International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Multigroup Classification using Privacy Preserving Data Mining

Alfian Erwinsyah, Jacky Chin, Irfan A. Palalloi, Phong Thanh Nguyen, K. Shankar

Abstract— Getting the useful and important data from a huge amount of information is known as data mining. It is a prominent field for search and research of data. To improve the communication between customers and organizations data mining is used. Preserve the private data is very necessary in data mining. It is the issue on which research developed their research in many different ways. For protecting the survey privacy and to avoid the bias answer the randomized response technique was developed. To prevent the data of the survey certain randomness will add with the answers. To improve the privacy level of the preserve data this research use multigroup methods. In this approach all the survey answer divided in multiple groups and then. For different groups data should randomize differently. Based on this multiple groups the decision tree used to classify the data. One group, two group and three group's techniques used to preserve the data.

Keywords: Data mining; randomized response; multi groupt; decision tree.

I. INTRODUCTION

A process in which data is gathered from many different sources and then convert this data in useful and important data is known as process of data mining. It is also called as KDD i.e. knowledge discovery in database. To predict the approaches of future, data mining process automatically receive the relevant patterns in large data set to get the existing and previous data. The tool of data mining can predict the behavior and trends of future so by using these tools the organization can make inactive and knowledge-driven decisions. Also it allows answering that questions quickly that takes too much time before. For analyzing the data the data mining software have several analytical tools [1]. The KDD is refers for Knowledge Discovery in Databases. It is the process to finding the useful information or knowledge from data. And it provides high level applications for methods of data mining. This process helps the scientists to provide the research in pattern recognition, data visualization, artificial intelligence, knowledge acquisition, statistics etc.

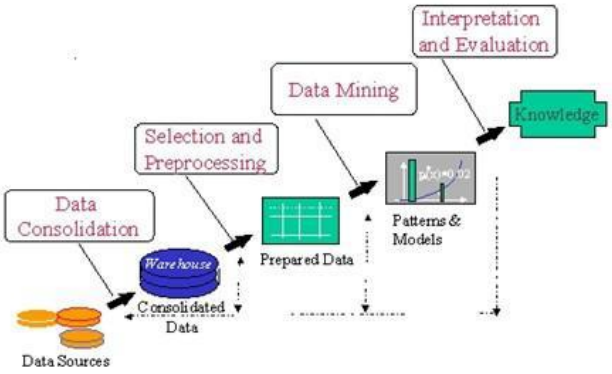

Figure 1: knowledge discovery process data mining steps[3]

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
Alfian Erwinsyah, Faculty of Tarbiyah and Teacher Training IAIN Sultan Amai Gorontalo, Indonesia.
Jacky Chin, Mercu Buana University, Indonesia.
Irfan A. Palalloi, Universitas Sulawesi Barat, Indonesia.
Phong Thanh Nguyen, Department of Project Management, Ho Chi Minh City Open University, Vietnam.
K. Shankar, Department of Computer Applications Alagappa University, India.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection(1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table 1: Data Mining Development

II. THE DATA MINING PROCES

To utilize and identify the hidden data in data mining there require three conditions [3]:

- The data that is finding from hidden data should have wide view rather than specific view.
- The data that is integrated data should be extracted.
- The data which get from hidden data should be organized in the way so it can use for decision making.

The process of data mining divided in to four steps. The already summarized data that find from data warehouse consist the transform of the information. And they use to provide the useful data. The process of data mining include following steps as given below [4]:

1. Selection of data
2. Transformation of data
3. Mining that data
4. Result interpretation

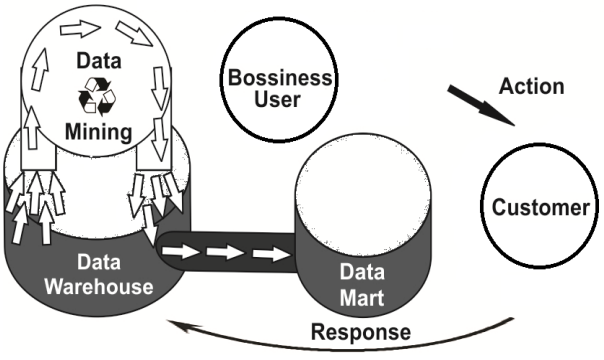

Figure 2: Process of Datamining

III. DATA MINING PRIVACY ISSUES

In many applications, it is consider that for device empowering disclosure of helpful patterns data mining is ground-breaking tool. As there is large data warehouse is available and it related to different approaches it is very necessary to preserve the information in many conditions, for example information about condition of patient, individual foundation data and client inclinations etc. It unavoidably creates private data of the client if it uncovers the private original data. In this way the main goal of data mining is to discover the approaches in which data can get with privacy preservation. Privacy preserving data mining use for solve this problem change the original data there are many privacy preserving techniques are used. The privacy preserving data mining measured on the basis of metrics of privacy protection, computation, accuracy and applicability. [5].

IV. DECISION TREES AND THE ID3 ALGORITHM

The decision tree contains edges and nodes. Decision tree is a rooted tree. In the decision tree every internal node is consider as test node and it relevant to an attribute. The edge of tree that leaves a node associate with the value that is taken from that attribute. For example if there is a attribute called home owner. form this tree there are two edges that leave the tree. One node is for "yes" and other node is for "no". For transactions matching a path from root to leaf the leaves of the tree contain the expected value. [6].

The concept of ID3 algorithm based on that each attribute contain discrete data and it divided in categories, with representation of continues data [7].
The ID3 algorithm is given below. In this algorithm tree is created in top down approach in recursive manner. Each attribute at the root check that how attribute can classify the transaction. In these attributes best attributes are chosen and the rest of the transactions were divided by it. After that on each partition ID3 is called recursively [8].

ID3(R,C,T)

1. If R is empty, return a leaf-node with the class value assigned to the most transactions in T.

2. If T consists of transactions which all have the same value c for the class attribute, return a leaf-node with the value c (finished classification path).

3. Otherwise,

(a) Determine the attribute that best classifies the transactions in T, let it be A.

(b) Let a1,….,am be the values of attribute A and let T(a1),…,T(am) be a partition of T such that every transaction in T(ai) has the attribute value ai.

(c) Return a tree whose root is labeled A (this is the test attribute) and has edges labeled a1,….,am such that for every i, the edge ai goes to the tree ID3(R-{A},C,T(ai)).

Example of ID3 algorithm [9]

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

V. RANDOMIZED RESPONSE TECHNIQUES

To protect the data of survey researches developed a technique called Randomized Response (RR) techniques. This technique helps to protect the data that is based on privacy. This method mainly avoids the bias answer. In 1965 Warner developed the method in which measure the people percentage in the population that has the specific attributes. In this method the respondent cannot give the incorrect answer or they will not give the answer [10].

VI. ONE-GROUP METHOD

As the name implies the in one group method all the attribute consider in one group. And all attributes stay together by getting the same value. For example when private data send to the central data then each answer of all the members are same. There are two options whether all tell the truth of all the questions or they tell lie of all the questions. It is shown that the probability of telling the truth about all the answers is θ and the probability of telling the lie about all the answers is (1-θ). For example a truth value of a user is A1,A2, and A3 i.e. 110. User creates a random number from 0 to 1. If the generated number is less than θ then the user send 001 to the collector. It shows the user is telling truth. If the generated number is greater than θ then the user sends 001. It shows the user is telling lie. To represent it we use P(001) as follows:

$$P(A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 0)$$

to present

$$P(A_1 = 0 \wedge A_2 = 0 \wedge A_3 = 1).$$

Some involvement of P*(110) and P*(001) is come from P (110) and some involvement come From P (001), the following equation derived [11]:

$$P^*(110) = P(110).\theta + P(001).(1-\theta)$$
$$P^*(001) = P(001).\theta + P(110).(1-\theta)$$

From above equation it will provide P(110). This is the information that requires to create the decision tree. The basic one group model is described as follows:

$$P^*(E) = P(E).\theta + P(\overline{E}).(1-\theta)$$
$$P^*(\overline{E}) = P(\overline{E}).\theta + P(E).(1-\theta)$$

The coefficiency matrix of above equation is shown below:

$$\begin{pmatrix} p^*(E) \\ P^*(\overline{E}) \end{pmatrix} = M_1 \begin{pmatrix} P(E) \\ P(\overline{E}) \end{pmatrix}, \text{ where } M_1 = \begin{bmatrix} \theta & (1-\theta \\ 1-\theta & \theta \end{bmatrix}$$

VII. TWO-GROUP METHOD

In the technique of one group the data collector or interviewer know that there is only one possibility of the answer for all the attributes that whether the respondent tell truth or lie [11-13]. So data privacy is in danger for this type of technique [14-18]. In this way to enhance the level of privacy the data will divide in two groups. After divide the groups, the randomized response techniques will apply on each group. In this scheme a one group can tell the truth and other group can tell the lie. If the data collector know answer of one group then it is possible to not have the information about the answer of another group. In this way privacy level improved in this scheme as compared to one group scheme.

Here it takes P* (E1 E2). To show the evaluation of P(E1 E2). The contribution of P*(E1 E2) consider in four parts. So it derived the equation given below:

$$P^*(E_1\ E_2) = P(E_1\ E_2).\theta^2 + P(E_1\ \overline{E_2}).\theta(1-\theta) + P(\overline{E_1}\ E_2).\theta(1-\theta) + P(\overline{E_1}\ \overline{E_2}).(1-\theta)^2.$$

As shown in the equation there are 4 variables that are not known

$$(P(E_1\ E_2), P(E_1\ \overline{E_2}), P(\overline{E_1}\ E_2), P(\overline{E_1}\ \overline{E_2}))$$

We need 3 more equations to solve the above equations. It can derive by suing the same approach:

$$\begin{pmatrix} P^*(E_1\ E_2) \\ P^*(E_1\ \overline{E_2} \\ P^*(\overline{E_1}\ E_2) \\ P^*(\overline{E_1}\ \overline{E_2}) \end{pmatrix} = M_2. \begin{pmatrix} P(E_1\ E_2) \\ P(E_1\ \overline{E_2}) \\ P(\overline{E_1}\ E_2) \\ P(\overline{E_1}\ \overline{E_2}) \end{pmatrix}$$

The derived matrix is as follows:

$$\text{Where } M_2 = \begin{bmatrix} \theta^2 & \theta(1-\theta) & \theta(1-\theta) & (1-\theta)^2 \\ \theta(1-\theta) & \theta^2 & (1-\theta)^2 & \theta(1-\theta) \\ \theta(1-\theta) & (1-\theta)^2 & \theta^2 & \theta(1-\theta) \\ (1-\theta)^2 & \theta(1-\theta) & \theta(1-\theta) & \theta^2 \end{bmatrix}$$

VIII. THREE-GROUP METHOD

$$\begin{pmatrix} P^*(E_1\ E_2\ E_3) \\ P^*(E_1 E_2\ \overline{E_3}) \\ P^*(E_1 \overline{E_2}\ E_3) \\ P^*(E_1 \overline{E_2}\ \overline{E_3}) \\ P^*(\overline{E_1} E_2 \overline{E_3}) \\ P^*(\overline{E_1}\ \overline{E_2}\ E_3) \\ P^*(\overline{E_1}\ \overline{E_2}\ \overline{E_3}) \end{pmatrix} = M_3 = \begin{pmatrix} P(E_1\ E_2\ E_3) \\ P(E_1 E_2\ \overline{E_3}) \\ P(E_1 \overline{E_2}\ E_3) \\ P(E_1 \overline{E_2}\ \overline{E_3}) \\ P(\overline{E_1} E_2 \overline{E_3}) \\ P(\overline{E_1}\ \overline{E_2}\ E_3) \\ P(\overline{E_1}\ \overline{E_2}\ \overline{E_3}) \end{pmatrix}$$

$$M_3 = \begin{bmatrix} \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta(1-\theta)^2 \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 & (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta(1-\theta)^2 & (1-\theta)^3 & \theta^3 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta(1-\theta)^2 \\ \theta(1-\theta)^2 & \theta^2(1-\theta) & (1-\theta)^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^3 & \theta(1-\theta)^2 & \theta^2(1-\theta) \\ \theta(1-\theta)^2 & (1-\theta)^3 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^3 & \theta^2(1-\theta) \\ (1-\theta)^3 & \theta(1-\theta)^2 & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta(1-\theta)^2 & \theta^2(1-\theta) & \theta^2(1-\theta) & \theta^3 \end{bmatrix}$$

IX. CONCLUSION

Getting the valuable and significant information from an immense measure of data is known as data mining. It is a conspicuous field for pursuit and research of information. To improve the correspondence among clients and associations data mining is utilized.

Save the private information is essential in data mining. It is the issue on which research built up their exploration from multiple points of view. For ensuring the overview protection and to maintain a strategic distance from the inclination answer the randomized reaction strategy was created. To avert the information of the overview certain haphazardness will include with the appropriate responses. To improve the security level of the protect information this examination use multigroup strategies.

REFERENCES

1. Giudici, P, “Applied Data-Mining: Statistical Methods for Business and Industry.” John Wiley and Sons (2003) West Sussex, England.
2. Edelstein, Herb. Data Mining News "Two Crows Releases 1999 Technology Report". Volume 2, number 18. 10 May 1999.
3. “Data Mining: An Introduction”, SPSS Whitepaper. SPSS. 2000.
4. Deependra Dwivedi, “Study Analysis of data mining Algorithms: case study” Researcher. 2012;4(2):16-19] 2012, http://www.sciencepub.net.
5. Greg., "E-Voting Milestones," IEEE Security and Privacy, Gayatri Nayak, Swagatika Devi, “A Survey On Privacy Preserving Data Mining Approaches And Techniques” , IJEST, Vol. 3 No. 3 March 2011.
6. G.Rama Krishna, G.V.Ajaresh, I.Jaya Kumar Naik, Parshu Ram Dhungyel, D.Karuna Prasad “A New Approach to Maintain Privacy And Accuracy In Classification Data Mining” IJCSET Volume 2, Issue 1, January 2012 Y.
7. An Overview of Data Mining Techniques Excerpted from the book by Alex Berson, Stephen Smith, and Kurt Thearling. Page no 2.
8. Lior Rokach and OdedMaimon, “Top-Down Induction of Decision Trees Classifiers – A Survey” IEEE Transactions On Systems, Man And Cybernetics: Part C, Vol. 1, No. 11, November 2002.
9. Md. Zahidul Islam and Ljiljana Brankovicv “DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique for Preserving Privacy in Data Mining”.
10. Zhouxuan Teng, Wenliang Du,”A Hybrid Multi-Group Privacy-Preserving Approach for Building Decision Trees”.
11. Gerty J. L. M. Lensvelt-Mulders, Joop J. Hox And Peter G. M. Van Der Heijden “How To Improve The Efficiency of Randomised Response Designs “, Springer 2005.
12. D. A. Puspito Sari, I. Listiyowati, T. Nefianto, and Lasmono, “The Discrepancy between The Programs and Disaster Management Policy in Klapanunggal District, Bogor, West Java,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 135, no. 1, p. 012011, Mar. 2018.
13. D. A. P. Sari, S. Innaqa, and Safrilah, “Hazard, Vulnerability and Capacity Mapping for Landslides Risk Analysis using Geographic Information System (GIS),” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 209, no. 1, p. 012106, Jun. 2017.
14. Susilo, D., Christantyawati, N., Prasetyo, I. J., & Juraman, S. R. (2019, March). Content analysis of LINE application user: intersecting technology and social needed. In *Journal of Physics: Conference Series* (Vol. 1175, No. 1, p. 012224). IOP Publishing.
15. Ahmar, A. S., Rusli, R., & Ihsan, N. (2017). Design and Development Website of Research Institute, Case Study: Universitas Negeri Makassar. *Jurnal Studi Komunikasi*, *1*(3), 271-279.
16. Lydia, E.L., Kumar, P.K., Shankar, K., Lakshmanaprabu, S.K., Vidhyavathi, R.M. and Maseleno, A., 2018. Charismatic Document Clustering Through Novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction. *International Journal of Parallel Programming*, pp.1-19.
17. Elhoseny, M., Shankar, K., Lakshmanaprabu, S.K., Maseleno, A. and Arunkumar, N., 2018. Hybrid optimization with cryptography encryption for medical image security in Internet of Things. *Neural computing and applications*, pp.1-15.
18. Maheswari, P.U., Manickam, P., Kumar, K.S., Maseleno, A. and Shankar, K., Bat optimization algorithm with fuzzy based PIT sharing (BF-PIT) algorithm for Named Data Networking (NDN). *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp.1-8