IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Providing Privacy Using M-Diversity Algorithm along with K-Anonymity

Ashok Reddyboina[1], Pavan Koushik Batchu[2], Naveen Kumar Chennakesi[3]

Asst Prof, Department of Computer Science & Engineering, VVIT, Guntur[1]

UG Student, Department of Computer Science & Engineering, VVIT, Guntur[2,3]

Abstract: The data of any person can be divided into two categories. They are non-sensitive data and sensitive data. Coming to banking details, the bank’s database consists of large dataset regarding customers of the bank and their transactional details which are treated as most sensitive data. Any Intruder can get the sensitive data of the customer even the customer’s personal information is removed.

Interlopers utilize a few information mining methods to recover the information and break down it and utilize the information for their own motivation. Gatecrashers can get the subtle elements of the managing an account clients through Quasi Identifier esteem which results to the homogeneity assault and membership disclosure assault. So to keep this information assaults this paper proposes new strategy which saves the security among the client's information through randomizing the client's information based on Quasi-Identifier.

Keywords: Anonymization, Regression, K-Anonymity, M-Diversity.

I. INTRODUCTION

Banking allows changing the behavior according to the customers. Banking data will be given to an organization such as address, phone numbers, profiles, visits the websites to gain the knowledge of the customers. Lending activities may be performed either directly or indirectly. Today banks are highly regulated in most countries. Banking is an area in which it contains large amount of data. Every bank concentrates to customer’s profitability. Most activities are done based on customer grade and their creditability. Customer profiling consists of customer details, customer attractiveness, and satisfaction. This information can be acquired from financial balance exchanges, credit applications, and advance reimbursements. In this manner, it keeps up tremendous measure of client data, Visa utilization example, exchanges etc. Client Relationship alludes to the strategies and devices that can help bank administrators to oversee client connections in a productive way; information mining is primarily used to process the information investigation and displaying procedures to develop connections and examples in information that can be utilized to make precise forecasts.

Successfully supporting managerial decision-making is depends upon the availability of integrated, quality of information organized and presented in a timely and easily understandable way. Some of the data mining techniques to solve business problems like classification, time series, clustering, association rules, regression and discovery of sequence patterns. Clustering can be used for either forecasting or description. Segmentation helps to better understand customer preferences and allocate resources based on the information. It helps banks to where their most regular customers are located and helps to allocate time and effort to generate the most profit.

Data mining techniques like k-means clustering can be applied to get the best investments based on customer’s profile. Privacy is the biggest challenge in banking. The privacy depends upon not revealing customer information to the third party. They don't uncover the data about their clients unless certain conditions are fulfilled. Security relies on not to uncover the client characters and can stay mysterious even after information mining testing. It was intended to lessen the level of hazard in basic leadership to a base through examination of existing individual credit clients. To accomplish this by making mysterious engineering to deal with the client data. Information examination procedures were arranged towards separating attributes of quantitative and measurable information. These methods encourage valuable information elucidations for the saving money division to maintain a strategic distance from client wearing down.

The keeping money area utilizes the idea of Management Information System to produce various types of reports which can be utilized and investigated for basic leadership. There is different information mining systems to recover required information from those immense databases. Diverse levels of idealistic outcomes can be gotten for suspicious exchange discovery by utilizing the systems from bunching to grouping. Banks often offers investment services to their

DOI 10.17148/IJARCCE.2018.7325

customers. There are a vast number of financial instruments in the market. Data mining techniques like clustering can be applied to select the correct investments based on account holders details. In the financial area, data mining used to determining the authorized customer for loan disbursement, and finding profitable customers, products, characterizing different product segments. These factors are used in business factors and forcing banks to consider reinventing themselves to win in the marketplace. Data mining techniques will be used in banking sector to find hidden patterns and discover unknown relationship in the data. Data mining helps business analyst to generate hypotheses but it does not validate the hypotheses.

II. PROBLEM STATEMENT

Large data and other information can be separated utilizing a few information mining procedures which in additionally utilized for breaking down public policy patterns and detailing open strategies.
In the event that the information uncovered demonstrates the specific individual's private data then that information is known as sensitive information. To give protection to the individual's private information, the information security saving system is utilized. Here K-anonymity utilizes clustering method where the comparable sort of information is assembled together and as a result of absence of diversity in the sensitive properties of the customer's, the real information can be anticipated by the interloper and may profit by the acquired information.
As the information is assembled together there are two assaults that can be utilized to get the real data of the banking clients. The assaults are homogeneity attack and membership disclosure attack. To beat this assaults new strategy known as M-diversity is utilized. A database is said to be l-diversified if and only if each cluster of the information comprises 'l' novel esteems for sensitive characteristics.
The past method known as k-anonymity gave security to information up to some degree and it doesn't have any learning on the previously mentioned attacks. New strategy L- Diversity for the most part centers on setting up protection up to greatest and this uses effective preprocessing method. At the point when the information is preprocessed then the superfluous information and excess information is evacuated and resultant information is most effective one.

III. RELATED WORK

Area information is utilized to discover better business manages effectively be created with a control based misrepresentation recognition and aversion framework for keeping money applications. Areas data can likewise be utilized as an elective verification metric notwithstanding secret key, security token and biometric measures. Unsupervised and directed strategies can be utilized for displaying false occasions [13].
Strategic relapse, neural systems, Bayesian conviction systems, choice trees and guileless Bayes are the main managed techniques utilized [1]. The overview learns about irregularity discovery Chandola et al. what's more, Gupta et al. additionally detailed that misrepresentation identification issue is tied in with recognizing abnormalities in discrete successions and the related information groupings [2], [3]. One of the mainstream methods which are connected to money related hazard administration is calculated relapse [4], [5]. It has been connected to misrepresentation recognition [6]. Strategic relapse is a sort of class with summed up direct models that connection the resultant variable to a straight indicator through a capacity called log ().
Information mining method, for example, Support Vector Machines used to anticipate the client esteem rate. This model additionally utilized as a part of hazard administration to help the records receivable. For the most part, a choice help demonstrate is utilized as a part of extensive banks to quantify the dangers in advance beneficiaries. A credit scoring model is displayed to survey account credit value. A calculated relapse model can be utilized to investigate debt claims hazard. The aftereffect of this model can be acquired and empowering records of sales chiefs to apply measurable investigation through information mining to deal with their hazard.

Choice trees are for the most part utilized as a part of monetary hazard administration [7], [8]. It speaks to a division of the information that is gotten by applying an arrangement of guidelines. The standards are connected consistently, bringing about chain of command of section inside portions. In [9], a cross breed procedure proposed which consolidates both grouping and characterization calculations to recognize exceptions from review logs. In [10], a scientific approach was proposed for money related exchanges to check the database review logs to stamp speculated exchanges assuming any. [14] Aimed to recognize the illicit client exchanges in banks that can prompt extortion by separating the conduct as non-suspicious and suspicious utilizing the grouping approach. In [15], a neuro-fluffy framework was proposed to identify unsafe clients in managing an account framework by thinking about their behavioral characteristics. Impartial systems used to make forecasts about certifiable issues utilizing nonlinear connections in information.

DOI 10.17148/IJARCCE.2018.7325

IV. PROPOSED METHODOLOGY

A. FEATURE SELECTION

Highlight Selection By the quick expanding of data innovation the part of the system administration and system activity assumes indispensable part. This prompts new sorts of assaults and interruptions and cause to framework disappointments and unapproved get to.

We concentrated on interruption recognition framework by choosing the valuable highlights by dispensing with the excess highlights. A significant number of the highlights have little significance amid the location procedure and these impacts computational proficiency amid testing and preparing the informational index. In this way, choosing the helpful highlights by the element pertinence investigation is principle essential advance for the entire framework.

IM is the measure which is normally utilized in building a choice tree. We utilize IM to remove the valuable highlights by separating the interruption and ordinary class. For a predetermined assault compose, the element with the most astounding IM is dealt with as the most pertinent component and which assume a key part in deciding the assault class.

B. INFORMATION MEASURE (M)

Here X is set of features within a data set and assumes Y is a set of attack classes. Consider relevance expressed as Information Measure (IM) between X and Y. Where P(X, Y) expressed as joint probability function and P(X) and P(Y) are the marginal density of X, Y respectively. The marginal density functions are

$$\mathrm{P(X)} = \int_{Y}^{X} P_{\mathrm{X}}\,\mathrm{Y}\ \mathrm{(Y, X)}\ dy \qquad (1)$$

$$\mathrm{P(Y)} = \int_{Y}^{X} P_{\mathrm{X}}\,\mathrm{Y}\ \mathrm{(Y, X)}\ dx \qquad (2)$$

Now combining the equation (1) & (2), we get IM as follows:

$$\mathrm{IM(X; Y)} = \int_{X}^{Y}\int_{Y}^{X} P_{\mathrm{X,}}\ \mathrm{Y(X, Y)\ dy\ log\ P_{X}Y(X, Y)\ dy} \qquad \text{-------- (3)}$$

Several features still carry same information. For wiping out those highlights we can ascertain the separation between the highlights and the classes with Normalized Information Measure (NIM). Than the condition of NIM inferred as:

$$\mathrm{NIM(X; Y)} = \frac{2IM(X,Y)}{P(X)+P(Y)} \qquad (4)$$

The value of NIM lies between 0<=NIM<=1. The component which contains NIM esteem is 0 is dealt with as repetitive element and the element which contains the NIM esteem 1 is dealt with as trustworthy element as per the regard assault class. Highlight IM for each class is ascertained by the over four equations. By utilizing the above recipes we recoil the database.

The dataset which we had considered is taken from UCI repository. The dataset consists of 17 attributes and some attributes in that dataset are unnecessary. The original database has attributes such as age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y1. These are reduced and listed below.

TABLE – I

S. No.	Attribute Name	Attribute Type
1	Age	Non-sensitive
2	Education	Non-sensitive
3	Job	Sensitive
4	Balance	Sensitive
5	Housing loan	Sensitive

DOI 10.17148/IJARCCE.2018.7325

6	Personal loan	Sensitive
7	Annual income	Sensitive
8	Assets	Sensitive
9	Proofs	Sensitive
10	Address	Non-sensitive

C. M- DIVERSITY

M-diversity is an efficient privacy technique that can be easily implemented. A class is said to be M- diversified if and only if there consists of 'm' different values for the sensitive information in the dataset. The main idea of the M-diversity is to randomize the data in a well-represented manner. As this technique supports monotonicity property frequent data mining algorithms can't analyze the sensitive data.

M-diversity algorithm:

1. Read the input file
2. Identify clusters in the dataset(based upon Non-sensitive data)
3. Identify ranges of Quasi-Identifier (QI) value
4. Let 'S' be a sensitive attribute in the dataset and check the ranges
5. If 'S' are same in a group then rearrange the QI value ranges
6. Apply step-5 until sensitive attributes in a group are diversified

The below table gives the micro data of the banking customers with attributes such as age, loan, balance, personal proofs.

TABLE – II: Banking Customer Data

S. No	**Age**	**Loan**	**Balance**	**Personal proofs**
1	20	Yes	2000	Voter ID
2	20	Yes	3000	Passport
3	24	Yes	1000	Aadhar
4	21	Yes	5000	Driving License
5	26	No	7000	Passport
6	26	Yes	10000	Voter ID
7	27	No	25365	Aadhar
8	29	No	222555	PAN
9	32	No	54663	PAN
10	31	No	254100	Driving License
11	34	No	80000	Voter ID
12	32	No	9000	Passport
13	23	Yes	5000	PAN
14	28	Yes	100000	Driving License
15	33	No	8000	Aadhar

When k=5 and 5-anonymous table that is formed and in this anonymous technique we apply generalization and suppression techniques.

DOI 10.17148/IJARCCE.2018.7325

TABLE – III: 5- Anonymous Customer Data

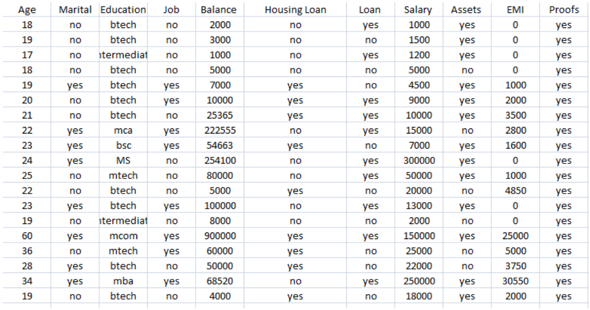
S. No	Age	Loan	Balance	Personal Proofs
1	20*	Yes	2000	Aadhar
2	20*	Yes	3000	Voter ID
3	20*	Yes	1000	PAN
4	20*	Yes	5000	Driving License
5	20*	Yes	5000	Passport
6	25*	No	7000	Voter ID
7	25*	Yes	10000	Passport
8	25*	No	25365	PAN
9	25*	Yes	222555	Aadhar
10	25*	No	100000	Driving License
11	30*	No	54663	Driving License
12	30*	No	254100	PAN
13	30*	No	80000	Aadhar
14	30*	No	9000	Passport
15	30*	No	8000	Voter ID

V. RESULTS

As this technique is an algorithm which can be implemented using java language and JVM. This algorithm takes a CSV file (comma separated file) as input and gives output as CSV file and in between this algorithm is implemented to give effective and efficient result.

Input File:

This file consists of preprocessed data where the external and unnecessary details are removed in preprocessing technique.

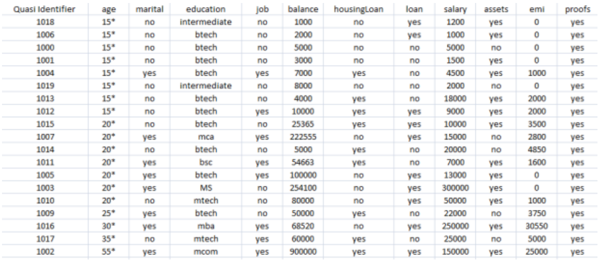


Age	Marital	Education	Job	Balance	Housing Loan	Loan	Salary	Assets	EMI	Proofs
18	no	btech	no	2000	no	yes	1000	yes	0	yes
19	no	btech	no	3000	no	no	1500	yes	0	yes
17	no	ıtermediat	no	1000	no	yes	1200	yes	0	yes
18	no	btech	no	5000	no	no	5000	no	0	yes
19	yes	btech	yes	7000	yes	no	4500	yes	1000	yes
20	no	btech	yes	10000	yes	yes	9000	yes	2000	yes
21	no	btech	no	25365	yes	yes	10000	yes	3500	yes
22	yes	mca	yes	222555	no	yes	15000	no	2800	yes
23	yes	bsc	yes	54663	yes	no	7000	yes	1600	yes
24	yes	MS	no	254100	no	yes	300000	yes	0	yes
25	no	mtech	no	80000	no	yes	50000	yes	1000	yes
22	no	btech	no	5000	yes	no	20000	no	4850	yes
23	yes	btech	yes	100000	no	yes	13000	yes	0	yes
19	no	ıtermediat	no	8000	no	no	2000	no	0	yes
60	yes	mcom	yes	900000	yes	yes	150000	yes	25000	yes
36	no	mtech	yes	60000	yes	no	25000	no	5000	yes
28	yes	btech	no	50000	yes	no	22000	no	3750	yes
34	yes	mba	yes	68520	no	yes	250000	yes	30550	yes
19	no	btech	no	4000	yes	no	18000	yes	2000	yes

As this file is taken as input to apply further processing. This input is sorted based on the age (non sensitive attribute). After sorting, the sorted file is considered as input to give output as generalized and suppressed file taken as file3.

In the next step, Quasi-Identifier values are generated and these are randomized. Now these values are merged with file3. The below figure is the output after applying M-diversity.

Output File:



Quasi Identifier	age	marital	education	job	balance	housingLoan	loan	salary	assets	emi	proofs
1018	15*	no	intermediate	no	1000	no	yes	1200	yes	0	yes
1006	15*	no	btech	no	2000	no	yes	1000	yes	0	yes
1000	15*	no	btech	no	5000	no	no	5000	no	0	yes
1001	15*	no	btech	no	3000	no	no	1500	yes	0	yes
1004	15*	yes	btech	yes	7000	yes	no	4500	yes	1000	yes
1019	15*	no	intermediate	no	8000	no	no	2000	no	0	yes
1013	15*	no	btech	no	4000	yes	no	18000	yes	2000	yes
1012	15*	no	btech	yes	10000	yes	yes	9000	yes	2000	yes
1015	20*	no	btech	no	25365	yes	yes	10000	yes	3500	yes
1007	20*	yes	mca	yes	222555	no	yes	15000	no	2800	yes
1014	20*	no	btech	no	5000	yes	no	20000	no	4850	yes
1011	20*	yes	bsc	yes	54663	yes	no	7000	yes	1600	yes
1005	20*	yes	btech	yes	100000	no	yes	13000	yes	0	yes
1003	20*	yes	MS	no	254100	no	yes	300000	yes	0	yes
1010	20*	no	mtech	no	80000	no	yes	50000	yes	1000	yes
1009	25*	yes	btech	no	50000	yes	no	22000	no	3750	yes
1016	30*	yes	mba	yes	68520	no	yes	250000	yes	30550	yes
1017	35*	no	mtech	yes	60000	yes	no	25000	no	5000	yes
1002	55*	yes	mcom	yes	900000	yes	yes	150000	yes	25000	yes

REFERENCES

DOI 10.17148/IJARCCE.2018.7325

[1] Christine Dartigue, Hyun Ik Jang and Wenjun Zeng, "A New Data-Mining Based Approach for Network Intrusion Detection," IEEESeventh Annual Communication Networks and Services Research Conference, pp. 372-377, May 2009.

[2] Freitas, Alex A, "Data mining and knowledge discovery withevolutionary algorithms," Springer, 2002.

[3] HG Kayacik, A N Zincir-Heywood, M I Heywood, "Selecting Featuresfor Intrusion Detection: A Feature Relevance Analysis on KDD 99Intrusion Detection Datasets," Third Annual Conference on Privacy,
Security and Trust, pp. 3-8, Oct. 2005.

[4] Al-Sharafat, W.S.Naoum, R.Al Al-Bayt Univ. and Jordan, “Significant of features selection for detecting network intrusions," IEEE ICITST:International Conference for Internet Technology and SecuredTransactions, pp. 1-6, 2009.

[5] AnirutSuebsing, NualsawatHiransakolwong, "Feature Selection UsingEuclidean Distance and Cosine Similarity for Intrusion DetectionModel," ACIIDS: First Asian Conference on Intelligent Information and
Database Systems, pp. 86-91, April 2009.

[6] Chou, Te-Shun Yen, Kang K. Luo, and Jun Pissinou, NikiMakki and Kia, "Correlation-Based Feature Selection for Intrusion Detection Design,"IEEE MILCOM: Military Communications Conference, pp. 1-7, 2008.

[7] Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok, "A Data MiningFramework for Adaptive Intrusion Detection," IEEE Symposium onSecurity and Privacy, pp.120-132, 1999.

[8] Yiming Yang and Jan O. Pedersen, "A Comparative Study on FeatureSelection in Text Categorization," In Proceedings of the FourteenthInternational Conference on Machine Learning (ICML '97), Douglas H.Fisher (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412-420, 1997.

[9] P Garciateodoro, J Diazverdejo, G Maciafernandez, E Vazquez,"Anomaly-based network intrusion detection: Techniques,systems andchallenges," Computers & Security, Elsevier, vol. 28,Issue 1-2, pp. 18-28, Feb.-Mar. 2009.

[10] Christina Leslie, EleazarEskin and William Stafford Noble, "TheSpectrum kernel: A string kernel for SVM protein classification," Procsof the Pacific Symposium on Biocomputing, pp. 564-575, January 2-7,2002.

[11] G. Fung, O. L. Mangasarian, "A Feature Selection Newton Method forSupport Vector Machine Classification," Computational Optimizationand Applications, Volume 28, Issue 2, Pp. 185-202, July 2004.

[12] Isabelle Guyon, JhonsonWestren and Vladimir Vapnik,"Gene selectionfor cancer classification using support vector machines," MachineLearning, Vol. 46, pp. 389-422, 2002.

[13] EmreÇomak, Ahmet Arslan,"A new training method for support vectormachines: Clustering k-NN support vector machines," Expert SystemAppl. Volume 35, Issue 3, pp. 564-568, 2008.

[14] Pingjie Tang, Rong-a Jiang, and MingweiZhao. 2010. FeatureSelection and Design of Intrusion Detection System Based on k-Meansand Triangle Area Support Vector Machine. In Proceedings of the 2010
Second International Conference on Future Networks (ICFN '10). IEEEComputer Society, Washington, DC, USA, 144-148.DOI=10.1109/ICFN.2010.4

[15] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” Decision Support Systems, vol. 50, no. 3, pp. 559 – 569, 2011, on quantitative methods for detection of financial fraud.