International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Normalization and Transformation Technique Based Privacy Preservation in Data Mining

[1] Prateek Kumar Singh; [2]Naazish Rahim
[1]prateek8030@gmail.com; [2]naazish.rahim786@gmail.com

ABSTRACT

Data mining is the process in which we extract the useful patterns and knowledge from the large amount of databases. Data mining has attracted a big deal of attention in the IT industry and in society in recent years, due to the availability of large amount of data and the imminent need for converting such data into useful information and knowledge. This information and knowledge can be used for the applications like fraud detection, ranging from market analysis, customer retention to production controls and science exploration. Data mining generally viewed as the result of the natural evolution of information technology. Now a day's everyone wants to store their data or information in the online media. When this stored data is transferred from one place to another we require privacy preserving techniques because different types or hackers or attackers can disclose our private data. In our work we provide two level security by using normalization and transformation technique. With the help of normalization technique we can convert given data values into the specified range and with the help of translation transformation technique we can change the position of the given data objects. For performing the clustering operation we use k means clustering technique. Our work gives the highest privacy as compared to the previous work.

Keywords-Data Mining; Normalization and Transformation Technique; K Means Clustering Technique

1. INTRODUCTION

Data mining is the process in which we extract the useful patterns and knowledge from the large amount of databases. Now a days the databases are very large which consists of so much information but what we want to find is the relevant data from large amount of databases or want to find some interesting patterns which becomes very difficult with normal database management systems but with the use of data mining techniques we can find the hidden patterns and knowledge from large database system. So conclusion is that data mining as the knowledge mining, pattern extraction etc. But before applying the data mining techniques we need to apply some other processes which we known as preprocessing of data. Although data mining is one of the step involved in process of knowledge discovery but still it becomes more popular by name then that.

The data mining techniques are also used on Bio-Database for analyzing and acquiring the different relations in the food condition of market or environmental conditions and many other conditions to find the relations which can tell the cause of any disease at very early stage so that proper precautions can be taken. Bio-Database is the collection of information of medical science which contains the information about the patients, diseases and cause of diseases and many

more things which are related to the medical science but this database of Bio-Database contains very huge amount of data or the information which is not easy to analyze and also finding out some useful information from that is also very difficult.

Medical science and market analysis is a field where large amount of data is gathered and collected from many sources now the challenge is to find the appropriate information and pattern from that data so that it can be used for further research to find some valuable results for the patients and customers but security is the major issue we should be very careful while sending data from one place to other otherwise it may create some harmful effects. This thesis work is mainly to provide privacy to such type of data so that the information remains safe while transferring data from one place to other. In this thesis work we are going to concentrate on finding the valuable information or patterns or relations between many things from large dataset which can be of any field and then security will be our major concern while transferring data from one environment to other environment for which we will use data modification techniques which will provide security to database and ensures secure transformation of valuable data.

Data mining is the process which is used to extract useful patterns and information from large databases. In this work we are going to take a database that is patient dataset. We now discuss about security issues as while communicating the data from one place to other we need to provide security to our database. When we need to communicate this important data with the admin first we need privacy as there is possibilities that someone in between the communication of data may change this important data which will cause many hazards so in order to secure our communications from intruders, we will modify our data. In our work we provide two level security by using normalization and transformation technique.

2. PROPOSED WORK

Data mining is the process which is used to extract useful patterns and information from large databases. In this work we are going to take a database that is patient dataset. We Now discuss about security issues as while communicating the data from one place to other we need to provide security to our database. When we need to communicate this important data with the admin first we need privacy as there is a possibility that someone in between the communication of data may change this important data which will cause many hazards so in order to secure our communications from intruders, we will modify our data. In our work we provide two level security by using normalization and transformation technique. In this technique we keep the original data as it is but before sending the valuable data to admin, we put changes in one copy and use that copy for communication in this copy we perform normalization and transformation technique due to which intruder will have to work a lot in order to crack this valuable data and our data will be secure for communication.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

Flow Chart

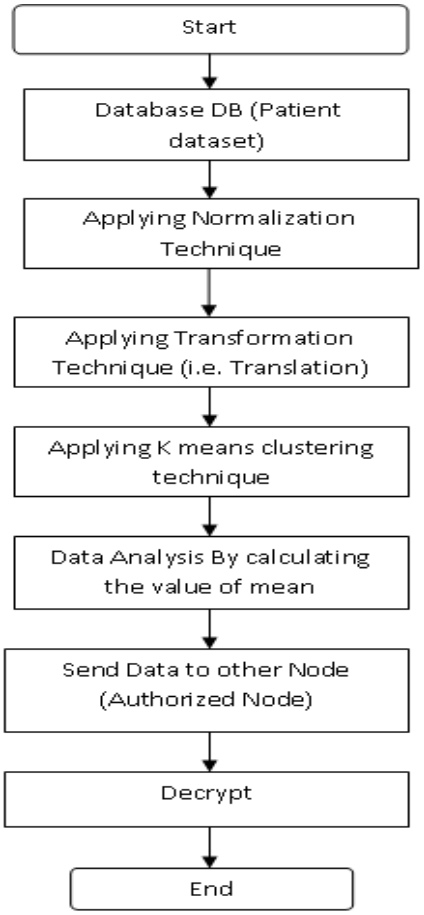

Figure 1. Flow chart of proposed work

3. Implementation & Results

In implementation work, we are taking the patient database (i.e. patient dataset) that contains four number of attributes like age, weight, height and class then we apply the min_max normalization technique and translation transformation technique on our dataset for providing the highest privacy. For analyzing the data we use the K means clustering technique and implemented this work with the help of weka data mining tool.

S. No.	Age(in year)	Weight(in kg)	height(in feet)	Class
1	2	15	2.1	1
2	10	18	4.6	1
3	20	49	4.9	1
4	25	65	6.1	1
5	12	42	4.8	1
6	30	59	5.9	1
7	20	49	5.7	1
8	18	68	6	1
9	32	71	5.9	1
10	28	52	5.3	1
11	26	51	5.7	1
12	31	48	5.0	1
13	17	53	5.5	1
14	30	57	5.9	1
15	15	64	5.1	1
16	23	55	5.6	1
17	37	70	5.9	1
18	30	61	6.1	1
19	24	54	6.0	1
20	16	67	5.6	1
21	13	62	4.8	1
22	19	73	5.7	1
23	34	82	5.8	1
24	20	77	6.1	1
25	42	47	5.8	1

Table 1: Patient dataset

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

After applying the normalization and transformation technique on age attribute the dataset is shown in table 2.

Serial Number	Age(in year)	Age(in year) After Normalization and Transformation
1	2	7
2	10	17
3	20	29.5
4	25	35.75
5	12	19.5
6	30	42
7	20	29.5
8	18	27
9	32	44.5
10	28	39.5
11	26	37
12	31	43.25
13	17	25.75
14	30	42
15	15	23.25
16	23	33.25
17	37	50.75
18	30	42
19	24	34.5
20	16	24.5
21	13	20.75
22	19	28.25
23	34	47
24	20	29.5
25	42	57

Table 2: Patient dataset after applying the normalization and transformation technique

We perform k means clustering on both original and modified data sets its results are shown in the following snapshots.

For k=2(For original dataset)

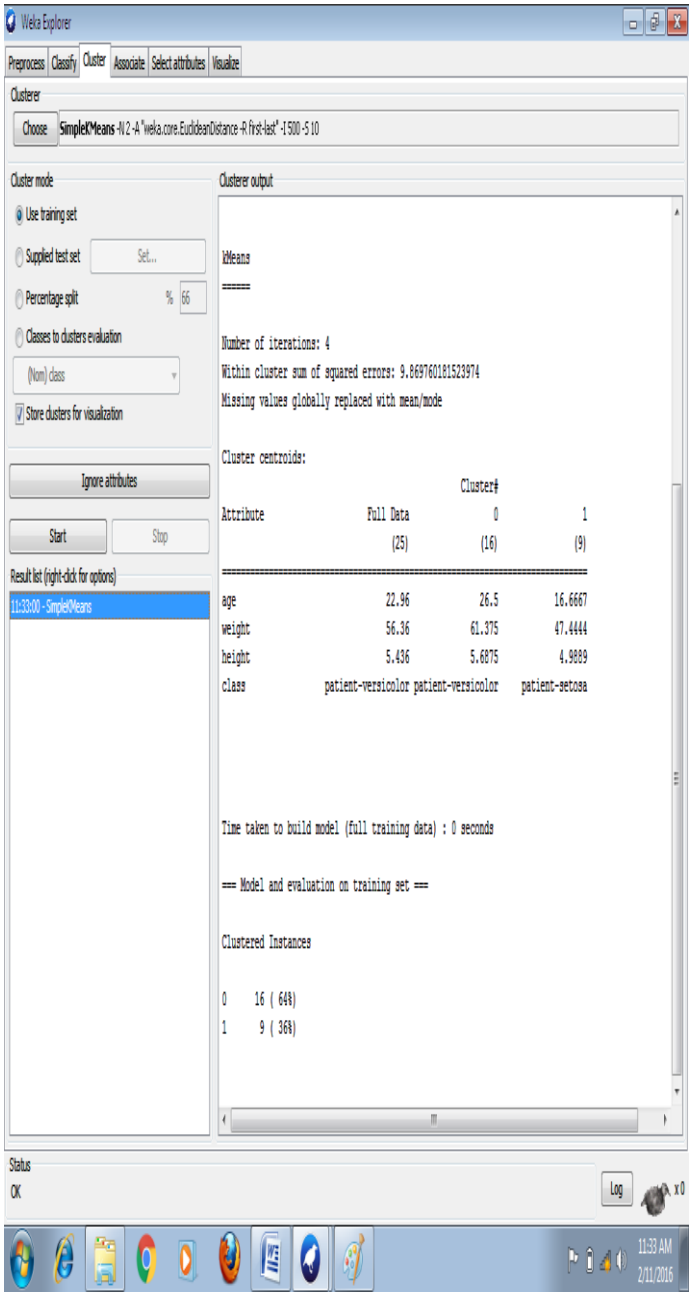

Figure 2: Clustering on original dataset (k=2)

For k=2(For modified dataset)

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

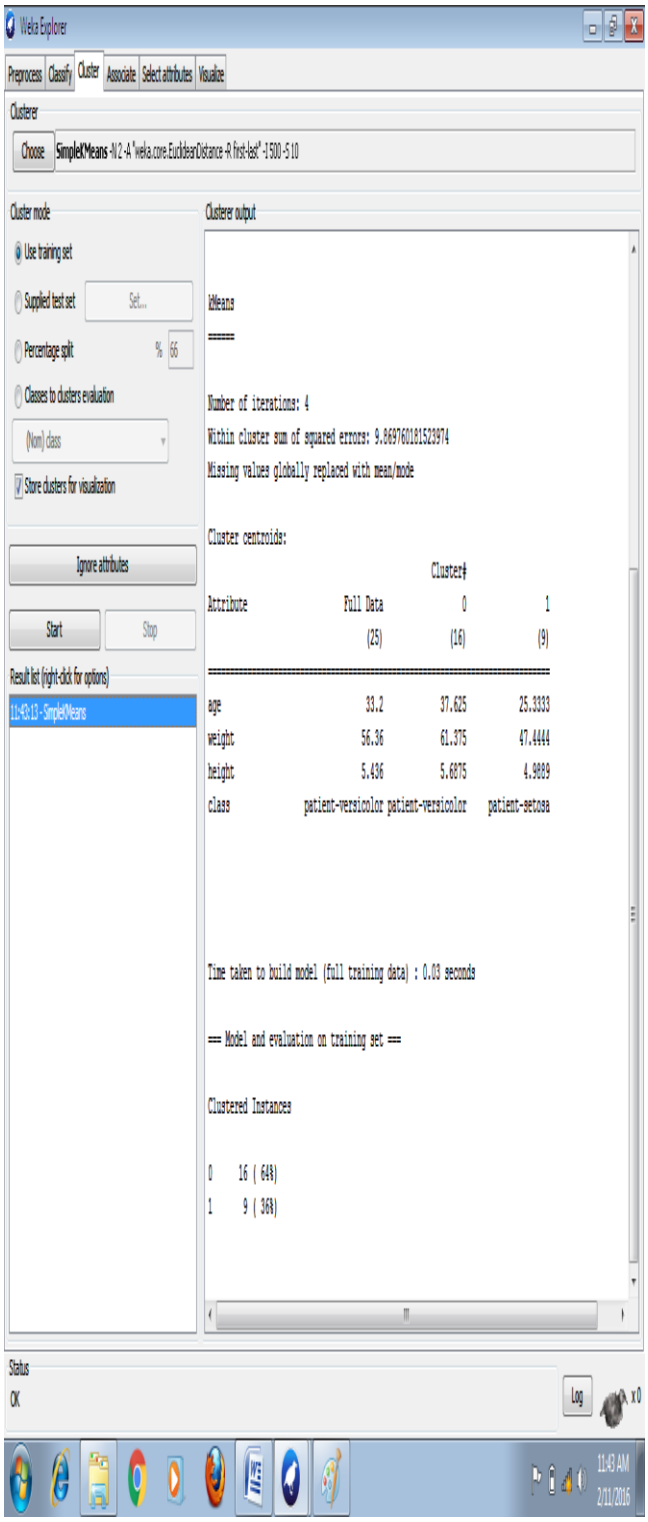

Figure 3: Clustering on modified dataset (k=2)

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 provides two level security and transforms the original data values into privacy-preserved data maintaining the inter relative distance among the data. The comparison between the base paper and proposed method (for attribute age) is shown in table 3 and its graph is shown in figure.4

S.No.	Original data values	Base paper	Proposed System
1	2	10	7
2	10	33	17
3	20	62	29.5
4	25	76	35.75
5	12	39	19.5
6	30	90	42
7	20	62	29.5

Table 3: Comparison table

4. Comparison

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 3, Issue 05, March 2016
Available at http://internationaljournalofresearch.org

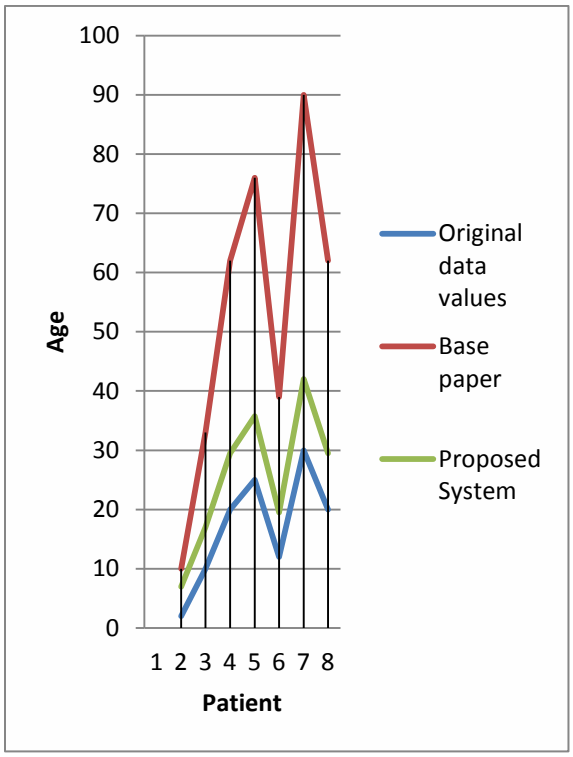

Figure 4: Comparison graph

5. Conclusion and future work

In this work we have dealt with normalization and transformation technique to preserve data privacy. Our approach convert the original data values into the privacy- preserved data maintaining the inter relative distance among the values. Our work have proven that performing the k- means clustering technique on the distorted data values produces same clustering results as original data values. So we can say we have succeeded for achieving both accuracy and highest privacy. We have tested this approach for the numerical data set.

In future work of this proposed approach is to extend the same over categorical data values and apply other techniques or approaches for preserving the privacy. We can also extend the this work by using the concept of distributed database in order to preserve privacy and for providing fault tolerance.

REFERENCES

[1] 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 2014.*

[2] 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 2010.

[3]. Agarwal, R., Imielinski, T., Swamy, A. "Survey on privacy preservation in data mining", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, 1993.

[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, 2000.

[6] Lindell, Y., Pinkas, B, "Privacy preserving data mining", In: Proceedings of 20th Annual International Cryptology Conference (CRYPTO), 2000.

[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, 2004.

[8] Han, J. Kamber, M, "Data mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.

[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, 2010.

[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, 2010.

[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 (IJDKP) 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 partitioneddata," Ph.D. dissertation, Purdue University, 2004.

[16] J. Vaidya, C. Clifton, M. Kantarcioglu, and A. S. Patterson, "Privacypreservingdecision trees over vertically partitioned data," in ACMTransactions on Knowledge Discovery from Data, vol. 2, no. 3, 2008,pp. 14–41.

[17] Y. Shen, H. Shao, and L. Yang, "Privacy preserving c4.5 algorithmover vertically distributed datasets," in International Conference onNetworks Security, Wireless Communications and Trusted Computing,vol. 2. Wuhan, Hubei: IEEE computer society, April 2009, pp. 446–448.

[18] O.Goldreich, S. Micali, and A. Wigderson, "How to play any mentalgame or a completeness theorem for protocols (extended majority36abstract)," in STOC '87 Proceedings of the nineteenth annual ACMsymposium on Theory of computing, New York, 1987, pp. 218–229.

[19] N. Adam and J. C. Wortmann. Security control methods for statistical databases: A comparative study. ACM Computing Surveys, 21 (4): 515-556, 1999.

[20] T. Dalenius and S. P. Reiss. Data Swapping: A technique for disclosure control. Journal of Statistical Planning and Inference, 6(1):73-85, 1982.

[21] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. Journal of Official Statistics, 21:309-323, 2005.

[22]K. Murlidhar and R. Sarathy.Data Shuffling –a new masking approach for numerical data. Management Science, Forthcoming, 2006.

[23]V.S.Iyenger.Transforming data to satisfy privacy constraints.InProc.Of SIGKDD'02, Edmonton, Alberta, Canada, 2002.

[24] S. Rizvi and J.R Hartisa.Maintaining data privacy in association rule mining. In Proc. of the 28th VLDB Conference, pages 682-693, Hong-Kong, China, 2002.

[25]Y. Saygin, V. S. Verykios and A. K. Elmagarmid. Privacy preserving association rule mining. In RIDE, pages 151-158, 2002.

[26]A.V.Evfimievski, R. Srikant, R. Agarwal and J.Gehrke.Privacy preserving mining of association rules.InProc. Of the Eighth ACM SIGKDD International Conference on Knowledge and Data Mining, pages 217-228, 2002.