Application of Data Mining Techniques using Internet of Things

Sania Talha[1], Devunuri Sharanya[2], Dr V Sravan Kumar[3]
[1,2]Student of computer science & engineering at Balaji Institute of Technology & Science, Warangal, Telangana, India
[3]Associate Professor, Dept. of. CSE, Balaji Institute of technology & science, Warangal district, Telangana, India.

Received: 7 June 2021/ Revised: 18 June 2021/ Accepted: 25 June 2021/ Published: 30-06-2021

Abstract— *The generation and growing power of computer science have boosted data collection, storage, and manipulation as data sets are broad in size and complexity level. Internet of Things (IOT) is the most popular term in describing this new interconnected world. The massive data generated by the Internet of Things (IoT) are considered of high business value, and data mining algorithms can be applied to IoT to extract hidden information from data. As more and more devices connected to IoT, the latest algorithms should be applied to IOT. This paper explores a systematic review of various data mining models as well as its applications in the Internet of things along with its advantages and disadvantages*.

***Keywords— Internet of things (IOT), Data mining, Applications of Data mining*.**

I. INTRODUCTION

The term Internet of Things is 16 years old. But the actual idea of connected devices had been around longer, at least since the 70s Back then, the idea was often called “embedded internet “or “pervasive computing”. But the actual term “Internet of Things” was coined by “Kevin Ashton” in 1999 during his work at Procter & Gamble. The internet was the hottest new trend in 1999 and because it somehow made sense, he called his presentation “Internet of things (IOT)”. The Internet of things describes the network of physical objects “things” that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

One of the most valuable technologies is data mining. Data mining helps in discovering novel, interesting and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. Data Mining began in the 1990s and is the process of discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information.

In order to make IoT smarter, lots of analysis technologies are introduced into IoT; one of the most valuable technologies is data mining. Data mining overlaps with other fields like statistics, machine learning, artificial intelligence, databases but mainly it focuses on automation of handling large heterogeneous data, algorithm and scalability of number of features and instances research ions for in process of big data mining from IoT comes with its own set of challenges such as disparate datasets, large volumes of data, and the integrity of data sources. With the increasing popularity of IoT, new solutions and data mining algorithms are being developed to tackle such problems. [1,2]

On the basis of the definition of data mining and the definition of data mining functions, a data mining process includes the following steps:

- **Data preparation:** prepare the data for mining. It includes 3 steps: integrate data in various data sources and clean noise from the data; extract some parts of data into the data mining systems; pre-process the data to facilitate the data mining
- **Data Mining:** apply algorithms to the data to find the patterns and evaluate patterns of discovered knowledge.

Data presentation: visualize the data and represent mined knowledge to the user.

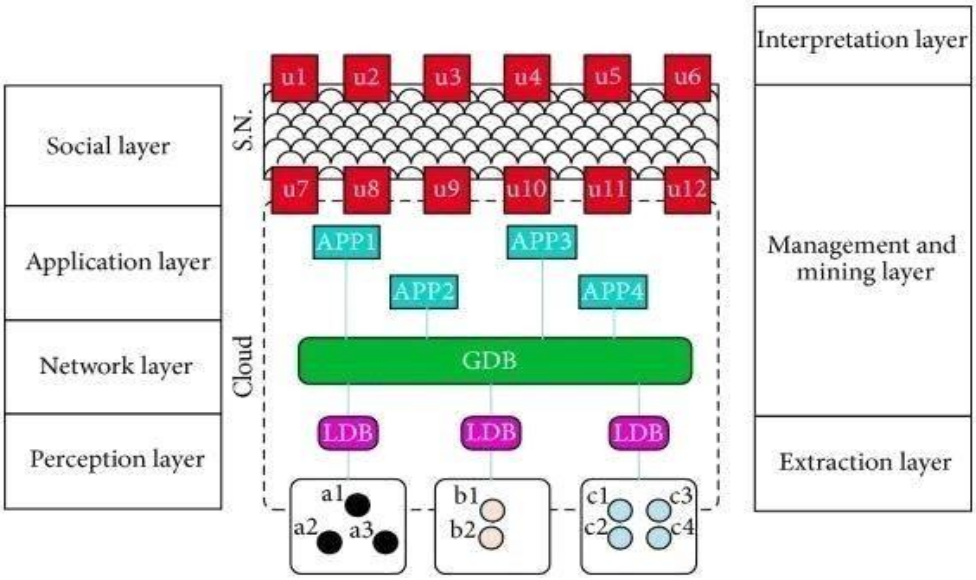

FIGURE 1: Architecture for data mining process

II. DATA MINING FUNCTIONALITIES

There are number of Data mining functionalities, they include:

- Characterization and discrimination
- Mining of frequent patterns, associations, and correlations.
- Classification and regression
- Clustering analysis
- Outlier analysis.

The key contribution of this paper includes:

- Introduction to IOT and Data mining.
- Process of Data mining.
- Functionalities of Data mining.
- Applications of data mining techniques in IOT.
- Advantages and Disadvantages

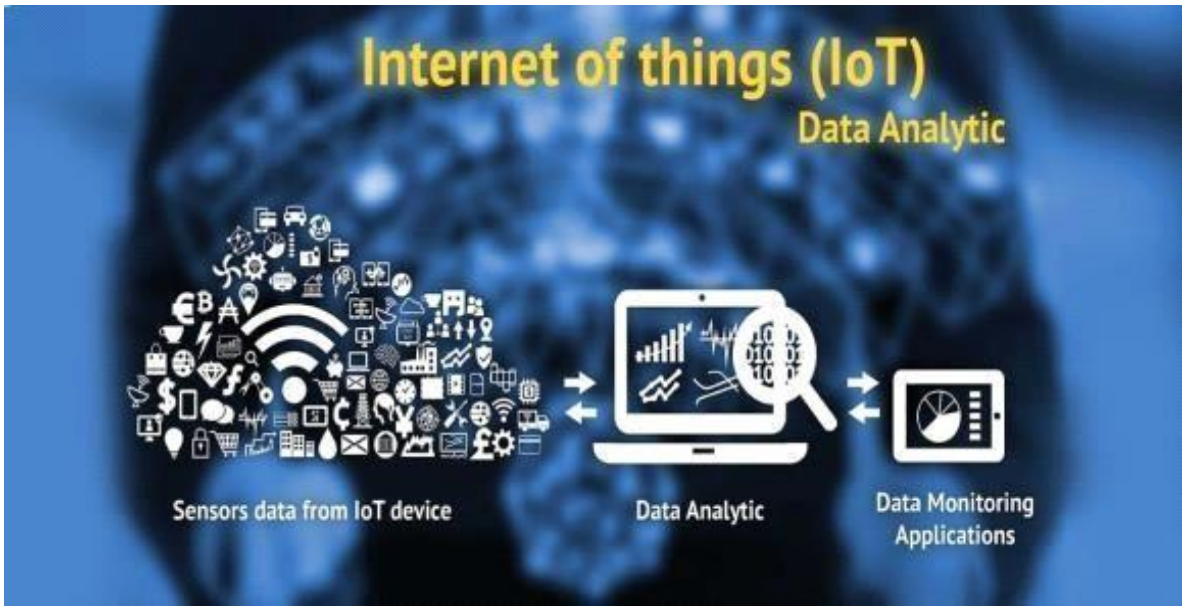

FIGURE 2: Data transfer through Internet of things (IOT)

Data Mining through IOT is primarily used today by companies with a strong consumer for retail, financial, communication, and marketing organizations, to drill down into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits.

With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments. [3,4]

III. DATA MINING TECHNIQUES IN FRAUD DETECTION IN CREDIT-DEBIT CARD TRANSACTIONS

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud Detections are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not. A fuzzy logic system incorporated the actual fraud evaluation policy using optimum threshold values. The result showed the chances of fraud and the reasons why an insurance claim is fraudulent. Another logic system used two approaches to imitate the reasoning of fraud experts, i) the discovery model, uses an unsupervised neural network to find the relationships in data and to find clusters, then patterns within the clusters are identified, and ii) the fuzzy anomaly detection model, which used Wang- Mendel algorithm to find how health care providers committed fraud against insurance companies.

Classification techniques have proved to be very effective in fraud detection and therefore, can be applied to categorize crime data. The distributed data mining model (Chen et al. 1999) uses a realistic cost model to evaluate CART, and naïve Bayesian classification models. The method was applied to credit card transactions. The neural data mining approach uses rule-based association rules to mine symbolic data and Radial Basis Function neural network to mine analog data. The approach discusses the importance of use of non-numeric data in fraud detection. It was found that the results of association rules increased the predictive accuracy. The Bayesian Belief Network (BBN) and Artificial Neural Network (ANN) study used the STAGE algorithm for BBN in fraud detection and back propagation for ANN. The different types of fraud detection are: internal, insurance, credit card, and telecommunications fraud detection. [6]

IV. BAYESIAN BELIEF NETWORK

Bayesian Belief Networks provide a graphic model of causal relationships on which class membership probabilities are predicted, so that a given instance is legal or fraud. Naïve Bayesian classification assumes that the attributes of an instance are independent, given the target attribute. The aim is to assign a new instance to the class that has the highest posterior probability. The algorithm is very effective and can give better predictive accuracy when compared to decision trees and back propagation However, when the attributes are redundant, the predictive accuracy is reduced. For the purpose of fraud detection and we construct two Bayesian networks to describe the behaviour of auto insurance. First, a Bayesian network is constructed to model behaviour under the assumption that the driver is fraudulent (F) and another model under the assumption the driver is a legitimate user (NF). The 'fraud net' is set up by using expert knowledge. The 'user net' is set up by using data from non-fraudulent drivers.[7]

During operation user net is adapted to a specific user based on emerging data. By inserting evidence in these networks (the observed user behaviour x derived from his toll tickets) and propagating it through the network, we can get the probability of the measurement x under two above mentioned hypotheses. This means, we obtain judgments to what degree observed user behaviour meets typical fraudulent or non-fraudulent behaviour. These quantities we call p(x|NF) and p(x|F). By postulating the probability of fraud P(F) and P(NF) = 1- P(F) in general and by applying Bayes' rule, we get the probability of fraud, given Journal of Digital Forensics, Security and Law, the measurement x,

$$P(F|x) = P(F)p(x|F)/ p(x)$$

Where, the denominator p(x) can be calculated as

$$P(x) = P(F)p(x|F) + P(NF)p(x|NF)$$

The chain rule of probabilities is: Suppose there are two classes C1, C2 for fraud and legal respectively. Given an instance X = (X1, X2, …, Xn) and each row is represented by an attribute vector A = (A1, A2, …, An) The classification is to derive the maximum P(Ci|X) which can be derived from Bayes' theorem as given in the following steps:

- P (fraud |X) = [P (fraud | X) P(fraud)] / P(X) P (legal |X) = [P (legal | X) P(legal)] / P(X)

  As P(X) is constant for all classes, only [P(fraud | X) P(fraud)] and [P(legal | X) P(legal)] need to be maximized.

- The class prior probabilities may be estimated by:

  P(fraud) = si / s

  Here, s is the total number of training examples and si is the number of training examples of class fraud.

- A simplified assumption of no dependence relation between attributes is made. Thus,

  P (X |fraud) = $\prod$= n k 1 P (x k |fraud) and

  P (X |legal) = $\prod$= n 1 P (x k |legal)

The probabilities P (x1 |fraud), P (x2 |fraud) can be estimated from the training samples:

P (x k |fraud) = si k / si

Here, si is the number of training examples for class fraud and si k is the number of training examples of class with value x k for Ak.

V. OUTPUT

We present Bayesian learning algorithm to predict occurrence of fraud. Using the "Output" classification results for Table 1. there are 17 tuples classified as legal, and 3 as fraud. To facilitate classification, we divide the age of driver attribute into ranges.

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	5	fraud
3	Jeremy Dejuan	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 Penson	M	36	0	0.66	6	legal
7	Collin Pyle	M	42	1	0.33	3	legal
8	Eric Penson	M	39	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 Vasconic	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 Dusek	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	?

Table 1 shows the counts and subsequent probabilities associated with the attributes. With these simulated training data, we estimate the prior probabilities:

The classifier has to predict the class of instance to be fraud or legal.

P(fraud) = si / s = 3/20 = 0.15

P(legal) = si / s =17/20 = 0.85

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

By using these values and the associated probabilities of gender and driver age, we obtain the following estimates:

P (X |legal) = 4/17 * 3/18 = 0.039

P (X |fraud) = 3/3 * 1/2 = 0.500

Thus, Likelihood of being legal = 0.039*0.9=0.0351

Likelihood of being fraud = 0.500 *0.1= 0.050

We estimate P(X) by summing up these individuals' likelihood values since X will be either legal of fraud:

P(X) = 0.0351 + 0.050 = 0.0851

Finally, we obtain the actual probabilities of each event:

P (legal | X) = (0.039 *0.9)/0.0851= 0.412

P (fraud |X) = (0.500 *0.1) / 0.0851= 0.588

Therefore, based on these probabilities, we classify the new tuple as fraud because it has the highest probability. Since attributes are treated as independent, the addition of redundant ones reduces its predictive power. To relax this conditional independence is to add derived attributes which are created from combinations of existing attributes. Missing data cause problems during classification process. [8,9]

Naïve Bayesian classifier can handle missing values in training datasets. To demonstrate this, seven missing values appear in dataset. The naïve Bayes approach is easy to use and only one scan of the training data is required. The approach can handle missing values by simply omitting that probability when calculating the likelihoods of membership in each class. Although the approach is straightforward, it does not always yield satisfactory results. The attributes usually are not independent. We could use subset of the attributes by ignoring any that are dependent on others. The technique does not handle continuous

data. Dividing the continuous values into ranges could be used to solve this problem, but the division of the continuous values is a tedious task, and this is done can impact the results.

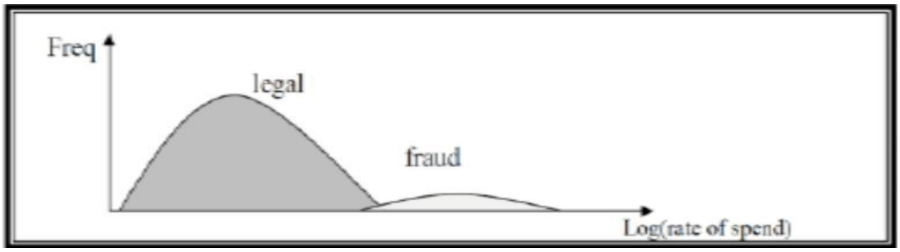

FIGURE 3: frequency distribution of legal and fraud transaction

VI. RESEARCH ANALYSIS

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The research can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequence and correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of data. Any technology available today has not reached its 100 % capability. It always has a gap to go. So, we can say that data mining through the Internet of Things has a significant technology in a world that can help other technologies to reach its accurate and complete 100 % capability as well.

6.1 Advantages of Mining through IOT:

Data mining has a lot of advantages when using it in a specific industry. Besides those advantages, e.g., Privacy, security, and misuse of information.

FIG 4: Industrial internet of things (IOT)

Data mining through the internet of things facilitates the several advantages in day-to-day life in the business sector. Some of its benefits are given below:

- **Efficient resource utilization:** If we know the functionality and the way that each device works, we definitely increase the efficient resource utilization as well as monitor natural. Minimize human effort: As the devices of data mining through IoT interact and communicate with each other and do a lot of tasks for us, then they minimize resources. The human effort.
- **Save time:** As it reduces the human effort then it definitely saves out time. Time is the primary factor which can save data mining through IoT platforms.
- **Enhance Data Collection:** Improve security: Now, if we have a system where all these things are interconnected then we can make the system more secure.

6.2 Disadvantages of Data Mining through IOT:

As the data mining through the Internet of things facilitates a set of benefits, it also creates a significant set of challenges. Some of the IoT challenges are given below:

- **Security:** As the data mining through IoT systems are interconnected and communicated over networks. The system offers little control despite any security measures, and it can lead the various kinds of network attacks.
- **Privacy:** Even without the active participation of the user, the system provides substantial personal data in maximum detail.
- **Complexity:** The designing, developing, and maintaining and enabling the large technology to system is quite complicated.

FIG 5: IOT offers Security to the system

VII. CONCLUSION

The output counts subsequent probabilities associated with the attributes. With these simulated training data, we estimate the prior probabilities. With simulated training data, we estimate the prior probabilities in fraud detection. Finally, we obtain the actual probabilities of each event. Due to seamless integration of classical networks with IOT. It enables a great vision that all things can be easily monitored and controlled which results in voluminous data. As a vital improvement of the next age of internet, the internet of things pulls in numerous considerations by industry world and scholarly circles. This makes the issue of information mining in IOT turn into a test process.

REFERENCES

[1] Mining with Big data: Jampalachaitanya, Fazi Ahmed parvez. International journal for technological research in engineering. Volume 4 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 515–518. IEEE, 2017.
[5] Sebastian Scholze Claudio CenedeseOliviuMatei, Carmen Anton. Multi- layered data mining architecture in the context of the internet of things. In IEEE. IEEE, 2017.
[6] Bhargava, B., Zhong, Y., & Lu, Y. (2003). Fraud Formalization and Detection. Proc. of DaWaK2003, 330-339.
[7] Bentley, P., Kim, J., Jung., G. & Choi, J. (2000). Fuzzy Darwinian Detection of Credit Card Fraud. Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society.
[8] Bolton, R. & Hand, D. (2001). Unsupervised Profiling Methods for Fraud Detection. Credit Scoring and Credit Control VII.
[9] Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M. (2002). Fraud Classification using Principal Component Analysis of RIDITs. Journal of Risk and Insurance 69(3): 341-371.
[10] Burge, P. &Shawe-Taylor, J. (2001). An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection. Journal of Parallel and Distributed Computing 61: 915-925.
[11] Bentley, P. (2000). Evolutionary, my dear Watson: Investigating Committee based Evolution of Fuzzy Rules for the Detection of Suspicious Insurance Claims. Proc. of GECCO2000.
[12] Ezawa, K. & Norton, S. (1996). Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts. IEEE Expert October: 45-51.