeport, CT 06604, USA

[13] statistical learning theory; optimisation theory; financial econometrics; support vector machine; SVM; kernel methods. DOI: 10.1504/IJBIDM.2019.10019195

[14] Dharmender Kumar 2017," Classification Using ANN: A Review," International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1811-1820

[15] Networks: A Review," (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No. 7, July 2016