IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Study of Incentive Compatible Privacy Preserving Data Analysis

Yuvraj Singh[1], Pankaj Pratap Singh[2] , Anirudh Kumar Tripathi [3], Amit kishor[4]

[1,2,3,4]Department of CSE, Swami Vivekanad Subharti University, Meerut, India

Corresponding Author: yuvrajsinghsolank03@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.737741 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

***Abstract*—** In corporate and government department's increasingly keeping large size electronic databases, which are accessed using internet or intranet. Important information implement from the data using Privacy data mining methods. While performing data mining steps, there is an inherent danger to the privacy of the data. The valuable data stored in the database should not be accessible to users. Most of the privacy preserving methods are based on reduction in the granularity of the implementing of the data. This ends to loss of information but it improves privacy. Therefore, in PPDM there is a conflict between loss of information and the privacy. Effective Methods are required which do not compromise the security mechanisms. Some of the methods proposed for privacy preserving data mining include randomization method, k-anonymity model, l-diversity and distributed privacy preservation. The k-anonymity model is based on a quasi-identifier, which is a collection of attributes in a database that is the identifier for the entire data. All the data in the database is assumed to be in a set of tables, and each tuple is information of an individual customer. K-anonymity Methods are based on the reduction of granularity in representation of data using pseudo identifiers. Major Methods used for granularity reduction are generalization and suppression. In generalization, the attribute values are converted into a range that reduces the granularity and reduces the risk of identifying individual values. In suppression, value of the attribute is removed completely. These methods introduce loss of detail which may affect the accuracy. This induces the search for anonymization algorithms that achieve the required level of anonymization while incurring a minimization of loss of information. Finding optimal anonymous datasets using generalization or suppression has been proved to be a NP hard problem. Therefore, some standard heuristic search Methods such as Genetic Algorithms (GAs), Particle Swam Optimization (PSO) and Ant Colony Optimization (ACO) can be used to find optimal datasets.

Keywords*—Data mining, Secure Multiparty Computation, Genetic Algorithm, Particle Swam Optimization, Ant Colony Optimization*

I. INTRODUCTION

With the advancement in hardware and software Technology, high data storage comes in existence and this leads to a rapid use of internet and so the data mining started rapidly from the users of internet and database which increase the transition of personal information related to users or companies hence the misuse of the data also increases rapidly to prevent this threat privacy of data become big concern for computer science fraternity. To preserve the privacy of person or a group requirement of new technology comes in existence called privacy preserving data mining this is being extensively used to maintain the privacy of data. Privacy preserving data mining (PPDM) algorithms developed so that confidential data which is mined remain protected from users. The main concerns of HPPDM is that sensitive raw data like names, ad dresses are modified from the original database, so that the users of the data will not be able to compromise another person's privacy. And also, sensitive knowledge obtained from mining which can compromise data privacy must be excluded. Privacy preservation is to be integrated at two levels, users' personal information and their collective activity. The former is known as individual privacy preservation and the latter as collective privacy preservation data mining [1]. Privacy preserving of data must safeguard from divulging sensitive data during publication of individual data. To maintain privacy, a number of techniques have been proposed for modifying or transforming the data. To avoid data misuse, the data is anonymized. Many data mining techniques are modified to ensure privacy. The techniques for PPDM are based on cryptography, data mining and information hiding. In general, statistics- based and the crypto-based approaches are used to tackling PPDM. In the statistics-based approach, the data owner's sanitize the data through perturbation or generalization before publishing. Knowledge models such as decision trees are used on the sanitized data. The advantage of statistics-based approach is that it efficiently handles large volume of datasets. In the crypto-based PPDM approach, data owners have to cooperatively implement specially designed data mining algorithms. Though these algorithms achieve verifiable privacy protection and better data mining performance, it suffers from performance and scalability issues . In recent years, privacy preserving data for a single database has been extensively studied. Data anonymization transforms a Data set to uphold privacy using methods such as k-anonymity using generalization or suppression techniques, so that individually identifiable information is masked [2].

II. RELATED WORK

METHODOLOGY: UCI Machine Learning Repository provides the 'Adult' dataset used for evaluation. It contains 48,842 instances, including categorical and integer attributes from 1994 Census information. It has about 32,000 rows with 4 numerical columns, the column which includes age {17 – 90}, fnlwgt {10000–1500000}, hrsweek {1 – 100} and edunum{1–16}. The age column and native country are anonymize using k-anonymization. Table 1 shows the original attributes of the Adult dataset.

Table 1: Attributes of the adult Dataset

Age	native-country	Class
39	India	<=50K
50	India	<=50K
38	India	<=50K
53	India	<=50K
28	Sri Lanka	<=50K
37	India	<=50K
49	Bangladesh	<=50K
52	India	>50K
31	India	>50K

In k-anonymity, the data is transformed to equivalence classes where each class has a set of k records that differs from others [22]. Generalization & suppression are used to reduce the granularity representation of the pseudo-identifiers techniques. The attributed values are generalized to a range so as to reduce the granularity (for example, date of birth generalized as year of birth) and it also reduces identification risk. The value of the attribute is removed completely to reduce the identification risk with public records (suppression). The k-anonymity is a good technique because of its simplicity in definition and also many algorithms are available to process the anonymization [23, 24].

Genetic Algorithm (GA)

In Genetic Algorithm (GA), a group of individuals called chromosomes forms the population that represents a complete solution to a defined problem [25, 2 6]. Each chromosome is encoded using a sequence ofH0s or 1s. The GA begins using a randomly generated set of individuals as population. In each iteration, a new population is generated which replaces all of members of the population. Though, certain number of the best individuals is kept from each generation and is copied with the new generation (this approach known as elitism). The best chromosome in the population is used to generate the next population. Based on the fitness functions, the population will transform into the future generation. On evaluation of population's fitness, fit test chromosomes are selected for reproduction. Lower fitness chromosomes or poor chromosomes might be selected in very less numbers or not at all. There are popular selection methods such as "Roulette-Wheel" selection, "Rank" selection and "Tournament" selection. In this study, tournament selection is used wherein two chromosomes are chosen randomly from the population. First, for a predefined probability p, the more fit of these two is selected and with the probability (1-p) the other chromosome with less fitness is selected [26].

The crossover operation in GA combines two chromosomes together to produce new offspring (child). Crossover occurs only with crossover probability.

Chromosomes remain the same when not subjected to crossover. The idea behind crossover is considering new solutions and exploiting of the old solutions.

As fittest chromosomes are selected more, good solutions are carried to the next generation. In this study, single-point crossover has been applied to produce new offspring for that a high value of crossover probability is used (between 0.80 and 0.90).Due to crossover operation, the new generation will contain only the character of the parents. This can lead to a problem saturation of finding a better population as no new genetic material is introduced in the offspring. Mutation operator introduces new genetic patterns into the new chromosomes. The new sequence of genes due to mutation may or may not produce desirable features in the new chromosome. The new mutated chromosome is kept if the fitness is better than the general population.

The Particle Swarm Optimization (PSO)

The Particle Swarm Optimization (PSO) algorithm is an adaptive algorithm made of population of individuals (commonly referred to as particles), adapting through returning stochastically toward previous successful regions [27, 28]. The two primary operators in PSO are Velocity update and Position update. During iteration, particle is accelerated toward the particles in the previous best position and the global best position. A new velocity value is updated for each particle at PSO begins with a group of random particles or solutions and searches for optima through updating of generations. The two "best" values, pbest and gbest, of the particle are updated in each iteration. 'pbest' is the best solution (fitness) achieved till the n and 'gbest' value is the best value obtained till then by any particle in the population. PSO is computationally simple as it requires only primitive mathematical operators. Particle positions and velocities are assigned randomly in the beginning of the algorithm. PSO updates all velocities and positions of all the particles iteratively as follows:

Hybrid GA-PSO

Cooperative search is a type of parallel algorithms, where several search algorithms are run in parallel to solve the optimization problem. As the search algorithms may be different, cooperative search technique is viewed as a hybrid algorithm [31]. In this work, it is proposed to implement a Hybrid Evolutionary Algorithm using Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Both GA and PSO in the proposed system work with the same population. Initially, Ps individuals which form the population are generated randomly. They can be considered chromosomes in GA, or as particles in PSO. After initialization, new next generation individuals are created by enhancement, crossover, and mutation operations.

III. RESULTS AND DISCUSSION

The generalization depends on the type of data; it can either be categorical or numeric. The generalization of the categorical data (gender, work, zip code) is described by a taxonomy tree as seen in Figure 1. The Figure shows an example for generalization of continuous data used in this work.

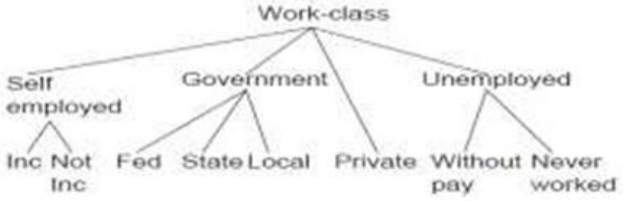

Fig. 1: Example for Generalization of continuous data as a taxonomy tree

For generalization of numeric data (age, income) is obtained by discretization of its values into a set of disjoint intervals. Various levels of discretization defined, for numeric data of age, the set of intervals:

{(0,10),(10,20),(20,30),..};

{(0,20),(20,40),(40,60),..};

{(0,30),(30,60),(60,90),..} are valid.

Experiments are conducted for different levels of k-anonymity (5, 10, …, 45, 50). Hybrid algorithm is used to find the optimal generalization feature set. Table 2 shows the parameter used for GA in this study. Following Figures and Tables give the results of classification, precision and recall for class label income. The precision and recall is shown for value greater than 50K and less than or equal to 50K

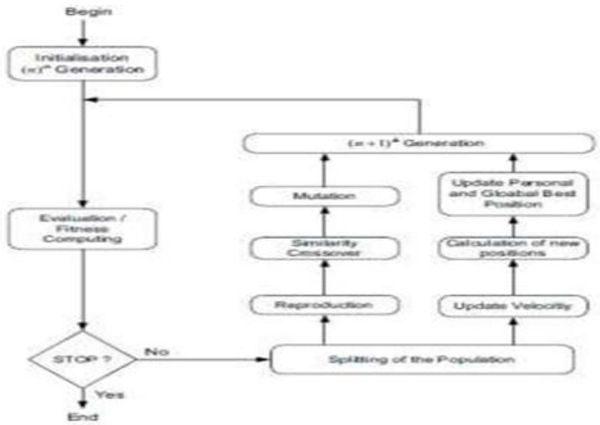
Begin
Initialisation (n)th Generation
(n + 1)th Generation
Mutation
Update Personal and Gloabal Best Position
Evaluation / Fitness Computing
Similarity Crossover
Calculation of new positions
Reproduction
Update Velocity
STOP ?
No
Splitting of the Population
Yes
End

Table 2: The Proposed Hybrid Algorithm Parameters

Initial population size	25
Maximum generations	20
Number of epochs	500
Momentum	Lower bound 0.5
Optimization	Upper bound 1.0
Step size	Lower bound 0.1
Optimization	Upper bound 0.5
Encoder mechanism	Roulette
Cross Over	Single point
Cross over probability	0.9

Mutation	0.1
Mutation probability	0.01

Table 3: Classification Accuracy for Different levels of k-anonymity

k-anonymity level	Classification Accuracy
K=50	0.832500717
K=45	0.833135416
K=40	0.836083698
K=35	0.840362803
K=30	0.847242128
K=25	0.855759387
K=20	0.862454445
K=15	0.870582695
K=10	0.875619344
K=5	0.880389828

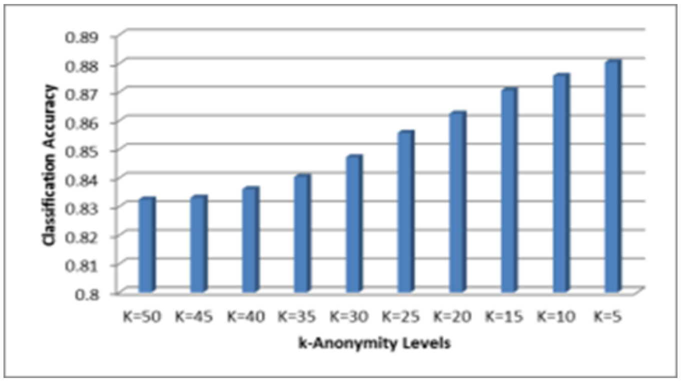

Fig. 2: Classification Accuracy for different levels of k-anonymity

It is observed from Figure 2, that the classification accuracy decreases with the increase in k-anonymity level. Figure 3 and 4 show the precision and recall for class label income greater than 50k and less than or equal to 50k respectively.

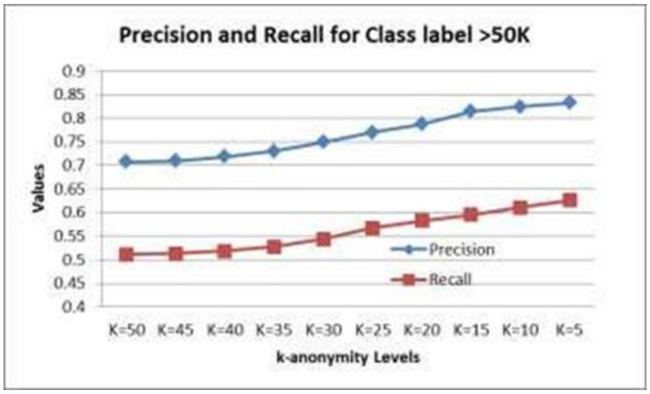

Fig. 3: Precision and Recall for different levels of k-anonymity for class label >50K

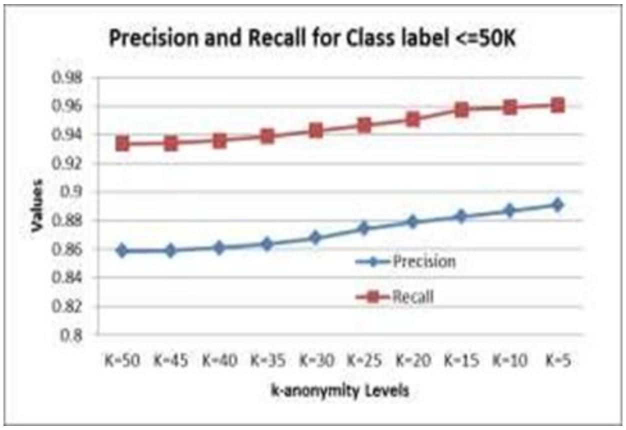

Fig 4: Precision and Recall for different levels of k-anonymity for class label<=50K

IV. CONCLUSION AND FUTURE SCOPE

Existing Evolutionary Algorithm (EA) solutions in privacy-preserving domain mainly deals with specific problems such as cost function evaluation. In this work, it is proposed to implement a Hybrid EA using Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Both GA and PSO complement each other to provide global optimization. In the proposed framework, k-anonymity is accomplished by generalization of the original dataset. The hybrid optimization is used to search for optimal generalized feature set. Experiments were conducted for different levels of k-anonymity and the results obtained are satisfactory.

V. REFERENCES

[1] Xinjing Ge and Jianming Zhu, (2011), Privacy Preserving Data Mining, New Fundamental Technologies in Data Mining.

[2] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

[3] Malin, B., Benitez, K., & Masys, D. (2011). Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association, 18(1), 3-10.

[4] Singh, M. D., Krishna, P. R., & Saxena, A. (2010, January). A cryptography based privacy preserving solution to mine cloud data. In Proceedings of the Third Annual ACM Bangalore Conference (p. 14). ACM.

[5] Patrick Sharkey, Hongwei Tian, Weining Zhang, and Shouhuai Xu, 2008, Privacy-Preserving Data Mining through Knowledge Model Sharing, Springer-Verlag Berlin Heidelberg, pp. 97– 115, 2008

[6] Pawel Jurczyk, Li Xiong, 2008, Privacy-Preserving Data Publishing for Horizontally Partitioned Databases, CIKM'08, October 26–30USA., ACM 978-1-59593-991- 3/08/10.

[7] Campan, A., & Truta, T. (2009). Data and structural k-anonymity in social networks. Privacy, Security, and Trust in KDD, 33-54.

[8] Nergiz, M. E., Clifton, C., & Nerg iz, A. E. (2009). Multirelational k- anonymity. Knowledge and Data Engineering, IEEE Transactions on, 21(8), 110 4-1117.

[9] Stokes, K., & Torra, V. (2012, March). N-Confusion: a generalization of k-anonymity. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 211-215). ACM.

[10] Cao, J., Karras, P., Kalnis, P., & Tan, K. L. (2011). SABRE: a Sensitive Attribute Bucketization and RE distribution framework for t-closeness. The VLDB Journal, 20(1), 59-81.

[11] Shi, P., Xiong, L., & Fung, B. (2010, October). Anonymizing data with quasi-sensitive attribute values. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1389-1392). ACM.

[12] A. Meyerson, R. Williams, On the complexity of optimal k-anonymity, in: Proc. Of the 23rd ACM SIGMOD-SIGCAT-SIGART Symposium, ACM, New York,NY, 2004, pp. 223–228.

[13] P.Samarati, Protecting respondents' identities in micro data release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.

[14] Van der Merwe, D., & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In IEEE congress on evolutionary computation (1) (pp. 215–220). New York: IEEE.

[15] Holden, N., & Freitas, A. (2008).A hybrid PSO/ACO algorithm for discovering classification rules in data mining. Journal of Artificial Evolution and Applications, 2008, 11 pages.

[16] Van den Bergh F. and Engelbrecht A.P., 'A Cooperative Approach to Particle Swarm Optimization', IEEE Transactions on Evolutionary Computation, 2004, pp. 225-239.

[17] Premalatha, K., & Natarajan, A.M. (2009). Hybrid PSO and GA for g lobal maximization. Int. J. Open Problems Compt. Math, 2(4), 597-608.

[18] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k-Anonymization. Proceedings of the ICDE Conference, pp. 217–228, 2005.

[19] Sakuma, J., & Kobayashi, S. (2007, July). A genetic algorithm for privacy preserving combinatorial optimization. In Proceedings of the 9th annual conference on Genetic and evolutionary computation (pp. 1372-1379). ACM.

[20] Dehkordi, M. N., Badie, K., & Z adeh, A. K. (2009). A novel method for privacy preserving in association rule mining based on genetic algorithms. Journal of software, 4(6), 555-562.

[21] Matatov, N., Rokach, L., & Maim on, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. Information Sciences, 180(14), 2696-2720.

[22] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, CMU, SRI, 1998.

[23] Lefevre, K., Dewitt, D., And Ramakrishnan, R. 2005. Incognito: Efficient full domain k-anonymity. In SIGMOD.

[24] Zhong, S., Yang, Z., And Wright,R. N. 2005. Privacy-enhancing k-anonymization of customer data. In Proceedings of the International Conference on Principles ofHData Systems (PODS).

[25] L. David, Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold. 1991.

[26] D.E. Goldberg, Genetic Algorithms: in Search, Optimization, and Machine Learning. New York: Addison-Wesley Publishing Co. Inc. 1989.

[27] Qing Cao, Tian He, and Tarek Abdelzaher, uCast: Unified Connectionless Multicast for Energy Efficient Content Distribution in Sensor Networks, IEEE Transactions On Parallel And Distributed Systems, Vol. 18, No. 2, February 2007.

[28] Latiff, N.M.A.; Tsimenidis, C.C.; Sharif, B.S., "Performance Comparison of Optimization Algorithms for Clustering in Wireless Sensor Networks," Mobile Adhoc and Sensor Systems, 2007. MASS 2007. IEEE International Conference on , vol., no., pp.1-4, 8-11 Oct. 2007.

[29] Matthew Settles," An Introduction to Particle Swarm Optimization", 2005.

[30] Eberhart, R. C., Shi, Y.: Particle swarm optimization: Developments, applications and resources, In Proceedings of IEEE International Conference on Evolutionary Computation, vol. 1 (2001), 81-86.

[31] El-Abd, M., & Kamel, M. (2005). A taxonomy of cooperative search algorithms.Hybrid Metaheuristics, 902-902.