Swarm Optimization Algorithm for Privacy Preserving in Data Mining

Sridhar Mandapati[1], Dr Raveendra Babu Bhogapathi[2] and Dr M.V.P.Chandra Sekhara Rao[3]

**[1]Department of Computer Applications,
R.V.R & J.C College of Engineering, Guntur, India.**

**[2]Department of Computer Science and Engineering,
VNR VJIET, Hyderabad, India.**

**[3]Department of Computer Science and Engineering,
R.V.R & J.C College of Engineering, Guntur, India.**

Abstract

Free competition in business can be compromised as data mining techniques reveal critical information about business transactions. Hence, there is a need to ensure prevention of disclosures both of confidential personal information which is contextually sensitive. Literature is abounding with state-of-the-art methods for privacy-preserving evolutionary algorithms (EAs) that give solutions to real-world optimization problems. Existing EA solutions are specific to cost function evaluation in privacy-preserving domains. This work proposes implementation of Particle Swarm Optimization (PSO) to locate an optimal generalized feature set. The proposed framework accomplishes k-anonymity by generalization of original dataset.

Keywords: *Privacy-Preserving Data Mining (PPDM), Swarm Intelligence, Particle Swarm Optimization (PSO), K-anonymity.*

1. Introduction

Data mining technology aimed to provide tools to automatically transform user relevant huge data. Extracted knowledge as association rules, decision trees or clusters, ensures location of interesting patterns and regularities buried in data and meant to ensure decision making. But such knowledge discovery procedures also return inadvertently, sensitive individual information which compromises their right to privacy [1]. Also, the data mining techniques can open up critical information on business transactions which in turn compromises free competition. Hence there is a need to prevent disclosure of both confidential personal information and that which is contextually sensitive. Research devoted much effort to offset this issue in data mining resulting in many data mining techniques, which included privacy protection mechanisms based on differing approaches [2-5]. An example is the proposal of various sanitization techniques to hide sensitive items/patterns based on removing reserved information/inserting noise into data. Privacy preserving classification methods prevent a miner from classifier construction which could predict sensitive data. Privacy preserving clustering techniques that distort sensitive numerical attributes, while preserving general features was proposed [6, 7].

PPDM was originally meant to extend traditional data mining techniques to work with data modified to hide sensitive information, but the major issue was how to modify data and how to recover data mining results from such modified data. Solutions were usually linked to data mining algorithms under consideration.

The main goals of a PPDM algorithm include:

1. Preventing discovery of sensible information.
2. Being resistant to various data mining techniques.
3. Being uncompromising in access and use of non-sensitive data.
4. Being usable on large amounts of data.
5. It should have less exponential computational complexity.

PPDM techniques usually need data modification to sanitize them from sensitive information (private data items/complex data correlations) or to anonymize them at some uncertainty level. Hence it is important when evaluating a PPDM algorithm to assess transformed data quality. This requires assessment methodologies to assess data quality of individual database items due to a privacy preserving technique, and also information quality from modified data through use of a specific data mining method. Notions of data quality are strictly related to use data for the purpose for which it was intended. Also some algorithms are computationally expensive and so cannot be

used when large data sets are to be released frequently released. So both, data quality and performance should be assessed.

Data mining techniques ensure privacy due to modification. PPDM techniques are based on cryptography, data mining and information hiding [8]. Statistics-based and the crypto-based approaches tackle PPDM. In the former, data owners sanitize data through perturbation or generalization before publishing. Knowledge models like decision trees are used on sanitized data. This approach's advantage is its efficiency in handling large datasets volumes. In the latter approach, data owners cooperatively implement specially designed data mining algorithms [9]. Though such algorithms manage verifiable privacy protection and better data mining performance, performance and scalability issues are its bugbears. Privacy methods like k-anonymity use generalization/suppression techniques to mask an individual's identifiable information. Data is transformed to equivalence classes by K-Anonymity with each class having a set of K- records indistinguishable from each other [10, 11]. Techniques like l-diversity and t-closeness [12] overcame problems mentioned above.

Discovering optimal k-anonymous datasets using generalization or suppression proved to be NP-hard [13, 14]. Minimum data loss is possible through optimizing an aggregated value over all features/records. Swarm intelligence based evolutionary algorithms (EA) use simple entities having limited memory to evolve into improved solutions.

Particle Swarm Optimization (PSO) is a population based heuristic search technique [15], which solves optimization problems modelled on EA. PSO optimises an objective function through a population-based search, the population including potential solutions, named particles that are metaphors for flocks of birds. Particles are randomly initialised flying across a multidimensional search space. Each particle during flight, updates its velocity and position based on best experience of its own and the whole population. Updating policy drives a particle swarm to move to regions with higher objective function values, with all particles finally gathering around a point with highest objective value.

Particle Swarm Optimization differs from other evolutionary algorithms, converging rapidly, with less parameters, encoding with real numbers and directly dealing with the problem domains, without conversion. Hence the algorithm is simple, easy-to implement, widely used, and particularly applicable to continuous function optimization problems. This work proposes Swarm Optimization (PSO) implementation to locate an optimal generalized feature set. K-anonymity is accomplished by original data set generalization in the proposed framework. The paper is organized as follows: Section 2 reviews some related works in the literature, section 3 details the methods, section 4 gives the results and discussions and section 5 concludes the paper.

2. Related works

Data owners must preserve privacy and guarantee valid data mining results to achieve an equitable solution to PPDM. Ravi et al [16] proposed a novel PSO trained auto associative neural network (PSOAANN) for privacy preservation. Then both decision tree and logistic regression are invoked for data mining, resulting in PSOAANN + DT and PSOAANN + LR hybrids. Hybrids efficacy is tested on five benchmark and four bankruptcy datasets with results being compared with those of Ramu and Ravi [17] and others. The proposed hybrids yielded better/comparable results leading to the conclusion that PSOAANN is viable for privacy preservation.

Ramu and Ravi [17] proposed hybridisation of random projection and random rotation methods to classify privacy-preservation the hybrid method being tested on six benchmark data set's and four bank bankruptcy data sets. These methods ensured privacy/secrecy of bank data with the resulting data set being mined without much accuracy loss. Multilayer perceptron, decision tree J48 and logistic regression are used as classifiers with results of a tenfold cross-validation and t-test proving improved average accuracies for the hybrid privacy preservation method compared to a standalone random projection. A reason for the hybrid privacy preservation method's superior performance has been highlighted.

The randomized response (RR) technique promises to disguise PPDM's private categorical data. Though many RR-based methods were proposed for various data mining computations, no systematic study has compared them to locate optimal RR schemes. Comparison difficulties arise due to the need to consider conflicting metrics - privacy and utility - when comparing two PPDM schemes. An optimal scheme based on one metric is the worst when based on the other. Huang et al [18] described a method to quantify privacy and utility. Quantification is formulated as estimate problems, and estimate theories are used to derive quantification. Then an evolutionary multi-objective optimization method finds optimal disguise matrices for randomized response technique. The results prove that the proposed scheme performed better when compared to current RR schemes.

Das [19] proposed a scalable, local privacy-preserving algorithm for distributed peer-to-peer (P2P) data aggregation for advanced data mining/analysis like

average/sum computation, decision tree induction, feature selection, and more. Unlike most multi-party privacy-preserving data mining algorithms, this works in an asynchronous manner through local interactions and is highly scalable dealing specifically with distributed computation of the sum of a numbers set stored at different peers in a P2P network in the context of a P2P web mining application. The proposed optimization-based privacy-preserving technique for computing permits various peers to specify different privacy requirements without adhering to global parameters for the chosen privacy model. As distributed sum computation is a constantly used primitive, the proposed approach can significantly impact many data mining tasks including multi-party privacy-preserving clustering, frequent itemset mining, and statistical aggregate computation.

3. Methodology

3.1 Adult Dataset

UCI Machine Learning Repository provides 'Adult' dataset for evaluation. It contains 48842 instances, including categorical and integer attributes from the 1994 Census and has around 32,000 rows with 4 numerical columns, the column and ranges including age {17 – 90}, fnlwgt {10000 – 1500000}, hrsweek {1 – 100} and edunum {1 – 16}. K-anonymization anonymizes the age column and native country. Table 1 reveals original Adult dataset attributes.

Table 1: Attributes of the Adult dataset

age	native-country	Class
39	United-States	<=50K
50	United-States	<=50K
38	United-States	<=50K
53	United-States	<=50K
28	Cuba	<=50K
37	United-States	<=50K
49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

3.2 K-Anonymity

Data is transformed to equivalence classes in k-anonymity with each class having k-records set different from others [20]. Generalization & suppression reduce granularity representation of pseudo-identifiers techniques. Attribute values are generalized to a range to reduce the granularity (date of birth generalized as year of birth) and reduce identification risk. Attribute value is removed completely to reduce identification risk with public records (suppression). K-anonymity is a good technique due to its simplicity in definition. Many algorithms are available to process anonymization [21, 22].

3.3 The Particle Swarm Optimization (PSO)

The Particle Swarm Optimization (PSO) algorithm is an adaptive algorithm of population of individuals (commonly called particles), adapting through returning stochastically towards earlier successful regions [23]. PSO's two primary operators are Velocity update and Position update. During every iteration, particle is accelerated towards particles in earlier best position and global best position. Each particle's new velocity value is updated at iterations, this being based on current velocity, distance from previous best position, and distance from global best position and this calculates the particle's next position in search space. The process stops either on iteration of specific number of times, or till obtainment of a minimum error [24, 25].

PSO starts with a group of random particles/solutions, searching for optima through updating generations. The two "best" values - pbest and gbest - of a particle is updated with each iteration. 'pbest' is best solution (fitness) achieved till then and 'gbest' best value obtained till then by any population's particle. PSO is computationally simple requiring only primitive mathematical operators. Particle positions/velocities are randomly assigned at the algorithm's beginning. PSO updates all velocities/positions of particles iteratively as follows:

$$v_i^d = wv_i^d + c_1 r_1 \left(p_i^d - x_i^d \right) + c_2 r_2 \left(p_g^d - x_i^d \right)$$
$$x_i^d = x_i^d + v_i^d \qquad (1)$$

Where

v_i^d - new velocity of the i^{th} particle computed based on the particle's previous velocity, distance between previous best position and current position and distance between best particle of the swarm

d - Number of dimensions,

i - Size of the population,

w - Inertia weight,

r1 and r2 are random values in the range [0, 1]

c_1, c_2 are positive constants,

x_i^d - The particle's new position.

In classical PSO, particles are trapped in a local optimum in gbest region if gbest is far away from global optimum. To overcome this, particles fly through a larger search space with a particle's pbest position being updated based on the pbest position of all swarm particles thereby improving swarm diversity and avoiding optimum. The particle's updating velocity is given by:

$$V_i^d = w * v_i^d + c * rand_i^d * \left(pbest_{fi(d)}^d - x_i^d \right) \quad (2)$$

Where $f_i = \left[f_i(1), f_i(2), ..., f_i(d) \right]$ refers to the pbest that the particle i used and $pbest_{fi(d)}$ is the dimension of particle's pbest. Two particles are chosen randomly and one whose velocity is updated is left out. To update velocity, particles pbest's fitness values are compared and best dimension selected.

4. Results and Discussion

Generalization depends on data type, which is either categorical or numeric. Categorical data (gender, work, zip code) generalization is described by a taxonomy tree as seen in Fig. 1. Figure shows a generalization example of continuous data in this work.

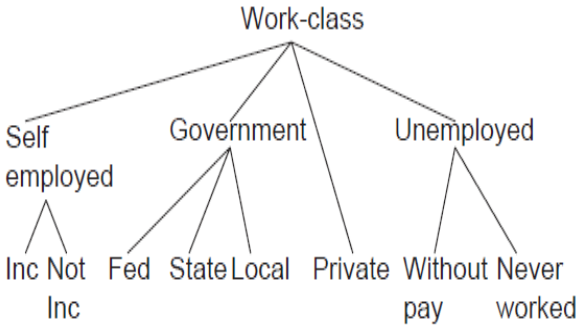

Fig. 1 Example for generalization of continuous data as a taxonomy tree

Generalization of numeric data (age, income) is obtained through discretization of 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). PSO algorithm finds optimal generalization feature set. The following Figures and Tables give results for classification, precision and recall for class label income. The precision/recall is shown for value greater than 50K and less than or equal to 50K.

Table 2: Classification Accuracy for different levels of k-anonymity

k-anonymity level	*Classification accuracy*
K=50	0.828590148
K=45	0.829613857
K=40	0.834978093
K=35	0.834507186
K=30	0.836227018
K=25	0.837045985
K=20	0.850272307
K=15	0.860345604
K=10	0.868043897
K=5	0.8798575

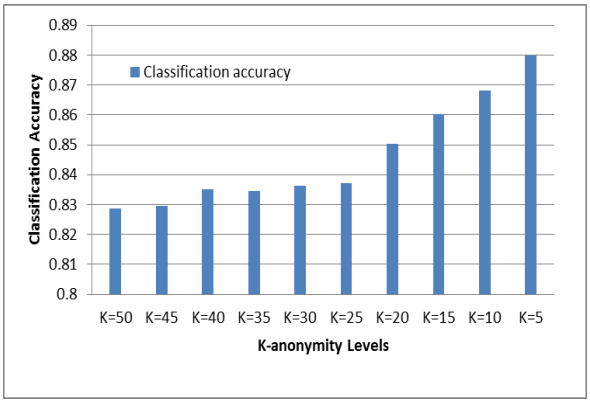

Fig. 2 Classification Accuracy for different levels of k-anonymity

Fig. 2 reveals that classification accuracy decreases when k-anonymity level increases. Fig. 3 and Fig. 4 show precision and recall for class label income greater than 50k and less than or equal to 50k respectively.

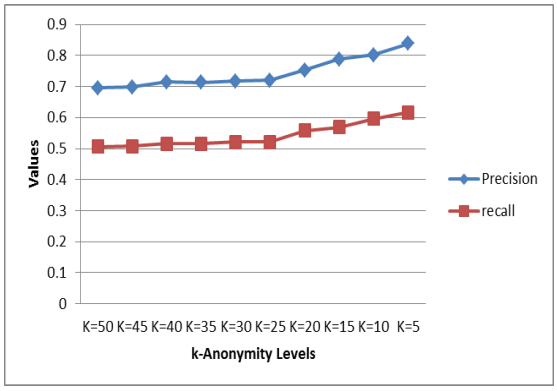

Fig. 3: Precision and Recall for different levels of k-anonymity for class label >50K

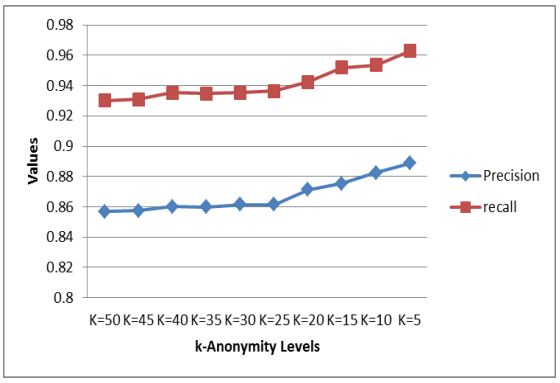

Fig. 4: Precision and Recall for different levels of k-anonymity for class label <=50K

5. Conclusion

Existing Evolutionary Algorithm (EA) solutions in privacy-preserving domain deal mainly with specific issues like cost evaluation. This work proposes Swarm Optimization (PSO) implementation to improve the PPDM process. K-anonymity is accomplished by generalization of the original dataset in the proposed framework. PSO Optimization searches for optimal generalized feature set which leads to better classification. Experiments conducted for various k-anonymity levels provided satisfactory results.

Acknowledgments

We thank immensely our management for extending their support in providing us infrastructure and allowing us to utilize them in the successful completion of our research paper.

References

[1] C. C. Aggarwal, J. Pei, and B. Zhang. On privacy preservation against adversarial data mining. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, August 2006.

[2] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In Proc. of the 24th IEEE International Conference on Data Engineering (ICDE), pages 376–385, April 2008.

[3] C. C. Aggarwal and P. S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. Springer, March 2008.

[4] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. Anonymity preserving pattern discovery. International Journal on Very Large Data Bases (VLDBJ), 17(4):703–727, July 2008.

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

[6] Ali Inan, Selim V. Kaya, Yucel Saygin, Erkay Savas, Ayca A. Hintoglu, and Albert Levi. Privacy preserving clustering on horizontally partitioned data. Data & Knowledge Engineering (DKE), 63(3):646–666, 2007.

[7] S. R. Oliveira and O. R. Zaiane. A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Journal on Computers and Security, 26(1):81–93, 2007.

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

[9] 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.

[10] Nergiz, M. E., Clifton, C., & Nergiz, A. E. (2009). Multirelational k-anonymity. Knowledge and Data Engineering, IEEE Transactions on, 21(8), 1104-1117.

[11] 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.

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

[13] 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.

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

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

[16] Ravi, V., Nekuri, N., & Rao, C. R. (2012). Privacy preserving data mining using particle swarm optimisation trained auto–associative neural network: an application to bankruptcy prediction in banks. International Journal of Data Mining, Modelling and Management, 4(1), 39-56.

[17] Ramu, K., & Ravi, V. (2009). Privacy preservation in data mining using hybrid perturbation methods: an application to bankruptcy prediction in banks.International Journal of Data Analysis Techniques and Strategies, 1(4), 313-331.

[18] Huang, Z., & Du, W. (2008, April). OptRR: Optimizing randomized response schemes for privacy-preserving data mining. In Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on (pp. 705-714). IEEE.

[19] Das, K., Bhaduri, K., & Kargupta, H. (2011). Multi-objective optimization based privacy preserving distributed data mining in Peer-to-Peer networks. Peer-to-Peer Networking and Applications, 4(2), 192-209.

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

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

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

[23] 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 Internatonal Conference on , vol., no., pp.1-4, 8-11 Oct. 2007

[24] 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.

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

Sridhar Mandapati obtained his masters degree in Computer Applications from S.V University, Tirupathi. He is currently working as Associate Professor in the Department of Computer Applications at R.V.R & J.C College of Engineering, Guntur. He has 14 years of teaching experience. At present he is pursuing Ph.D. from Acharya Nagarjuna University, Guntur. His are of research interest include Data Mining, Information Security and Image Processing.

Dr.**B.Raveendra Babu** obtained his Masters in Computer Science and Engineering from Anna University, Chennai. He received his Ph.D. in Applied Mathematics at S.V University, Tirupati. He is now working as Professor in the Department of Computer Science and Engineering, VNR VJIET, Hyderabad. He has 27 years of teaching experience. He has more than 55 International and National publications to his credit. His area of research interest includes Data Mining, Image Processing, Pattern Analysis and Information Security.

Dr.**M.V.P.Chandra Sekhara Rao** obtained his Masters in Computer Science and Engineering from JNTU, Kakinada. He received his Ph.D. in Computer Science and Engineering from JNTU, Hyderabad. At present working as Professor in the Department of Computer Science and Engineering at R.V.R & J.C College of Engineering, Guntur. He has 16 years of teaching experience. He has six international publications and attended several international conferences. His area of research interest is Data Warehouse and Data Mining.