

Achieving Privacy Preservation in Data Mining Using Hybrid Transformation and Machine Learning Technique

Pinkal Jain

Pinku029jain@gmail.com

Research Scholar

Amity University Gwalior, MP, India

Dr. Harish Kumar Shakya

hkshakya@gwa.amity.edu

Assistant Professor

Amity University Gwalior, MP, India

Article Info

Page Number: 7883 – 7889

Publication Issue:

Vol 71 No. 4 (2022)

Article History

Article Received: 25 March 2022

Revised: 30 April 2022

Accepted: 15 June 2022

Publication: 19 August 2022

Abstract

Extraction of meaningful patterns and knowledge from a vast number of datasets is known as data mining. Due to the availability of vast amounts of data and the need to turn that data into meaningful information, data mining has received a lot of attention in the IT sector in recent years. This crucial data can be applied to a number of fields, including fraud detection, market analysis, customer retention, manufacturing controls, and scientific research. We need privacy-preserving procedures when this data is moved from one location to another since different sorts of hackers or attackers may leak our private information to the public. In our study, we use a hybrid transformation strategy to achieve high degree privacy preservation. The transformation approach allows us to alter the given data objects' position, size, shape, and orientation. We employ the k means clustering technique to carry out machine learning operations. We use a dataset (the patient dataset) for experimental purposes, and WEKA is used for all operations. Comparing our work to the prior work, it provides the maximum level of anonymity.

Keywords: Data mining, Orange tool, Privacy, Clustering, K means clustering.

1. Introduction

The method by which we draw relevant patterns and knowledge from the numerous databases is known as data mining. Today's databases are very large and contain a lot of data, but we want to find the relevant information from all of those databases, or we want to find some interesting patterns. Using traditional database management systems, this is very challenging. However, by using data mining techniques, we can uncover the hidden patterns and knowledge from large database systems. In light of this, data mining can be used for knowledge discovery, pattern recognition, etc. However, several other procedures known as pretreatment of data must be used before using data mining techniques. Even if data mining is one of the steps in the process of knowledge discovery, its name nonetheless makes it more well-known. The analysis and acquisition of various relationships in the food, market, environmental, and many other conditions is done

using data mining techniques on bio-databases in order to find relationships that can reveal the origin of any disease at an extremely early stage and allow for the appropriate mitigation of risk. The difficulty is to identify the relevant information and pattern from that data so that it can be used for further research to uncover some significant results for the field of medical science and market analysis, which is one where a lot of data is acquired and collected from many sources.

2. Proposed Work

This work uses a database of patient records. If this important data has to be communicated to the administrator first, then privacy is needed because someone could change this important data during the data transmission, which could pose a lot of dangers. Modify data to protect communications from intruders. In our work, we provide two levels of security by applying transformation techniques. This technique keeps the original data intact, but before sending the valuable data to the administrator records the changes in a copy and uses this copy for communication. In this copy, an intruder would have to do a lot of work to crack this precious data, so they would perform conversion techniques, but the data could be safely communicated. Then send this valuable copy to your administrator. The administrator uses the same method that he used for the database on the client side.

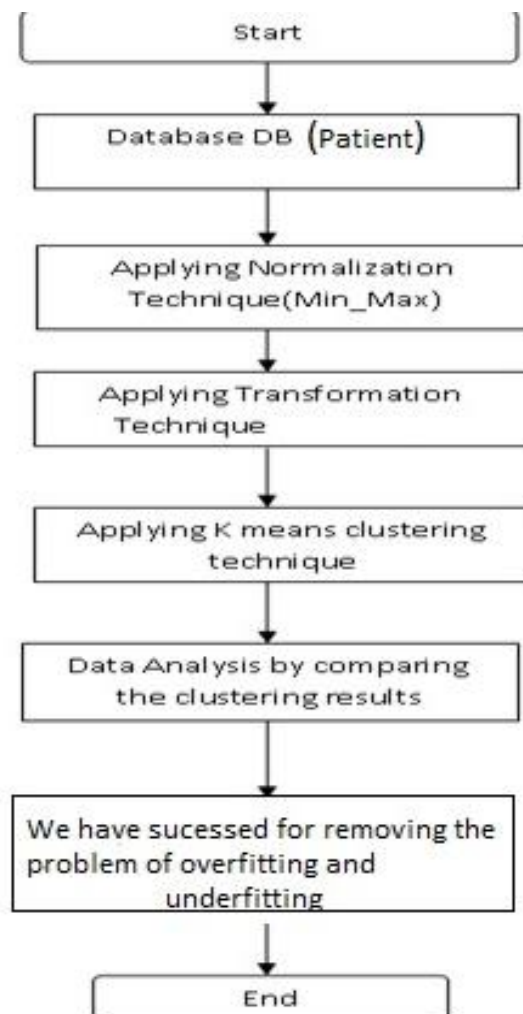


Fig. 1 Flow chart

3. IMPLEMENTATIONWORK

In implementation work, we are taking the Patient dataset that contains three attributes like age, weight and height then we apply hybrid transformation technique on our dataset for providing privacy and preserving the privacy. For analyzing the dataset we use the K means clustering technique and implemented this work with the help of weka tool. Patient dataset is shown in table 1.

Serial Number	Customer Age (in year)	Customer Weight (in kg)	Customer Height (in feet)
1	2	12	2.7
2	10	22	4.8
3	20	47	4.6
4	25	63	5.1
5	12	40	5.8
6	30	57	5.4
7	20	48	5.5
8	18	68	5.7
9	12	70	5.8
10	28	50	5.8
11	26	53	5.5
12	31	43	5.7
13	17	59	5.7
14	30	55	5.8
15	15	63	5.5
16	23	52	5.8
17	37	72	5.5
18	30	67	6
19	24	54	6.1
20	16	63	5.6
21	13	61	4.7
22	19	73	5.7
23	34	82	5.5
24	20	75	6
25	42	44	5.7

Table 1 Patient dataset

After applying the transformation technique the dataset is shown in table 2.

Serial Number	Customer Age(in year)	Customer Age(in year) After Hybrid Transformation
---------------	-----------------------	--

1	2	5.0
2	10	25.0
3	20	50.0
4	25	62.49
5	12	30.0
6	30	75.0
7	20	50.0
8	18	70.0
9	32	80.0
10	28	45
11	26	65.0
12	31	47.5
13	17	42.5
14	30	75.0
15	15	37.5
16	23	57.5
17	37	67.5
18	30	75.0
19	24	60.0
20	16	40.0
21	13	31.829
22	19	70.853
23	34	83.048
24	20	102.560
25	42	73.292

Table2 After applying the transformation technique

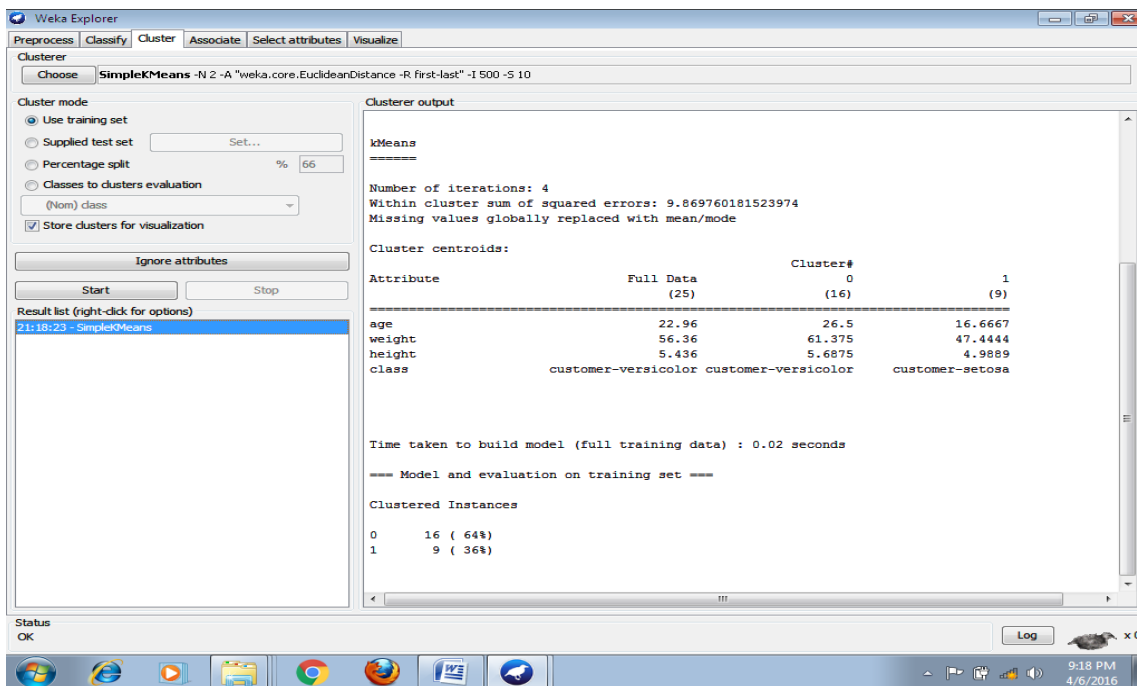


Fig. 2: Clustering on original dataset

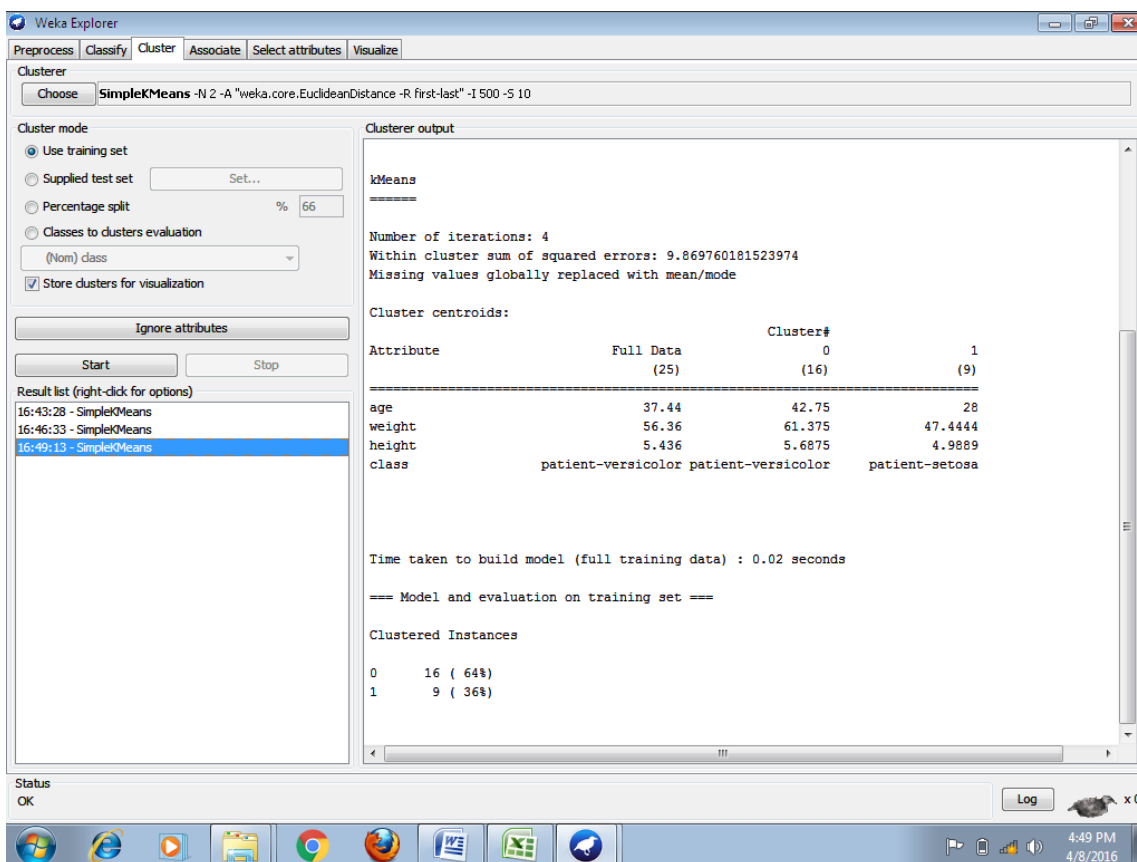


Fig. 3: Clustering on modified dataset

4. Result & Comparison

Obtained results have been compared with the base paper [1] in which author has proposed privacy preservation in data mining based on min _max normalization technique. Proposed approach reduces the problem of over fitting and under fitting. The comparison between the base paper and proposed method is shown in table 3.

S.No.	Original data values	Base paper	Proposed System
1	2	10	6
2	10	33	18
3	20	62	33
4	25	76	40.5
5	12	39	21
6	30	90	48
7	20	62	33

Table 3: Comparison between the base paper and proposed work.

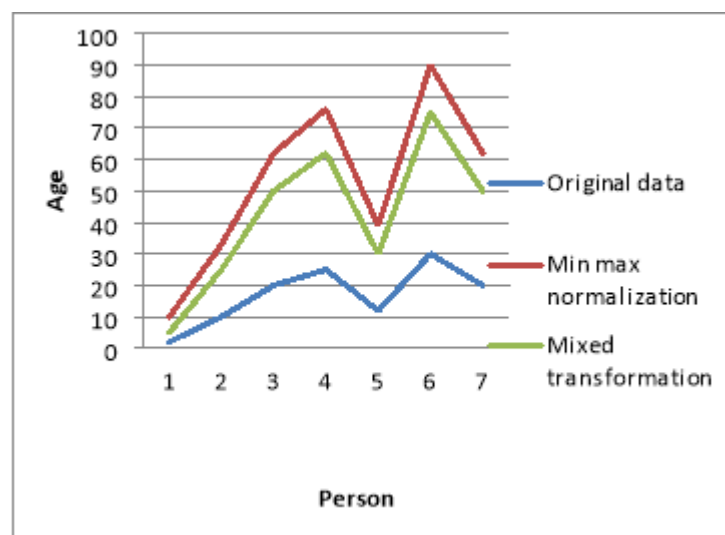


Fig.3: Comparison between the base paper and proposed work.

5. Conclusion & Future Work

In this work, we explored a hybrid conversion technique that mitigates privacy protection concerns. Our experiments proved that performing k-means clustering on skewed data yields the same clustering results as on the original data. Therefore, it can be said that the problem of over fitting and under fitting has been mitigated, and privacy has also been achieved. The proposed work can be extended by using the concept of distributed databases to protect privacy and provide default tolerance. We can also apply other advanced machine learning and privacy techniques to provide the highest security and completely eliminate over-and-under-fitting issues.

References

- [1]. **Adekitan, A.I., Abolade, J. & Shobayo, O.** Data mining approach for predicting the daily Internet data traffic of a smart university. *International Journal of Big Data* **6**, 11 (2019).
- [2]. **Syed Md. Tarique Ahmad**, et al “Privacy Preserving in Data Mining by Normalization” . IN: *Proc. Of International Journal of Computer Applications* (0975 – 8887), Volume 96– No.6, June 2017.
- [3]. **S. Vijayarani, et al** “Data Transformation Technique for Protecting Private Information in Privacy Preserving Data Mining”. In: *Proc. of Advanced Computing: An International Journal (ACIJ)* , Vol.1, No.1, November 2016.
- [4]. **Srikant, R., Agarwal, R** “Mining generalized association rules”, In: *VLDB’95*, pp.479-488, 1994.
- [5]. **Agrawal, R., Srikant, R**, “Privacy-Preserving Data Mining”, In: *proceedings of the 2000 ACM SIGMOD on management of data*, pp. 439-450, 2015.
- [6] **Lindell, Y., Pinkas, B**, “Privacy preserving data mining”, In: *Proceedings of 20th Annual International Cryptology Conference (CRYPTO)*, 2014.
- [7] **Kantarcioglu, M., Clifto, C**, “Privacy-Preserving distributed mining of association rules on horizontally partitioned data”, In *IEEE Transactions on Knowledge and Data Engineering Journal*, IEEE Press, Vol 16(9), pp.1026-1037, 2013.
- [8] **Han, J. Kamber, M**, “Data mining Concepts and Techniques”. Morgan Kaufmann, San Francisco, 2013.
- [9] **Sheikh, R., Kumar, B., Mishra, D, K**, “A Distributed k- Secure sum Protocol for Secure Multi Site Computations”. *Journal of Computing*, Vol 2, pp.239-243, 2012.
- [10] **Sheikh, R., Kumar, B., Mishra, D, K**, “A modified Ck Secure sum protocol for multi party computation”. *Journal of Computing*, Vol 2, pp.62-66, 2011.
- [11] **Jangde, P., Chandel, G, S., Mishra, D, K.,:** ‘Hybrid Technique for Secure Sum Protocol’ *World of Computer Science and Information Technology Journal (WCSIT)* ISSN: 2221-0741 vol 1, No. 5, 198-201, (2011).
- [12] **Sugumar, Jayakumar, R., Rengarajan, C (2012)** “Design a Secure Multi Site Computation System for Privacy Preserving Data Mining” .*International Journal of Computer Science and Telecommunications*, Vol 3, pp.101-105.
- [13] **N V Muthu Lakshmi, Dr. K Sandhya Rani** , “Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database”, *International Journal of Data Mining & Knowledge Management Process (IJDMP)* Vol.2, pp.17-29, 2012.
- [14] **N V Muthulakshmi, Dr. K Sandhya Rani**, “Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques”, *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 3 (1) , PP. 3176 – 3182, 2012.
- [15] **J.Vaidya**, “Privacy preserving data mining over vertically partitioned data,” Ph.D. dissertation, Purdue University, 2004.