Fuzzy Systems and Soft Computing
ISSN : 1819-4362

EXPLORING THE APPLICATION OF DATA MINING TECHNIQUES IN INTERNET OF THINGS ENVIRONMENTS

[#1]SAILAYA THALLA,
[#2]AKHIL SILLA,
[#3]Dr.N.SRINIVAS, *Associate Professor*,
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

ABSTRACT: The advancement and spread of computer science has significantly enhanced data collection, storage, and analysis techniques. This progress has occurred in tandem with the rise and complexity of datasets. The term "Internet of Things" (IOT) is used to denote the expanding sphere of interconnectivity. The significant economic repercussions associated with the massive amounts of data created by the Internet of Things give them tremendous value. Data mining approaches are capable of extracting latent information hidden in data emanating from the Internet of Things (IoT). The increasing popularity of Internet of Things (IoT) devices emphasizes the significance of advanced algorithms. This research study examines various data mining approaches, their use in the context of the Internet of Things, and the advantages and disadvantages of each methodology

.KEYWORDS: Internetof things(IOT), Datamining, Applicationsof Datamining

1. INTRODUCTION

The notion of the "Internet of Things" has been around for over sixteen years. Nonetheless, networked devices have been around since the 1970s. The term "embedded internet" or "pervasive computing" was originally used to refer to the previously mentioned notion. Kevin Ashton, a renowned Procter & Gamble employee, coined the phrase "Internet of Things" in 1999. As a result, the organization has been praised for fully implementing this notion. The presenter properly titled his presentation "Internet of Things (IoT)" in reference to the World Wide Web's fascinating novelty when it was first introduced in 1999. The "Internet of Things" (IoT) is a networked infrastructure that connects everyday objects with electronic elements such as sensors and software. These devices can communicate with other devices and instruments via the internet.

Through R&D, data extraction has the potential to generate enormous profits. Data mining algorithms can extract hidden information from massive datasets, allowing for the discovery of unexpected, intriguing, and potentially significant patterns. In the 1990s, data mining emerged as a fresh field of study, leveraging computing power to unearth and identify unique, intriguing, and potentially valuable patterns inside massive databases.

Data mining is being used as an analytical tool to improve the intelligence of the Internet of Things (IoT). Databases, artificial intelligence, machine learning, and statistics are all part of data mining. The primary focus of this strategy, however, is on the development of algorithms to achieve scalability in terms of instances and attributes, as well as the mechanization of the analysis of enormous volumes of heterogeneous data. Managing massive amounts of data, authenticating data sources, and processing heterogeneous datasets are all difficulties associated with obtaining big data from the Internet of Things (IoT). As the Internet of Things (IoT) grows in popularity, new data mining methodologies and technologies are being developed to meet these challenges.

Data mining definitions and functions govern the phases of the data mining process.

Data preparation:

To prepare the data for mining, organize it logically. Data mining systems use a three-step procedure to mine data: collecting data from

diverse sources, decreasing data contamination, and extracting accurate data segments for preparation.

DataMining:

The use of computer techniques to analyze data in order to uncover latent associations. When datais visualized, users are given with graphical representations of the data and the information that may be gleaned from it.

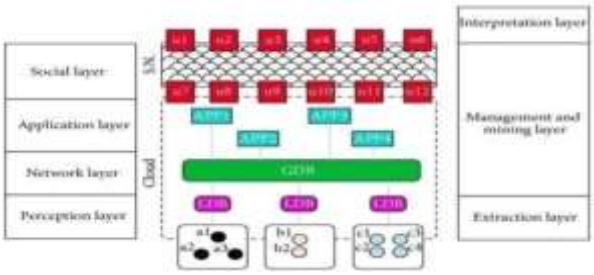

FIGURE1: Architecture for data mining process

DATA MINING FUNCTIONALITIES

- Data mining entails the description, extraction, and differentiation of recurring patterns, associations, and correlations.
- Analysis of outliers is a statistical technique that can be applied to both classification and regression investigations.

The keycontribution of this paper includes:

A data mining and Internet of Things overview. Techniques for Data Mining. Data mining's benefits. Applications of Data Mining in the World of the Internet of Things.

What are the benefits and drawbacks?

FIGURE2: DatatransferthroughInternetofthings(IOT)

- Data mining via the Internet of Things (IoT) is rapidly being used by businesses to establish pricing and product placement, study customer preferences, and evaluate revenue, profitability, and satisfaction. This is especially true for businesses in industries that rely heavily on customer satisfaction, such as banking, marketing, retail, and communication.
- A company can produce personalized merchandise and promotions for certain segments of its customers by using data mining algorithms to data collected at the point of sale (POS).

2.DATAMININGTECHNIQUESINFRAUDDETECTIONINCREDIT-DEBITCARDTRANSACTIONS

Significant financial damages have been incurred as a result of the fraud. Detection is a time-consuming and labor-intensive process. The practice of extracting relevant information from enormous databases by recognizing patterns is known as data mining. Knowledge includes any significant and proper information. All user data should be secure, with an unbreakable technique for detecting deception. Sample data aggregation is an example of supervised learning. The authenticity of these documents demonstrates their legality. Based on this data, an algorithm and model are developed to determine whether a record is false. A fuzzy logic technique with properly calibrated threshold values was used to incorporate the existing fraud screening procedure. The findings shed light on the probability of bogus insurance claims and the factors that contribute to them. In order to mimic the cognitive processes utilized by fraud professionals, an alternative logical system devised two strategies. The initial methodology is the discovery model, which employs an unsupervised neural network to detect patterns within data clusters and linkages. In the second method, the fuzzy anomaly detection model, the Wang-Mendel algorithm is utilized to assess how healthcare providers mislead insurance companies.

Classification algorithms are well-suited for the categorization of crime-related data due to their constant ability to detect fraudulent conduct. The distributed data mining model (Chen et al., 1999) employs a practical cost model to allow comparisons between the naive Bayesian classification algorithm and the CART algorithm. Every credit card transaction was handled in this

manner. The neural data mining technique captures both conceptual and analog information using a Radial Basis Function neural network and rule-based association algorithms. This method evaluates the ability of non-numerical data to detect fraudulent activity. It was discovered that applying association principles might greatly improve prediction accuracy. The STAGE algorithm has been utilized in research for the construction of both Bayesian Belief Network (BBN) and Artificial Neural Network (ANN); BBN utilizes it to detect fraud, while ANN employs it for back propagation. Internal fraud detection, credit card fraud detection, and insurance fraud detection are just a few of the various types of fraud detection.

3. BAYESIANBELIEFNETWORK

In Bayesian Belief Networks, the visual representation of causal relationships can be utilized to determine the likelihood of belonging to a specific class. This makes determining the truth or falsity of a particular case simple. When the target attribute is the focus, naïve Bayesian categorization is based on the assumption that the properties of an instance are independent of one another. The purpose is to add the most likely new instance to the class, as suggested by the posterior distribution. The proposed strategy increases projected accuracy while being far more efficient than decision trees and back propagation. Repeating qualities undermine the forecast's veracity. In order to detect likely incidences of motor insurance fraud, we employ two Bayesian networks to assess the attributes and patterns of auto insurance policies. The behavior is reflected by two Bayesian networks, one based on the assumption that the driver is a fraudulent individual (F) and the other on the assumption that the driver is a real user (NF). A "fraud net" is constructed by experts in their respective disciplines. The "user net" is made up of credible data sources. The text entered by the user is "[7]".

By leveraging newly received data, the user network is changed in real time to suit a particular user. The probability of the measurement x, according to the two hypotheses presented above, can be demonstrated by the public presentation of evidence within these networks, specifically the observed user behavior x as collected from their toll tickets. This demonstrates that it may be capable of determining the truth or falsity of observed user activity. They are represented by the values P[x|NF] and P[x|F]. The likelihood of fraud for a given measurement The Bayes rule and the assumption that P(F) represents the probability of fraud and P(NF) represents the probability of non-fraud can be used to compute P(F|x) = P(F)p(x|F)/ p(F).

P(x) = P(F)p(x|F) + P(NF)p(x|NF) can be used to compute the denominator p(x).

The rule of the probability chain is as follows: In this study, we will look at two distinct classes, C1 and C2, which represent legality and fraudulent behavior, respectively. The purpose of this research is to maximize P(Ci|X) in order to categorize an instance X = (X1, X2,..., Xn), with each row represented by an attribute vector A = (A1, A2,..., An). The following are the procedure steps:

In the equation, [P (fraud | X) P(fraud)] is the same as [P (fraud | X) P(fraud)]. The probability of event P divided by the probability of event X returns the probability of event P divided by the probability of event X.

Because P(X) remains constant across all classes, the primary goal should be to maximize [P(fraud | X) P(fraud)] and [P(legal | X) P(legal)].

P(fraud) = si/s is used to get the class prior probabilities.

In this scenario, s stands for the total number of training instances, whereas si stands for the number of training cases connected with the fraud class.

To idea that characteristics exist independently of one another is simplistic. P(X | legal) = n 1 P(x k | legal), but P(X | fraud) = n 1 P(x k | fraud).

P (x1 |fraud) and P (x2 |fraud) can be calculated using the training samples.

"si" and "si k" reflect the number of training samples for the classes "fraud" and "x k."

4.OUTPUT

Our organization provides a Bayesian learning technique for predicting fraud instances. Table 1 displays the categorization results for the "Output"

classification. Three tuples are incorrect, while the remaining 17 are correct. We divide the driver's age attribute into discrete intervals to speed up the categorization procedure.

TABLE1TRAININGSET

	Name	Gender	Age_driver	fault	Driver_rating	Vehicle_age	Output
1	David Oliyere	M	25	1	0	2	legal
2	Beau Jackson	M	32	1	1	7	fraud
3	Jeremy Dejean	M	60	0	0	7	legal
4	Robert Howard	M	35	1	0.33	3	legal
5	Crystal Smith	F	22	1	0.66	8	legal
6	Chibuike Preston	M	36	0	0.66	6	legal
7	Collin Pyle	M	42	1	0.33	3	legal
8	Eric Preston	M	19	1	1	2	fraud
9	Kristina Green	F	29	1	0	4	legal
10	Jerry Smith	M	33	1	1	9	legal
11	Maggie Fourier	F	42	1	0.66	3	legal
12	Justin Howard	M	21	1	0	2	fraud
13	Michael Vanvuin	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	9	legal
17	Aaron Drozk	M	40	1	0.33	8	legal
18	Bryan Sanders	M	49	1	0	3	legal
19	Derek Garrett	M	32	0	0	3	legal
20	Jeanine Jackson	F	27	0	1	2	legal
X	Crystal Smith	F	31	1	0	2	?

The characteristics are linked to the likelihood. Using this simulated training data, we compute the prior probability:

The classifier must determine if an instance is legitimate or fake.

TABLE2

PROBABILITIESASSOCIATEDWITHATTRIBUTES

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)	6	0	6/18	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	3	0	3/17	0
	1	12	3	12/17	3/17
driver rating	0	6	1	6/17	1/3
	0.33	3	0	3/17	0
	0.66	3	0	3/17	0
	1	3	2	3/17	2/3

Following the information provided and the probabilities associated with the driver's age and gender, the following approximations are calculated: A dependable source provides the probability of X's occurrence as 0.039; this value is calculated by dividing 4/17 by 3/18.

Multiplying 1/2 by 3/3 yields a probability of 0.500 (X | fraud). Consequently, the likelihood of legality is 0.0351, which is determined through the multiplication of 0.039 by 0.90.

The probability of deception is calculated by multiplying 0.1 by 0.500, which yields 0.050.

By summing the likelihood values of these individuals, the probability of occurrence X is represented as P(X). P(X) is calculated as the product of 0.050 and 0.0351, yielding 0.0851.

Finally, the exact probability of each occurrence is calculated:

The result obtained by multiplying 0.039 by 0.9 and dividing by 0.0851 is 0.412.

In order to classify the new tuple as fraudulent, the probabilities are utilized, given that it carries the highest likelihood. The predictive efficacy of repetitive features is diminished as a result of the perception that they are unique. By combining preexisting qualities to produce new ones, derived attributes facilitate the relaxation of the liberty restriction. The challenge of classification arises from the lack of available data.

The naïve Bayesian classifier possesses the ability to handle absent values in training sets. In order to represent this, the dataset is missing seven values. The naïve Bayes method is characterized by its simplicity and efficiency, as its implementation necessitates a single iteration over the training data. When calculating the probabilities for individual classes, the absence of any values is corrected for by completely disregarding the likelihood. Even in the case of an uncomplicated and easily understandable method, there is no assurance of positive results. In general, the attributes are interconnected. We could only put a limited number of the qualities to use; none of them were interdependent. Continuous data are inapplicable to this method. Dividing continuous data into intervals represents one possible resolution to this dilemma. However, the lengthy duration of this procedure could potentially impact the ultimate outcomes.

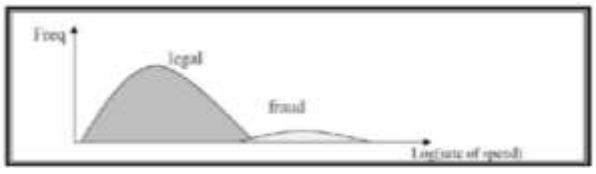

FIGURE 3:frequencydistributionoflegal andfraudtransaction

5. RESEARCHANALYSIS

Over time, profound innovations that radically modified existing paradigms have happened in the subject of research. Data mining is a powerful tool for executing processes such as database integration, data preparation, and data purification. The database contains possibly relevant data that could have an impact on the research. It is possible to determine the sequential organization and interrelationship of acts. Data visualization and visual data mining

provide an accurate and observable representation of the data. At this point, it is conceivable to argue that no technology has reached its entire potential. Every occurrence necessitates the fulfillment of a requirement. As a result, data mining via the Internet of Things is a vital technology in an area where its aid can bring other technological breakthroughs with unsurpassed precision and comprehensiveness

Advantages of Mining through IOT:

FIG4:Industrialinternetofthings(IOT)

- There are several benefits to adopting data mining in a particular organization. There are other advantages to consider, such as security, privacy, and the risk of information misuse.
- The internet of things makes data mining possible, which helps corporate operations in a variety of ways. Some of its benefits are as follows:

Efficient resource utilization:

A thorough grasp of each device's operation and attributes enables improved resource management and environmental monitoring. Reduce the amount of human labor required: By interacting and talking with one another, as well as completing a range of tasks on our behalf, data mining devices reduce the need for human labor. Human work

Save time:

It saves time since it eliminates the need for human work. IoT data extraction technologies have the potential to dramatically cut data processing time.

EnhanceDataCollection:

Implementing a network infrastructure that connects all of these components helps increase system security. The following drawbacks exist when using IoT for data mining:

Although data extraction via the Internet of Things offers many benefits, it also poses some new challenges. Here are a few examples of IoT-related problems:

Security:

Data mining for the Internet of Things (IoT) relies heavily on networking and data transmission across interconnected networks. Despite the implementation of security measures, the system has limited control and can be exploited to launch a range of network attacks.

Privacy:

Technology sends detailed and complete personal data due to the lack of human intervention.

Complexity:

Designing, building, managing, and enabling a system that mixes new technologies is a challenging process.

FIG5:IOToffersSecurity to the system

6.CONCLUSION

A list of the probabilities associated with the attributes is presented in the output. Using the simulated training data, the prior probability is determined. The initial likelihood of identifying fraudulent activity is determined by using simulated training data. Finally, the precise probability of each event is computed. Because of the seamless integration of traditional networks and the Internet of Things (IoT). It produces a holistic image in which every component is easily viewable and manipulable, resulting in massive volumes of data accumulation. The internet of things, a critical component for the internet's future, is causing widespread concern in academia and industry. As a result, the problem of data extraction in the context of the Internet of Things (IoT) is reframed as an investigative methodology.

REFERENCES

1. MiningwithBigdata:Jampalachaitanya,FaziAhmedparvez.Internationaljournalfortechnologicalresearchinengineering.Volume4 issuetooct2016.

2. DataMiningfortheInternetofThin:LiteratureReviewandChallengeS.Fengchen,Pandeng,Jiafu Wan,AthanasiosV.Vasilakos,Xiaohui.
3. SaralNigam,ShikhaAsthana,andPunitGupta.Iot basedintelligentbillboardusingdatamining.InInnovationandChallengesinCyberSecurity(ICICCS-INBUSH), 2016International Conferenceon pages107–110.IEEE,2016.
4. AlexanderMuriukiNjeru,MwanaSaidOmar,SunYi,SamiullahParacha,andMuhammadWannous.Using iottechnology toimproveonlineeducationthrough datamining.InAppliedSystemInnovation(ICASI),2017InternationalConferenceon,pages515518.IEEE,2017.
5. SebastianScholzeClaudioCenedeseOliviuMatei,CarmenAnton.Multi-layereddataminingarchitectureinthecontextofthe internetofthings. InIEEE.IEEE, 2017.
6. Bhargava,B.,Zhong,Y., &Lu,Y.(2003).Fraud Formalizationand Detection.Proc. ofDaWaK2003,330-339.
7. Bentley,P.,Kim,J.,Jung.,G.&Choi,J.(2000).FuzzyDarwinianDetectionofCreditCardFraud.Proc.of14thAnnualFallSymposiumoftheKoreanInformationProcessingSociety.
8. Bolton,R.&Hand,D.(2001).UnsupervisedProfilingMethodsforFraudDetection.CreditScoringand CreditControlVII.
9. Brockett,P.,Derrig,R.,Golden,L.,Levine,A.&Alpert,M.(2002).FraudClassificationusingPrincipalComponentAnalysisofRIDITs.Journal ofRiskandInsurance69(3): 341-371.
10. Burge,P.&ShaweTaylor,J.(2001).AnUnsupervisedNeuralNetworkApproachtoProfilingtheBehaviorofMobilePhoneUsersforUseinFraudDetection.Journal ofParallel andDistributed Computing61:915-925.
11. Bentley,P.(2000).Evolutionary,mydearWatson :InvestigatingCommitteebasedEvolutionofFuzzyRulesfortheDetectionofSuspiciousInsurance Claims. Proc. of GECCO2000.
12. Ezawa,K.&Norton,S.(1996).ConstructingBayesianNetworkstoPredictUncollectibleTelecommunicationsAccounts.IEEEExpert October:45-51.