Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 44 (2023), No. 3, pp. 331–340
DOI : 10.47974/JIOS-1344

Optimized data analysis through hybrid approach over trusted security environment

Satyajeet Sharma *
Bhavna Sharma †
Department of Computer Science and Engineering
JECRC University
Jaipur
Rajastthan
India

Abstract

For those who can uncover the knowledge hidden inside this data, this offers enormous opportunity, but it also creates new difficulties. In this study, we explore how the contemporary discipline of data mining might be used to glean usable information from the data that surrounds us. k Machine learning techniques include genetic algorithms, Bayesian approaches, and nearest neighbor. By combining these approaches and algorithms, a hybrid method is created in this study. The goal is to successfully categorize data by removing any information that makes it harder to learn. According to solid facts at hand, a novel data set formation strategy is suggested. Five datasets for machine learning from UCI are used in the testing procedure. These data sets pertain to the iris, breast cancer, glass, yeast, and wine. The success of the research is taken into consideration when test findings are analyzed in conjunction with prior efforts.

Subject Classification: *Primary 68T07, Secondary 68M25.*

Keywords: *Cloud, Machine learning, ID3, KNN, RF, SVM.*

1. Introduction

Consequent to the ever-increasing rate of networks of computers and it related usage, increment to global information infrastructure is on high demand resulting to large repository of information which if left unguarded can lead to dangerous situations all around the globe. The

* *E-mail:* **sharma.satyajeet24@gmail.com (Corresponding Author)**
† *E-mail:* **bhavna.sharma@jecrcu.edu.in**

protection of these networks of computers and the information saved or passing across it from cyber-attacks, computer viruses and worm, potential information theft and leakage is of utmost importance. Defending against these malicious hazards had saw to the development of a lots of computer security techniques including firewalls, cryptography, anomaly and misuse intrusion detection system among all others, wherein intrusion detection had strong ground for protection amidst them all due to its defending complex and dynamic intrusion behaviors[1]. Interruption location framework, a kind of safety the board framework utilized to gather and break down data inside a PC organization or in a PC to assist with building a security component from different assaults, is recording gigantic exploration deals with its framework utilizing information mining which has been overflowed by numerous specialists as of late as it has the power of characterizing and distinguishing irregularities inside an organization [2]. IDS detection efficiencies are becoming better day by day as improvements are being discovered. However, most IDS that are being developed are created using classifier and little effort are made on clustering technique in comparison to the rate of classification technique. Clustering technique which is an unsupervised form of classification [3], a data mining category of machine learning algorithms that separate the training dataset based on similar characteristics [4] can be also be used for classification purposes – which is the aim of this research. Clustering algorithms breaks dataset into groups in respect of the information found in the data describing it. It ensures that the attributes of a certain group are alike and distinct to other attributes in other groups [5].

Conclusively, this paper will present an IDS model that will be developed using classification via clustering technique and the considered algorithms will be two popular bunching calculations which are Assumption Amplification (EM) and K-Means calculations. More so, the performance evaluation of these algorithms was carried out on KDDCup'99 datasets and these algorithms performance will be enhanced by preprocessing the dataset using feature reduction technique.

IDS are at present getting the chance to be discernibly basic bit of our security system, and its legitimacy besides expands the estimation of the whole structure. Data mining frameworks can be associated with increment observing learning of intrusion balancing activity parts. They can help recognize new vulnerabilities and interferences, find earlier cloud cases of attacker practices, and give decision help to intrusion organization. Interference disclosure is all things considered the acknowledgment of intrusion lead; moreover it assembles information of the key bit of PC

framework and system, by then reviews them to perceive whether happen the action of resist security strategy.

1.1 *Classifiers*

Different classifiers will divide the extracted words into different groups.

A gullible Bayes classifier

The method we handle the volume of words is via a naive bayes classifier. By assuming that every word in the text is original, this is implied. The probability of each class is calculated as the first step in categorization.

$$P(c) = f/c$$

The total number of training papers, fd, is equal to the number of training documents that include the label "c" in them. As a result, the likelihood of a document belonging to each class is determined using the formula $P(c \mid d) = P(c)\ P(xid \mid c)$

The class for the document d will be determined using this formula, with the greatest probability.

Support Vector Machine (b)

The multiclass classifier is what it is. The 3-d plane will be used to distribute the complete amount of data entities in this instance. The scores for each word are indicated once each value from the dataset is taken. The data item will be placed in the three-dimensional plane based on the results. distinct data elements will be gathered and organized into distinct classes. The SVM line will be created according on the criteria, which will separate the data items and gather them into various classes.

Particle swarm optimization, or PSO

It is another another strategy built on optimization. It is a soft computing technology where different components are inserted into the many levels of chromosomes. The threshold value is used to compare each chromosome. This optimum element identification method will continue until the optimized element's ultimate result cannot be determined. The data elements will be divided into two groups by this whole procedure. The components that are more closely matched to the ideal value fall into one group. And yet another category that is distant from the value that is optimized.

K-Nearest Neighbor (KNN)

It's a different approach where the data objects are identified based on measuring the distance. The Euclidean distance is this distance. Sort

these distances according to increasing distances. The top k neighbors will be chosen and assembled into one set. The other group will get the rest.

2. Theoretical Background

The way to progress in the field of stock exchange relies upon the capacity of trading stocks at the right time where thinking abilities and forecasts assume a significant part. The goal is to "Sell high and purchase low" which sounds simple yet is hard to accomplish and mistakes in supposition can prompt a broad deficiency of cash. Information mining is a prevalently utilized procedure that has assisted with settling dynamic issues in this field and this rising exactness in the precision of expectations in stock trading.[3] In this paper, a near examination of the utilization of different order calculations specifically SVM, Credulous Bayes classifier, and Truck and Fellow are investigated which have been a well-known selection of calculations in stock trade information. At last, a half-breed model is created considering the upsides and downsides of the Credulous Bayes classifier and SVM calculation which has been fruitful in delivering improved exactness in contrast with the current methodologies. Late exploration shows that the significant reason for death because of coronary illness is expanding these days. Around billion individuals will lose their life due to cardiovascular breakdown in a year. Coronary illness otherwise called Cardio Vascular infection can be forestalled by analysis at the beginning phase. Through expectation in AI, we can analyze this sickness. The AI calculation utilizes a few significant highlights for expectation. Highlight choice assumes a significant part in expectation since in a dataset all properties can't be utilized. Just the important elements are utilized. In this paper, we talk about a portion of the component choice techniques utilized in information digging for the expectation of coronary illness. [8] It is of incredible importance to accomplish the expectation of building energy utilization. In any case, AI, as a promising procedure for some, reasonable applications, was seldom used in this field. The main explanation is that the prescient design with best execution is difficult to be resolved. To fill the hole, this paper offers one top to bottom audit, which centers on the precision investigations and model correlations. In particular, the exactness investigations were conducted based on various kinds of structures (for example private structure, business building, government building or instructive structure), diverse kind of transient granularity (for example sub-hourly, hourly, daily or yearly), just as information/yield factors and recorded information assortments. Further,

neural network (ANN) and backing vector machine (SVM), as the plague models, were analyzed in terms of their intricacy of forecast measures, exact nesses of results, the measures of required historical information, the quantities of data sources, and so on at that point the mixture and single AI method were laid out and thought about regarding their qualities and shortcomings. Moreover, a few vitalfacts and further exploration bearings are introduced from a multivariate viewpoint.[7] Data burrowing offers strong methodology for different territories including tutoring. In the preparation field the assessment is growing quickly extending in view of enormous number of understudy's information which can be used to envision significant model relating learning conduct of understudies. The foundations of preparing can utilize enlightening data mining to examine the introduction of understudies which can maintain the association in seeing the understudy's presentation. In data mining plan is an unmistakable strategy that has been completed by and large to find the presentation of understudies. In this examination another assumption computation for assessing understudy's show in academic local area has been made ward on both portrayal and grouping techniques and been tried on areal time premise with understudy dataset of various educational orders of higher informative foundations in Kerala, India. The outcome demonstrates that the hybrid computation uniting gathering and demand approaches yield results that are far dominating in stating of achieving precision in supposition for scholastic execution of the students.

3. Methodology

Hybrid Algorithm:

The half-breed calculation which has been executed is essentially a crossover of two information mining calculations in particular Help Vector Machines and Credulous Bayes Classifier. The two calculations have been consolidated to obtain improved brings about terms of Exactness as well as Kappa and a few different boundaries like Accuracy, Review, and F-Measure.

$$\text{Exactness} = (TP+TN)/N \quad (1)$$

$$\text{Accuracy} = TP/(TP+FP) \quad (2)$$

$$\text{Review} = TP/(TP + FN) \quad (3)$$

$$\text{F-Measure} = (2 * \text{Accuracy} * \text{Review})/(\text{Accuracy} + \text{Review}) \quad (4)$$

where, TP = Genuine Positive, TN = Genuine Negative,

FP = Misleading Positive, FN = Bogus Negative, N = no. of occurrences

Algorithm:

Stage 1: In this step, the information is being taken and ready.

Stage 2: Information is then pre-handled consequently changing the mathematical worth over completely to discrete worth.

Stage 3: After pre-handling the information, Element Determination is applied to the informational index to think of the principal ascribes.

Stage 4: Insignificant traits are being dispensed with from the dataset and afterward crossover SVM calculation is being run on the dataset.

Stage 5: SVM at first does the rate split of the dataset containing in excess of 1000 columns of values to 80% and 20%.

Stage 6: Rate split will be followed up by 10 overlay Cross Approval.

Stage 7: Upon fruitful finish, the half-breed Gullible Bayes classifier will be applied on the resultant dataset of the mixture SVM to create improved results.

4. Result

In our examination, we applied different grouping calculations by utilizing two unique ways to deal with handling the missing data instances, mode-attribution and bitwise erasure strategies with the help of Information mining Calculations. During the experiment, a pre-handled informational index comprises 869 information examples.

The characterization calculations result acquired agreeing on totwo test choices which are:

1. The Grouping Exactness: is the rate number of correctly characterized examples (the number of right forecasts from all expectations made)
2. Accuracy: is a proportion of classifier precision (utilized as a measure for the pursuit viability)
3. Review: is a proportion of classifier fulfillment.
4. F-Measure: likewise called F-Score, it conveys the balance between the accuracy and the review.

Assessment of grouping calculations by test split

In the event of partitioning the info informational index into 66% for the training information and the leftover 34% for testing the classifiers;

Table 1

Hybrid Classifier report on Diabetics Data [10]

	Precision	Recall	f1-score	Support
0	0.80	0.86	0.83	99
1	0.71	0.62	0.66	55
Accuracy			0.77	154
Macro avg	0.76	0.74	0.74	154
weighted avg	0.77	0.77	0.77	154

AUC score: 0.7383838383838384

the results are displayed in Table 1 which give a clear comparison among the chosen classifiers as per exactness, Accuracy, review, and F-measure which shows that:

Hybrid Classifier report
Hybrid Stage One Score: 0.7727272727272727
Hybrid Stage Two Score: 0.7727272727272727
Training score = 0.7931596091205212
Test score = 0.7727272727272727

Here the code output generating the range of max depth values to identify which value provides the best training fit. Fitting the Decision tree model for k = 1 to 20 it found that the lambda = 3 provides the best output. so we use max depth = 3. So on the Lambda values and accuracy scores the plotting accuracy values w.r.t depths of tree is shown in Fig 1

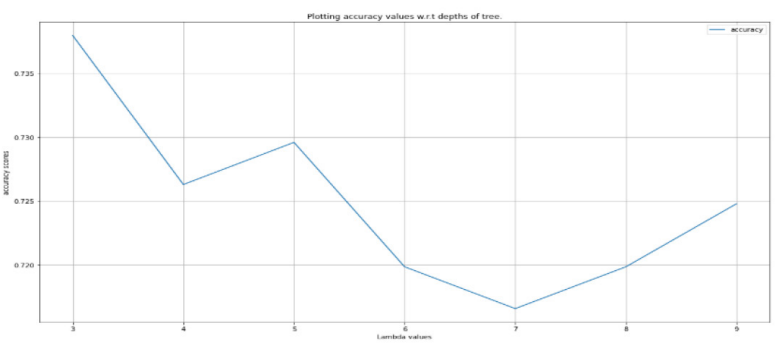

Figure 1

Accuracy Value depth tree

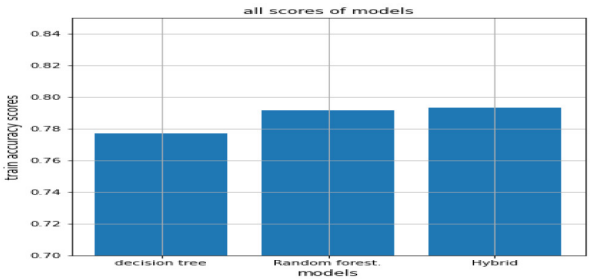

Figure 2

Training Accuracy Score on Patient Dataset

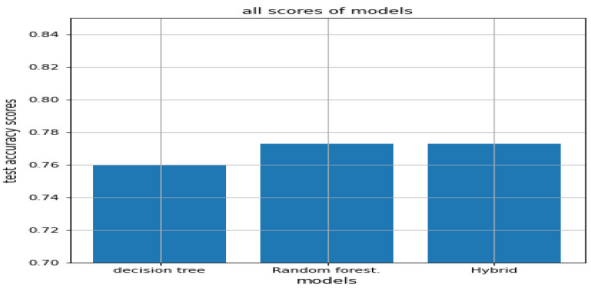

Figure 3

Test Accuracy Score on Patient Dataset

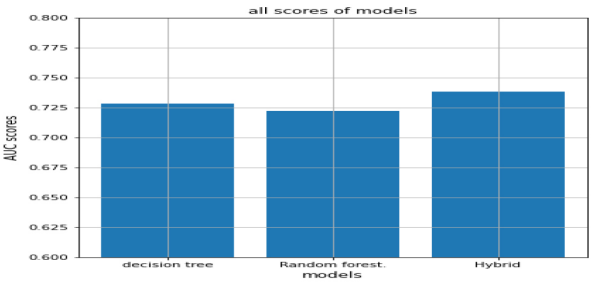

Figure 4

AUC Score on Patient Dataset

5. Discussion and Conclusion

Currently, there are several hardware applications that may provide a vast amount of information. This volume of data will be very difficult to manage. This knowledge is so vast that it would be impossible to keep it

all in one memory. The sequence of the news needs to be handled properly for better handling. These courses are completed in two different contexts, such as intra-news characterisation and buried news categorization. The news is divided into five distinct categories, such as politics, entertainment, sports, and so on, in the entomb news characterization. The intra-news arrangement categorizes the sports news into several categories, such as rugby, football, cricket, and so on. For both burial and intra-news categorization, a hereditary-based half-breed technique has been used. Analysis is being done on the result on several borders. Sensitivity, specificity, exactness, review, and accuracy detective are some of these limits. According to Figures 2, 3 and 4, it is clear that results of the hybrid technique performed better in every boundary when compared to other methods like SVM, KNN Choice trees, and so on.

6. Future Work

Currently, an intra-news characterisation and mixing strategy to news grouping have been carried out. Comparing the half-breed method's results to those of other approaches like SVM, KNN, and Choice tree, the results are noticeably better. Other than the categorization of games, the approach may be used in the future for the intra-news grouping for other classes.

References

[1] Mishra, Narendra, R. K. Singh, and S. K. Yadav. "Detection of DDoS Vulnerability in Cloud Computing Using the Perplexed Bayes Classifier." *Computational Intelligence and Neuroscience* 2022 (2022).

[2] Cheema, Ammarah, et. al. "Prevention Techniques against Distributed Denial of Service Attacks in Heterogeneous Networks: A Systematic Review." *Security and Communication Networks* 2022 (2022).

[3] Samani, Mishti D., et. al. "Intrusion detection system for DoS attack in cloud." *International Journal of Applied Information Systems*. Vol. 10. FCS (2016).

[4] Domeke, Afra, Bruno Cimoli, and Idelfonso Tafur Monroy. "Integration of Network Slicing and Machine Learning into Edge Networks for Low-Latency Services in 5G and beyond Systems." *Applied Sciences* 12.13 : 6617 (2022).

[5] Chakraborty, Meghna, Md Shakir Mahmud, and Timothy J. Gates. "Child Restraint Use and Seating Position of Child Passengers in Motor Vehicles and Their Correlations: Application of a Random-Effects Bivariate Probit Model." *Transportation Research Record* (2022): 03611981221095088.

[6] Gulzar, Zameer, A. Anny Leema, and I. Malaserene. "Human activity analysis using machine learning classification techniques." *Int. J. Innov. Technol. Explor. Eng.(IJITEE)* (2019).

[7] Sharma, Gajanand, et. al. "Self-healing topology for DDoS attack identification & discovery protocol in software-defined networks." *Journal of Discrete Mathematical Sciences and Cryptography* 24.8 : 2221-2232 (2021).

[8] Sharma, Gajanand, et. al. "Video tempering detection assessment in full reference mode using difference matrices." *Journal of Discrete Mathematical Sciences and Cryptography* 22.4 : 645-659 (2019).

[9] Abdullah, Ako Muhammad, and RozaHikmat Hama Aziz. "New approaches to encrypt and decrypt data in image using cryptography and steganography algorithm." *International Journal of Computer Applications* 143.4 : 11-17 (2016).

[10] https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

[11] Amit Kumar Gupta, Pushpa Gothwal, Dinesh Goyal & Carlos M. Travieso-Gonzalez. IoT-Galvanized pandemic special E-Toilet for generation of sanitized environment, *Journal of Discrete Mathematical Sciences and Cryptography*, (2022). DOI: 10.1080/09720529.2022.2068607.

[12] Amit Kumar Gupta, Vijander Singh, Priya Mathur & Carlos M. Travieso-Gonzalez. Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario, *Journal of Interdisciplinary Mathematics*, (2020). DOI: https://doi.org/10.1080/09720502.2020.1833458.

[13] Jitendra Singh Yadav, Amit Kumar Gupta &Arjit Saxena. A review on gender identification using machine learning based on fingerprints, *Journal of Information and Optimization Sciences*, 40:5, 1121-1129 (2019), DOI: https://doi.org/10.1080/02522667.2019.1638002.