Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 1 : 2023
ISSN : **1906-9685**

LEVERAGING THE INTERNET OF THINGS FOR DATA MINING TECHNIQUES

[#1]Mr.VANGAPALLI RAVITEJA, *Assistant Professor*
[#2]Mrs.PINGILI SHANTHI, *Assistant Professor*
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

ABSTRACT
As datasets have grown larger and more complex, the development and expanding capabilities of computer science have facilitated improvements in data collecting, storage, and analysis. This new interconnected world is sometimes referred to as the "Internet of Things" (IOT). Massive amounts of data created by the IoT are valued highly because of their potential business applications. Information hiding in IoT data can be mined using data-mining techniques. Since more and more gadgets are connecting to the IoT, cutting-edge algorithms are essential. This paper offers a thorough examination of many data mining techniques, along with their implementations in the Internet of Things, and assesses their individual strengths and drawbacks.
Keywords — Internet of things (IOT), Data mining, Applications of Data mining.

1. INTRODUCTION

The concept of the "Internet of Things" has been around for over 16 years already. However, the concept of connected devices has been around at least since the 1970s. The term "embedded internet" or "pervasive computing" was first used to describe the idea. Although "Kevin Ashton" of Procter & Gamble fame coined the phrase "Internet of Things" back in 1999, the company has been widely credited for its widespread adoption. Since the World Wide Web was the most exciting new development in 1999, he appropriately titled his presentation "Internet of Things (IoT)". The Internet of Things (IoT) refers to a system in which everyday things are embedded with electronics such as sensors and software. These things can connect to the internet and exchange information with other machines and gadgets.

Data extraction is a lucrative field of study and development. Discovering novel, fascinating, and potentially relevant patterns in large data sets and employing algorithms to extract hidden information are both possible thanks to data mining. The 1990s saw the birth of a new field: data mining, which entails the use of algorithms to unearth previously hidden information and the discovery of novel, interesting, and possibly profitable patterns in large data sets.

Data mining is just one of several analytical tools being included into IoT to increase its intelligence. Data mining encompasses several topics such as statistics, machine learning, artificial intelligence, and databases. However, its primary concern is assuring scalability in terms of attributes and instances through the use of algorithms and the automation of the analysis of enormous volumes of heterogeneous data. Extraction of big data from the IoT presents a number of challenges, including but not limited to dealing with a wide variety of datasets, handling massive amounts of data, and guaranteeing the authenticity of data sources. Novel technologies and data mining techniques are being developed to address these issues as the Internet of Things (IoT) grows in popularity.

Based on the definitions of data mining and data mining functions, the data mining process consists of

the following steps

Data preparation:

Prepare the data for mining by arranging it in a format that makes sense. The data mining process in data mining systems consists of three consecutive steps: aggregating data from multiple sources, minimizing data contamination, and retrieving specific data segments for preprocessing.

Data Mining:

Applying computational methods to information in order to discover hidden relationships. In data visualization, the data and the information derived from it are presented visually to the user

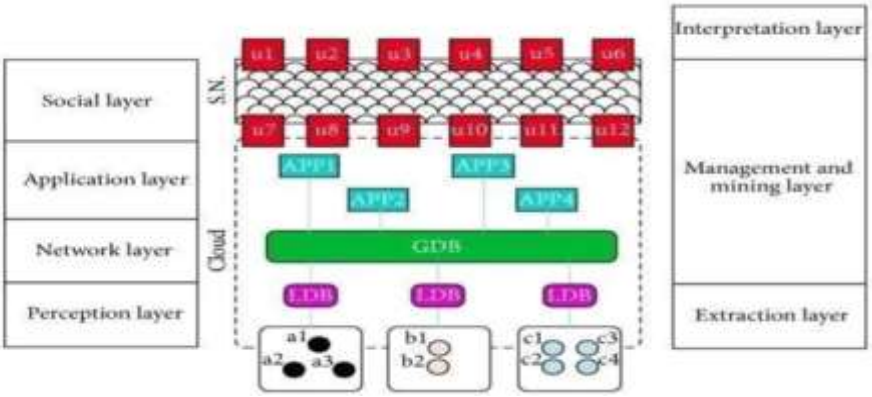

FIGURE 1: Architecture for data mining process

DATA MINING FUNCTIONALITIES

Characterizing, extracting, and differentiating common patterns, linkages, and correlations from data are all activities included in data mining.

The statistical method of outlier analysis can be applied to both classification and regression research.

The key contribution of this paper includes:

Introduction to the world of data mining and the IoT. Data Mining Methodology. Data Mining's Capabilities. Use Cases for Data Mining in the World of the Internet of Things.

What Are the Pros and Cons?

FIGURE 2: Data transfer through Internet of things (IOT)

Data mining via the IoT is already commonplace in businesses as a means to understand consumer tastes, set prices and product placement, and gauge the results on revenue, satisfaction, and bottom line. This is especially the case for businesses operating in highly consumer-oriented industries like retail, banking, communication, and marketing.

Using information collected at the point of sale (POS), a store can apply data mining techniques to develop niche items and offers for certain subsets of its clientele

2.DATA MINING TECHNIQUES IN FRAUD DETECTION IN CREDIT-DEBIT CARD

TRANSACTIONS

The financial losses caused by the fraud are substantial. The procedure for detection is laborious and intricate. Data mining is a process that helps find useful patterns in large datasets and extracts actionable knowledge from it. Knowledge covers any true and significant information. All user information should be safe if you utilize a foolproof approach to detect fraud. Sample record aggregation is an example of supervised learning. The legitimacy of these papers is determined by their authenticity. A model and algorithm are developed to determine whether or not a record is fraudulent based on this data. A fuzzy logic system with carefully determined threshold values was used to include the preexisting fraud evaluation policy. The findings shed light on the likelihood of fraudulent insurance claims and the causes of such fraud. Another logic system used two strategies to mimic the thought processes of fraud specialists. The discovery model is the first method; it uses an unsupervised neural network to find data clusters and linkages, and then to find patterns within those groups. The second method, the fuzzy anomaly detection model, employs the Wang-Mendel algorithm to identify how healthcare providers defraud insurance companies.

Classification strategies are well suited for classifying crime-related data due to their proven effectiveness in identifying fraudulent activity. In order to compare the CART and naive Bayesian classification methods, the distributed data mining model (Chen et al., 1999) uses a realistic cost model. All credit card transactions were handled in this fashion. To extract both conceptual and analog information, the neural data mining method employs rule-based association rules and a Radial Basis Function neural network. This technique evaluates the value of non-numerical information for spotting fraud. It was found that by employing association rules, prediction accuracy could be greatly enhanced. Both Bayesian Belief Network (BBN) and Artificial Neural Network (ANN) research made use of the STAGE algorithm, with the former using it for fraud detection and the latter for back propagation. Internal fraud detection, insurance fraud detection, credit card fraud detection, and telecom fraud detection are all examples of the many types of fraud detection

3. BAYESIAN BELIEF NETWORK

Using the visual representation of causal links provided by Bayesian Belief Networks, we can infer the probability of belonging to a particular class. This helps us decide if a specific incident is authentic or phony. Assuming that an instance's attributes are unrelated to one another when focusing on the target attribute is key to naive Bayesian categorization. The goal is to allocate a new instance to the class that has the highest probability based on the posterior distribution. When compared to decision trees and back propagation, the suggested method is much more efficient and provides better prediction accuracy. Redundancy in the attributes reduces the reliability of the prediction. We employ two Bayesian networks to simulate the characteristics and trends in auto insurance policies in order to spot potential cases of motor insurance fraud. The behavior is represented by two Bayesian networks, one based on the hypothesis that the driver is a fraud (F) and the other based on the hypothesis that the driver is a legitimate user (NF). The "fraud net" is built by people with specialized knowledge. The "user net" is made up of reliable carriers' data. "[7]" is the user-entered text.

During operation, the user network is modified for a certain user using recently collected data. We can calculate the likelihood of the measurement x based on the aforementioned two hypotheses by giving evidence within these networks (the observed user behavior x obtained from their toll tickets) and publicizing it. This means that we have the potential to detect if the observed user behavior is false or genuine. The values are denoted as p(x|NF) and p(x|F). By employing the Bayes rule and assuming the probability of fraud is indicated as P(F) and the probability of non-fraud is P(NF) = 1 - P(F), we can compute the likelihood of fraud given the measurement x as P(F|x) = P(F)p(x|F)/ p(F).

The equation P(x) = P(F)p(x|F) + P(NF)p(x|NF) provides the denominator p(x).

The probability chain rule is as follows: Let's consider two classes, C1 and C2, which represent fraud and legality, respectively. To categorize an instance X = (X1, X2,..., Xn) with each row represented by an attribute vector A = (A1, A2,..., An), the goal is to determine the maximum P(Ci|X) using Bayes theory. The following steps outline the process:

The equation [P (fraud | X) P(fraud)] is equal to [P (fraud | X) P(fraud)]. The probability of event P

being legal given event X, multiplied by the probability of event P being legal, is equal to the probability of event P being legal given event X divided by the probability of event X.
Since the probability P(X) remains constant for all classes, the focus should be on maximizing the values of [P(fraud | X) P(fraud)] and [P(legal | X) P(legal)].
The class prior probabilities can be computed by using the formula P(fraud) = si / s.
Here, s is the overall number of training instances, while si denotes the number of training examples related to the class of fraud.
The assumption of independence between attributes is oversimplified. The conditional probability P(X | fraud) is equal to the sum of n k 1 P(x k | fraud), while the conditional probability P(X | legal) is equal to the sum of n 1 P(x k | legal).
The training samples can be utilized to approximate the probabilities P (x1 |fraud) and P (x2 |fraud).
The number of training instances for the class "fraud" is represented by the variable si, and for the class with the value x k for Ak, it is represented by the variable si k.

4.OUTPUT

Our company offers a Bayesian learning system designed to predict instances of fraud. The categorization findings of the "Output" classification are presented in Table 1. There are a total of 17 valid tuples and 3 invalid tuples. In order to streamline the categorization process, we partition the driver's age attribute into distinct intervals

TABLE 1 TRAINING SET

	Name	Gender	Age_driver	fault	Driver_rating	Vehicle_age	Output
1	David Okyere	M	25	1	0	2	legal
2	Beau Jackson	M	32	1	1	7	fraud
3	Jeremy Dejean	M	40	0	0	7	legal
4	Robert Howard	M	35	1	0.33	1	legal
5	Crystal Smith	F	22	1	0.66	8	legal
6	Chibuike Peason	M	36	0	0.66	6	legal
7	Collin Pyle	M	42	1	0.33	3	legal
8	Eric Pearson	M	30	1	1	2	fraud
9	Kristina Green	F	29	1	0	4	legal
10	Jerry Smith	M	33	1	1	5	legal
11	Maggie Frazier	F	42	1	0.66	3	legal
12	Justin Howard	M	21	1	0	2	fraud
13	Michael Vasovasc	M	37	0	0.33	4	legal
14	Bryan Thompson	M	32	1	0.33	4	legal
15	Chris Wilson	M	28	1	1	6	legal
16	Michael Pullen	M	42	1	0	5	legal
17	Aaron Dotek	M	48	1	0.33	8	legal
18	Bryan Sanders	M	49	1	0	3	legal
19	Derek Garrett	M	32	0	0	3	legal
20	Jasmine Jackson	F	27	0	1	2	legal
X	Crystal Smith	F	31	1	0	2	?

Probability linked to the attributes. By utilizing these simulated training data, we calculate the prior probabilities:
The classifier is required to determine if an instance belongs to the fraudulent or lawful class.

TABLE 2
PROBABILITIES ASSOCIATED WITH ATTRIBUTES

Attribute	Value	Count		Probabilities	
		legal	fraud	legal	Fraud
Gender	M	13	3	13/17	3/3
	F	4	0	4/17	0/3
age driver	(20, 25)	3	0	3/18	0
	(25, 30)	4	0	4/18	0
	(30, 35)	3	1	3/18	1/2
	(35, 40)	3	1	3/18	1/2
	(40, 45)	3	0	3/18	0
	(45, 50)	2	0	2/18	0
fault	0	5	0	5/17	0
	1	12	3	12/17	3/17
driver rating	0	6	1	6/17	1/3
	0.33	5	0	5/17	0
	0.66	3	0	3/17	0
	1	3	2	3/17	2/3

Based on the provided data and the probability linked to the driver's gender and age, we calculate the following estimations: The probability of X given the authoritative source is 0.039, calculated by dividing 4/17 by 3/18.

The product of 3/3 and 1/2 is equivalent to 0.500 P (X |fraud). Hence, the chance of legality is calculated as the product of 0.039 and 0.90, resulting in 0.0351.
The probability of deception is calculated as 0.500 multiplied by 0.1, resulting in 0.050.
To get P(X), we aggregate the likelihood values of these persons, as X has the potential to be either lawful or deceitful: The probability of event X, denoted as P(X), is equal to the sum of 0.0351 and 0.050, which results in 0.0851.
Ultimately, we calculate the precise odds for each individual event:
The result of multiplying 0.039 by 0.9 and then dividing the product by 0.0851 is 0.412.
Due to its highest likelihood, we classify the new tuple as fraudulent using these probabilities. As characteristics are deemed independent, the inclusion of redundant ones reduces predictive efficacy. The introduction of derived attributes relaxes this constraint of independence by merging existing qualities to form new ones. Insufficient data hinders the classification process.
The naïve Bayesian classifier has the capability to accommodate missing values in training datasets. The dataset includes seven missing values to reflect this. The implementation of the naïve Bayes technique is simple and efficient, as it only necessitates a single pass over the training data. When computing the probabilities for each class, any missing values are handled by excluding that likelihood altogether. While the method is simple and easy to understand, it may not always provide favorable outcomes. In general, the attributes are interdependent. We could employ solely a subset of the attributes, excluding those that are contingent on others. The approach is not compatible with continuous data. The resolution of this problem can be achieved by dividing the continuous data into intervals. Nevertheless, this approach is laborious and can impact the ultimate outcomes

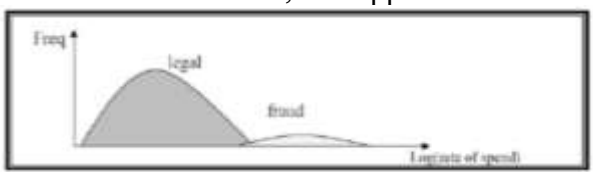

FIGURE 3: frequency distribution of legal and fraud transaction

5. RESEARCH ANALYSIS

Research has experienced paradigm-shifting changes over the course of history. Data mining is valuable for the tasks of data purification, preparation, and database integration. Potentially relevant data from the database that could impact the research can be found. Any sequence and link between acts can be identified. Visual data mining and data visualization offer a clear and precise representation of the data. No existing technology has fully realized its maximum potential. There is consistently a requirement that must be fulfilled. Hence, it can be asserted that data mining through the Internet of Things is a crucial technology in a realm where other technologies can get absolute precision and comprehensiveness with its aid.

Advantages of Mining through IOT:

FIG 4: Industrial internet of things (IOT)

Data mining offers several advantages when applied in a certain business. Furthermore, there are additional benefits to consider, such as privacy, security, and the potential for information misuse.
The utilization of data mining, enabled by the internet of things, provides multiple benefits to corporate operations. Some of its advantages include

Efficient resource utilization:

By fully understanding the functionality and operation of each device, we can greatly improve resource management and environmental monitoring. Minimize human exertion: Through the interaction and communication of data mining equipment, as well as their execution of diverse duties on our behalf, they diminish the need for human effort. Human endeavor

Save time:

It is time-efficient as it necessitates reduced human exertion. IoT data extraction solutions have the ability to significantly reduce the amount of time required for data processing

Enhance Data Collection:

By establishing an interconnected system integrating all of these components, we may enhance its security. Disadvantages of Data Mining through IOT:

Although data mining through the Internet of Things presents certain benefits, it also presents various novel obstacles. Here are some instances of challenges in the field of IoT:

Security:

The process of IoT-based data mining involves the networking and communication of data over interconnected networks. Irrespective of any security measures, the system provides only a restricted level of control and can be employed to carry out various network assaults

Privacy:

Even in the absence of human engagement, the system offers comprehensive and intricate personal information

Complexity:

Developing, building, upkeeping, and facilitating a system with substantial technology is a challenging endeavor

FIG 5: IOT offers Security to the system

6.CONCLUSION

The output includes a tally of the probabilities linked to the attributes. By utilizing these simulated training data, we ascertain the prior probability. By utilizing simulated training data, we establish the initial probability for the purpose of detecting fraudulent activities. Ultimately, the accurate probability of each event are computed. Thanks to the smooth integration of traditional networks and the Internet of Things (IoT). It offers a holistic perspective where all aspects are readily observed and managed, leading to the accumulation of extensive amounts of data. The internet of things, being a pivotal progression in the future of the internet, garners significant apprehensions from both the commercial community and academic spheres. Therefore, the problem of data extraction in the Internet of Things (IoT) becomes a process of examination

REFERENCES

1. Mining with Big data: Jampalachaitanya, Fazi Ahmed parvez. International journal for technological research in engineering. Volume4 issue to oct 2016.
2. Data Mining for the Internet of Thin: Literature Review and Challenge S. Feng chen, Pandeng, JiafuWan, Athanasios V.Vasilakos,Xiaohui.
3. Saral Nigam, Shikha Asthana, and Punit Gupta. Iot based intelligent billboard using data mining.

In Innovation and Challenges in Cyber Security (ICICCS-INBUSH), 2016 International Conference on pages 107–110. IEEE, 2016.
4. Alexander Muriuki Njeru, Mwana Said Omar, Sun Yi, Samiullah Paracha, and Muhammad Wannous. Using iot technology to improve online education through data mining. In Applied System Innovation (ICASI), 2017 International Conference on, pages 515–
5. 518. IEEE, 2017.
6. Sebastian Scholze Claudio CenedeseOliviuMatei, Carmen Anton. Multi- layered data mining architecture in the context of the internetof things. In IEEE. IEEE, 2017.
7. Bhargava, B., Zhong, Y., & Lu, Y. (2003). Fraud Formalization and Detection. Proc. of DaWaK2003, 330-339.
8. Bentley, P., Kim, J., Jung., G. & Choi, J. (2000). Fuzzy Darwinian Detection of Credit Card Fraud. Proc. of 14th Annual FallSymposium of the Korean Information Processing Society.
9. Bolton, R. & Hand, D. (2001). Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control VII.
10. Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M. (2002). Fraud Classification using Principal Component Analysis ofRIDITs. Journal of Risk and Insurance 69(3): 341-371.
11. Burge, P. &Shawe-Taylor, J. (2001). An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Usersfor Use in Fraud Detection. Journal of Parallel and Distributed Computing 61: 915-925.
12. Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee based Evolution of Fuzzy Rules for the Detection ofSuspicious Insurance Claims. Proc. of GECCO2000.
13. Ezawa, K. & Norton, S. (1996). Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. IEEEExpert October: 45-51.