Privacy Preservation and Data Mining

Gauri Deo[1], Dr. Prof. G. A. Kulkarni[2]

[1] PG Student, Dept. of Communication Engineering, S.S.G.B.C.O.E.T, Maharashtra, India
[2] Assistant professor, dept. of E & TC, S.S.G.B.C.O.E.T, Maharashtra, India.

ABSTRACT

The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. The author is trying to reduce the privacy risk brought by data mining operations. Privacy among the data provided is maintained. That is by preventing the any data leak, data movement or any third party access to it. Also maintain the refresh rate for third party access and performed the data mining using the k-means clustering. The collected data is showed in graph format which shows the histogram of average improvement of structure.

Keyword: - *PPDM , Data Mining, K-means Clustering, Privacy Preservation*

1. Introduction

In modern days organizations are extremely dependent on Data Mining results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. Data Mining deals with automatic extraction of previously unknown patterns from large amounts of data sets. This data includes sensitive data about Individuals or organizations. While running Data Mining algorithm against such data, the algorithm not only extracts the knowledge but it also reveals the information which is considered to be private. The real threat is that once information gets exposed to unauthorized party, it will be impractical to stop misuse. Privacy can for instance be threatened when Data Mining techniques uses the identifiers which themselves are not very sensitive, but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information. Privacy is very important for trusted collaboration and interactions. Because of these privacy and data security concerns in data mining, the data owner hesitates while sharing data for data mining activities. And this creates barrier in data mining task. Privacy preserving data mining technique gives new direction to solve this problem [1].

Solution to this problem is provided by Privacy preserving in data mining (PPDM). PPDM is a specialized set of Data Mining activities where techniques are evolved to protect privacy of the data, so that the knowledge discovery process can be carried out without barrier. The objective of PPDM is to protect sensitive information from leaking in the mining process along with accurate Data Mining results. In PPDM, new techniques are invented to provide privacy for the knowledge discovered in Data Mining. It also takes care that knowledge discovery process should not be banned because of privacy reason [2].

2. System description

The data miner applies mining algorithms to the data provided by data collector, and he wishes to extract useful information from data in a privacy-preserving manner. PPDM covers two types of protections, namely the protection of the sensitive data themselves and the protection of sensitive mining results. With the user role-based methodology proposed, it is consider that the data collector should take the major responsibility of protecting sensitive data, while data miner can focus on how to hide the sensitive mining results from untrusted parties. To perform data mining on the available information in this project we used the k-means clustering algorithm on the data available to us.

The system architecture is composed of three parts, it’s also called as Three-tier architecture. It is a well-established software application architecture that organizes applications into three logical and physical computing tiers:

I] Presentation tier (JSP)

The presentation tier is the user interface and communication layer of the application, where the end user interacts with the application. Its main purpose is to display information to and collect information from the user. This top-level tier can run on a web browser, as desktop application, or a graphical user interface (GUI), for example. Web presentation tiers are usually developed using HTML, CSS, JSP.

II] Application tier (Servlet)

The application tier, also known as the logic tier or middle tier, is the heart of the application. In this tier, information collected in the presentation tier is processed - sometimes against other information in the data tier - using business logic, a specific set of business rules. Servlets, through JDBC, can interact with database systems. Uses SQL for queries, and JDBC drivers handle the specifics of interacting with each database system.

III] Data tier (PostgreSQL)

The data tier, sometimes called database tier, data access tier or back-end, is where the information processed by the application is stored and managed.

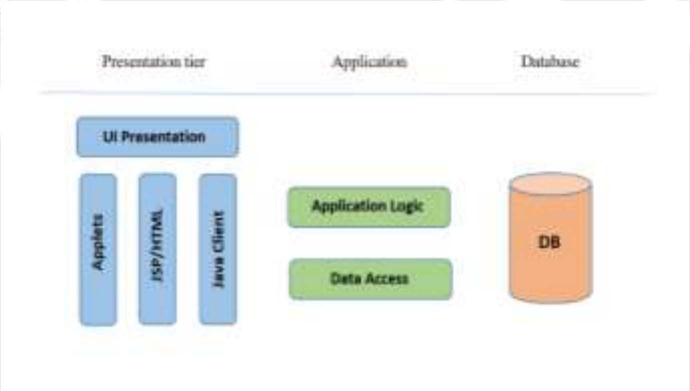


Fig -1: System Architecture

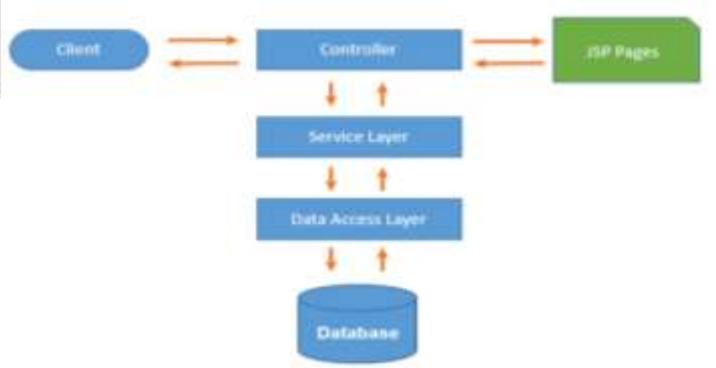


Fig -2: Workflow

2.1 k-means clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster refers to a collection of data points aggregated in together because of certain similarities. The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

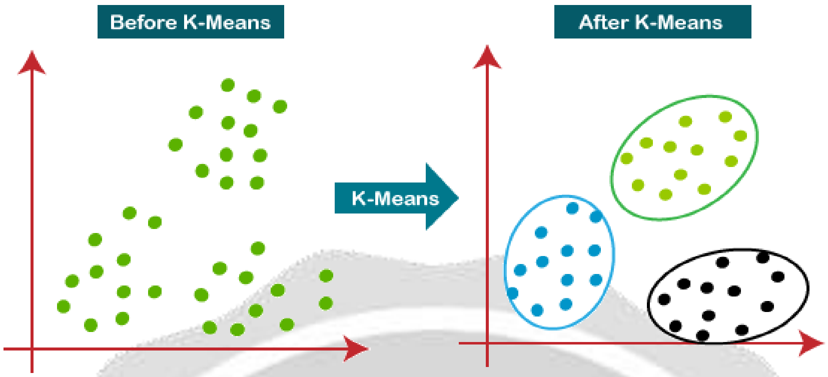
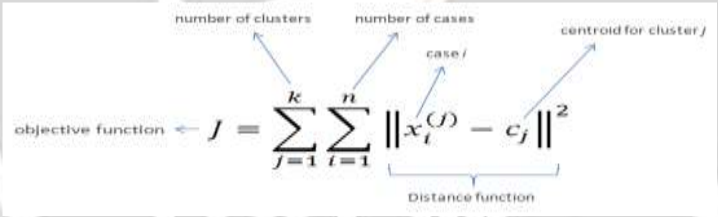


Fig -3: K-means Clustering

The working of the K-Means clustering algorithm:
Step-1: Select the number K to decide the number of clusters.
Step-2: Select random K points or centroids. (It can be other from the input dataset).
Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
Step-4: Calculate the variance and place a new centroid of each cluster.
Step-5: Repeat the third steps, which means reassign each data point to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
Step-7: The model is ready.



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

3. Result

Admin and user both are allowed to create the website by which means they are providing the data, making them the data provider. Privacy among the data provided by them is maintained. That is by preventing the any data leak, data movement or any third party access to it. The data available is of smaller amount and the distributed database improves the efficiency of the algorithm and reduces the time complexity of algorithm.

By using java library of the k-means clustering algorithm, the time spend by user on the particular site is collected from database and showed in mini session which is access by admin. The collected data is showed in graph format which shows the histogram of average improvement of website structure.

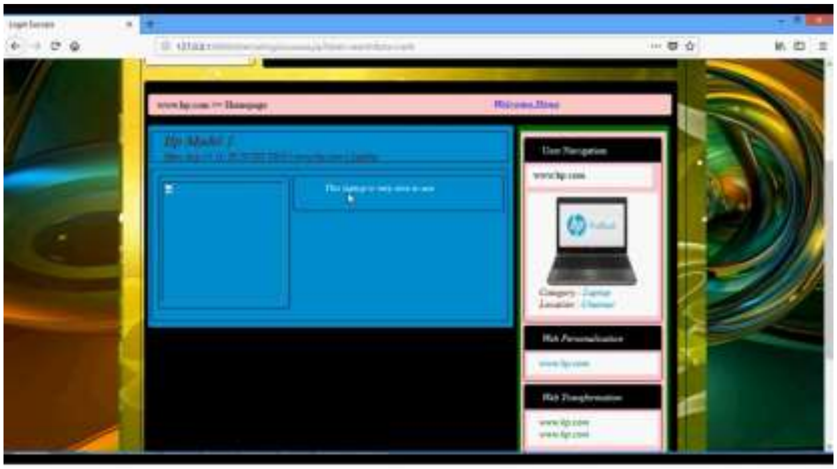
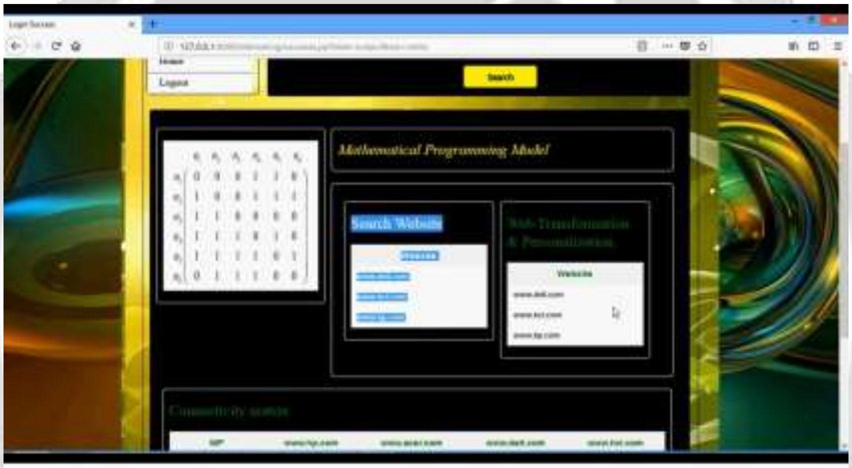
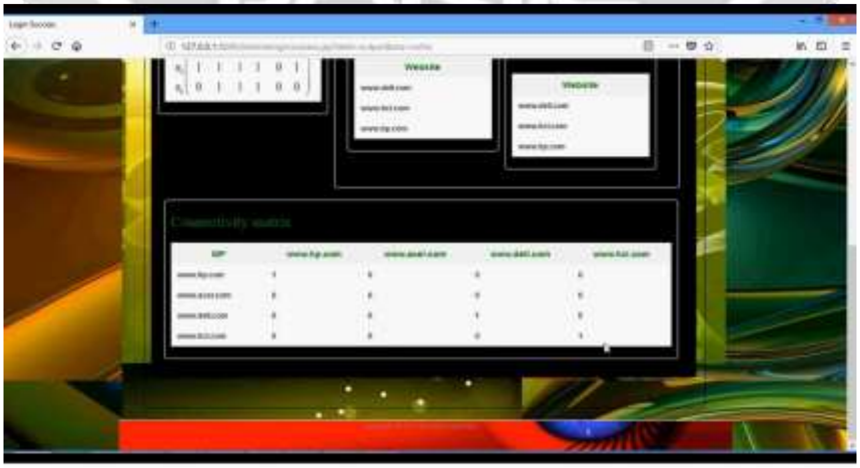


Fig -4: User searched page.



$$\begin{array}{c|cccccc} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \hline e_1 & 0 & 0 & 0 & 1 & 1 & 0 \\ e_2 & 1 & 0 & 0 & 1 & 1 & 1 \\ e_3 & 1 & 1 & 0 & 0 & 0 & 0 \\ e_4 & 1 & 1 & 1 & 0 & 1 & 0 \\ e_5 & 1 & 1 & 1 & 1 & 0 & 1 \\ e_6 & 0 & 1 & 1 & 1 & 0 & 0 \end{array}$$

Fig -4: Mathematical Programming model



SIP	www.hp.com	www.asus.com	www.dell.com	www.hcl.com
www.hp.com	1	0	0	0
www.asus.com	0	0	0	0
www.dell.com	0	0	1	0
www.hcl.com	0	0	0	1

Fig -6: Mathematical Programming Model

Fig -7: Mini Session

Fig -8: Mini Session

4. CONCLUSIONS

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. We differentiate four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. Each user role has its own privacy concerns, hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others: For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal, he can utilize security tools to limit other's access to his data, sell his data at auction to get enough compensations for privacy loss, or falsify his data to hide his true identity.

Also there is no guarantee that every participator will follow the protocol or truthfully share his data. Interactions among different participators need to be further investigated. Considering the nature of the data miner, game theory may be a proper tool for such problems. Some game theoretical approaches have been proposed for distributed data mining.

6. REFERENCES

[1]. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS100473 www.ijert.org (This work is licensed under a Creative Commons Attribution 4.0 International License.) Vol. 4 Issue 10, October-2015.

[2]. Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego: Academic Press,2001.

[3]. Neelamadhab Padhy, Dr.Pragnyaban Mishra and RasmitaPanigrahi, "The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology(IJCSEIT)", vol.2, no.3, June.

[4]. Wanliang Du, Mikhail J. Atallah "Secure Multi-Party Computation Problems and Their Applications; A Review and Open Problems". Proeedings of new security paradigms workshop, September 2001.

[5]. Ioannis Ioannidis, Ananth Grama "An Efficient Protocol for Yao's Millionaires' Problem". Proceedings of the 36th Hawaii International Conference on System Sciences IEEE 2002.

[6] Yehuda Lindell and Benny Pinkas, "Secure Multipart Computation for privacy-preserving data mining". The Journal of Privacy and Confidentiality, Vol. 1, 2009, pp. 59-98.

[7]. Durgesh Kumar Mishra, Rashid Sheikh, Beerendra Kumar, "Privacy-Preserving k-Secure Sum Protocol". (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, 2009.

[8]. Durgesh Kumar Mishra, Rashid Sheikh, Beerendra Kumar, "A Distributed k-Secure Sum Protocol for Secure Multi-Party Computations". Journal of Computing, Vol 2, Issue 3, March 2010.

[9]. Durgesh Kumar Mishra, Rashid Sheikh, Beerendra Kumar, "Changing Neighbors k-Secure Sum Protocol for Secure Multi- Party Computation". (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[10]. Gayatri Nayak and Swagatika Devi, "A servey on privacy preserving data mining: approaches and techniques". International Journal of Engineering Science and Tchnology (IJEST), Vol. 3, march 2011.

[11]. Mehmet Ercan Nergiz, Abdullah Ercument Cicek, Thomas B. Pedersen, and Yucel Saygin, "A Look-Ahead Approach to Secure Multiparty Protocols". IEEE Transactions on Knowledge and Data Engineering, Vol. 24, July 2012.

[12]. Alexandre Evfimievski, Tyrone Grandison, "Privacy Preserving Data Mining". IBM Almaden Research Center.

[13]. Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, Michael Y. Zhu "Tools for Privacy Preserving Distributed Data MIning". Vol. 4.

[14]. [IEEE 2013 Nirma University International Conference on Engineering (NUiCONE) - Ahmedabad, India (2013.11.28-2013.11.30)] 2013 Nirma University International Conference on Engineering (NUiCONE) - An efficient method for privacy preserving data mining in secure multiparty computation

[15] [IEEE 2019 IEEE Congress on Evolutionary Computation (CEC) - Wellington, New Zealand (2019.6.10-2019.6.13)] 2019 IEEE Congress on Evolutionary Computation (CEC) - A Swarm-based Data Sanitization Algorithm in Privacy-Preserving Data Mining

[16] [IEEE 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY 2011) - Subotica, Serbia (2011.09.8-2011.09.10)] 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics - Privacy preserving in data mining - Experimental research on SMEs data.

[17] [IEEE 2010 International Conference on Artificial Intelligence and Education (ICAIE) - Hangzhou, China (2010.10.29-2010.10.30)] 2010 International Conference on Artificial Intelligence and Education (ICAIE) - Application of Data Mining for emotional intelligence based on cluster analysis

[18] [IEEE 2014 Recent Advances in Engineering and Computational Sciences (RAECS) - Chandigarh, India (2014.03.6-2014.03.8)] 2014 Recent Advances in Engineering and Computational Sciences (RAECS) - Association rule sharing model for privacy preservation and collaborative data mining efficiency

[19] M. B. Malik, M. A. Ghazi, and R. Ali, ''Privacy preserving data mining techniques: Current scenario and future prospects,'' in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26–32.

[20] [IEEE 2012 4th International Conference on Computational Intelligence and Communication Networks (CICN) - Mathura, Uttar Pradesh, India (2012.11.3-2012.11.5)] 2012 Fourth International Conference on Computational Intelligence and Communication Networks - Privacy Preserving in Data Mining Using Hybrid Approach

[21] [IEEE 2016 7th International Conference on Computer Science and Information Technology (CSIT) - Amman, Jordan (2016.7.13-2016.7.14)] 2016 7th International Conference on Computer Science and Information Technology (CSIT) - Privacy preserving data mining on published data in healthcare: A survey

[22] M. A. Sheela and K. Vijayalakshmi, ''A novel privacy preserving decision tree induction,'' in Proc. IEEE Conf. Inf. Commun. Technol. (ICT), Apr. 2013, pp. 1075–1079.

[23] Malina L, Hajny J (2013) Efficient security solution for privacy-preserving cloud services. In: 36th international conference on telecommunications and signal processing (TSP), pp 23–27. http://doi.org/10.1109/TSP.2013.6613884

[24] Kamakshi P (2012) Automatic detection of sensitive attribute in PPDM. In: IEEE international conference on computational intelligence & computing research (ICCIC)

[25] Mukkamala R, Ashok VG (2011) Fuzzy-based methods for privacy-preserving data mining. In: IEEE eighth international conference on information technology: new generations (ITNG)

[26] Patil BB, Patankar AJ (2013) Multidimensional k-anonymity for protecting privacy using nearest neighborhood strategy. In: IEEE international conference on computational intelligence and computing research (ICCIC)

[27] Wang H, Hu C, Liu J (2010) Distributed mining of association rules based on privacy-preserved method. In: Internationalsymposium on information science and engineering (ISISE), pp 494–497. http://doi.org/10.1109/ISISE.2010.125

[28] Tai C-H, Huang J-W, Chung M-H (2013) Privacy preserving frequent pattern mining on multi-cloud environment. In: 2013 international symposium on biometrics and security technologies (ISBAST)

[29] W. Gan, J. C. W. Lin, H. C. Chao, S. L. Wang, P. S. Yu, "Privacy preserving utility mining: a survey," IEEE International Conference on Big Data, pp. 2617–2626, 2018.

[30] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, ''Information security in big data: Privacy and data mining,'' IEEE Access, vol. 2, pp. 1149–1176, 2014

[31]Youguo Li, Haiyan Wu," A Clustering Method Based on K-Means Algorithm", 1875-3892 © 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee doi: 10.1016/j.phpro.2012.03.206

[32]Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C," Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010

[33]Usama M.Fayyad cory A.Reina Paul S.Bradley,Initialization of Iterative Refinement clustering algorithms[C].Proc.4th International Conf.On Knowledge Discovery & Data Mining,1998.

[34]DUDA R O, HART P E.Pattern classification and scene analysis[M].New York:John Wiley & Sons,1973.

[35] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining," in Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 89_99.

[36] R. Agrawal and R. Srikant, ``Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439_450, 2000.

[37] Y. Lindell and B. Pinkas, ``Privacy preserving data mining," in Advances in Cryptology. Berlin, Germany: Springer-Verlag, 2000, pp. 36_54.

[38] C. C. Aggarwal and S. Y. Philip, A General Survey of Privacy- Preserving Data Mining Models and Algorithms. New York, NY, USA: Springer-Verlag, 2008.

[39] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects," in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26_32.