Research Journal of Applied Sciences, Engineering and Technology 9(8): 616-621, 2015
DOI:10.19026/rjaset.9.1445
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 10, 2014 Accepted: November 3, 2014 Published: March 15, 2015

Research Article

Privacy Preserving Data Mining

[1]A.T. Ravi and [2]S. Chitra
[1]Department of Computer Science and Engineering, SSM College of Engineering, Komarapalayam, India
[2]Department of Computer Science and Engineering, Er. Perumal Manimekalai College of Engineering, India

Abstract: Recent interest in data collection and monitoring using data mining for security and business-related applications has raised privacy. Privacy Preserving Data Mining (PPDM) techniques require data modification to disinfect them from sensitive information or to anonymize them at an uncertainty level. This study uses PPDM with adult dataset to investigate effects of K-anonymization for evaluation metrics. This study uses Artificial Bee Colony (ABC) algorithm for feature generalization and suppression where features are removed without affecting classification accuracy. Also k-anonymity is accomplished by original dataset generalization.

Keywords: Adult dataset, Artificial Bee Colony (ABC) algorithm, data mining, K-anonymization, Privacy Preserving Data Mining (PPDM)

INTRODUCTION

Data mining techniques extract knowledge to support various domains like marketing, medical diagnosis, weather forecasting and national security. Even then it is a challenge to mine some data types without violating data owners ’privacy. Most organizations collect information about individuals for their specific needs, but must ensure that individual privacy is not violated or sensitive business information revealed. To avoid these violations, various PPDM techniques are needed (Nayak and Swagatika, 2011).

PPDM is a research area from 1991 in the public and private sectors. PPDM refers to data mining that tries to safeguard sensitive information from unsolicited/unsanctioned disclosure. Traditional data mining techniques analyze and model data set statistically in aggregation, while privacy preservation is about protecting against disclosure of individual data records. This domain separation points to PPDM’s technical feasibility (Evfimievski and Grandison, 2009). Recently, collecting and monitoring data using data mining technology raised concerns about privacy issues for security and business-related applications (Singh *et al.*, 2011; Mandapati *et al.*, 2013). PPDM algorithms extract relevant knowledge from large voluminous data while protecting sensitive information simultaneously. An important aspect in such algorithms design is identifying evaluation criteria and developing related benchmarks (Bertino *et al.*, 2008).

Discretization is resorted to hide individual values. Value Distortion Return a value xi+r instead of xi where r is a random value from some distribution. Two random distributions like uniform and Gaussian are considered. In uniform distribution, random variable is between (-a.+a) where mean is 0. In Gaussian distribution, mean $\mu = 0$ and standard deviation (Jha and Barot, 2014). Most privacy computation methods use some transformation on data to perform privacy preservation. But, they reduce representation granularity to reduce privacy. Reduction in granularity leads to loss of data management or mining algorithms effectiveness, a trade-off between information loss and privacy. Such techniques are: randomization method (Agrawal and Srikant, 2000; Agrawal and Aggarwal, 2002), k-anonymity model and l-diversity (Machanavajjhala *et al.*, 2007), distributed privacy preservation and downgrading application effectiveness.

k-anonymity captures protection of released data against re-identification of respondents to whom released data refers. k-anonymity demands that each tuple in a private table being released be indistinguishably related to k respondents. As it seems impossible or impractical and limiting to assume as to which is a potential attacker and can (re-) identify respondents, k-anonymity requires that respondents be indistinguishable (within a given individuals set) in the released table itself regarding attributes set, called quasi-identifier, which can be exploited for linking.

In k-anonymity techniques, pseudo-identifiers granularity representation is reduced with techniques like generalization and suppression. In generalization, attribute values are generalized to a range so as to

Corresponding Author: A.T. Ravi, Department of Computer Science and Engineering, SSM College of Engineering, Komarapalayam, India

reduce representation granularity. Generalization, replaces attribute values with their generalized version and it is generalization hierarchy based and a corresponding value generalization hierarchy on domain's values. Domain generalization hierarchy is total order and the corresponding value generalization a hierarchy tree, where a parent/child relationship represents direct generalization/specialization relationship (Machanavajjhala *et al*., 2007). In suppression, attribute value is removed completely. These methods reduce risk of identification with public records, while reducing applications accuracy on transformed data (Aggarwal and Philip, 2008).

LITERATURE REVIEW

A new PPDM framework of multi-dimensional data proposed by Aggarwal and Yu (2008) developed a new and flexible PPDM approach without needing new problem-specific algorithms, as it mapped original data set to new anonymized data set. The anonymized data closely matches characteristics of original data including correlations among different dimensions.

An anonymization of query logs using micro-aggregation was proposed by Navarro-Arribas *et al.* (2012) which ensured the k-anonymity of users in all query logs, while preserving utility. This system evaluated real query logs, showing privacy and utility achieved and ensured estimations for use of data in clustering based data mining processes.

A strategy to protect data privacy during decision tree analysis of data mining process was proposed by Kadampur (2010). This adds specific noise to numeric attributes after exploring original data decision tree. Obfuscated data is presented to second party for decision tree analysis. Decision tree got on original data and obfuscated data are similar but the new method fails to reveal data proper to second party during mining, thereby preserving privacy.

A perturbation-based PPDM for Multi-Level Trust (MLT-PPDM) was proposed by Li *et al.* (2012). In this setting, a malicious data miner could access differently perturbed copies of same data and combines diverse copies to jointly infer additional information about original data that a data owner did not plan to release. The new system addresses this by properly correlating perturbation across copies at various trust levels.

A framework for multiple parties to do privacy-preserving association rule mining was presented by Zhan (2008). Issues of privacy preserving collaborative data mining were considered with binary data sets input. The new system ensured efficient association rule mining to ensure such computation. A secure protocol, called number product was developed, for multiple-parties to jointly conduct desired computations.

A new perturbation based technique proposed by Liu *et al.* (2009) modified data mining algorithms to ensure their being used directly on perturbed data i.e., it directly builds a classier for original data set from perturbed training data set. The new algorithm decreases communication and computation cost compared to cryptography based approaches. The algorithm is based on perturbation scheme and increased privacy protection with reduced computation time.

An evolutionary privacy-preserving data mining technology to locate an appropriate method to perform secure transactions in a database was proposed by Patel *et al.* (2013) to present popular PPDM approaches, namely: randomization, suppression, cryptography and summarization. Privacy guarantees, advantages and disadvantages of every approach provided a balanced view of the state of the art technique.

Three representative multiplicative perturbation methods like rotation perturbation, projection perturbation and geometric perturbation were proposed by Chen and Liu (2008). The new system discussed appropriate privacy evaluation models design for multiplicative perturbations, giving an overview of how privacy evaluation measures privacy guarantee levels in different types of attacks context.

A different approach to achieve k-anonymity by partitioning original dataset into many projections so that each adhered to k-anonymity was proposed by Matatov *et al.* (2010). The new Data Mining Privacy by Decomposition (DMPD) algorithm uses a genetic algorithm to locate optimal feature set partitioning. Ten separate datasets were evaluated with DMPD to compare its classification performance with other k-anonymity-based methods. Results suggest that DMPD performed better than current k-anonymity-based algorithms. Also, there was no need to apply domain dependent knowledge.

How one ensured an individual's privacy regarding his location and spatiotemporal behavioral patterns was proposed by Ho (2012) through differential privacy mechanism which assumes that data trajectory is secure and users can only query knowledge derived from it. The proposed system demonstrated privacy preserving approach on frequent location pattern mining tasks.

Distributed homogenous database algorithm, a modification of privacy preserving association rule mining was proposed by Hussein *et al.* (2008) which was faster than the earlier one which modified privacy preservation with accurate results. The modified algorithm was based on a semi-honest model with negligible collision probability. It was possible to extend flexibility to many sites without implementation changes.

An individually adaptable perturbation model, enabling individuals to choose their own privacy levels was proposed by Liu *et al.* (2008). The new approach's electiveness was demonstrated by experiments on synthetic and real-world data sets. It gave a simple but effective/efficient technique to build data mining models from perturbed data based on this experiment.

The design and security requirements for large-scale PPDM systems in fully distributed settings, where each client possesses own private data records were discussed by Magkos *et al*. (2009). This framework was based on classical homomorphic election model and specifically on an extension to support multi-candidate elections.

A problem of collusions, where some parties collude and share records to deduce private information of other parties, proposed by Yang *et al*. (2010), was a method that entailed high level of full-privacy security. This method ensured that no sensitive information of a party would be revealed even when other parties colluded. Also, this method was efficient with a running time of O (m). This general method was applicable on many PPDM problems which are solved with enhanced security.

A work that followed the line of research suggesting that many data mining problems are realized in a privacy-preserving setting by designing techniques and can be decomposed into secure evaluations of addition, multiplication, comparison and division was proposed by Blanton (2011). The new system revealed how efficient solutions secure in semi-honest and malicious models are developed in a framework.

A comparative study between multi agent based data mining and high-performance privacy preserving data mining was discussed by Farooqui *et al*. (2010). This study provides a detailed analysis of agent framework for data mining along with overall architecture and functionality. Challenges in developing PPDM algorithms with existing frameworks, motivating design of a new infrastructure based on these challenges is also discussed.

METHODOLOGY

The proposed framework models ABC for feature generalization and suppression. In the new algorithm, features removable without affecting classification accuracy are suppressed. Also, k-anonymity is accomplished by original dataset generalization. ABC finds features range during generalization for varying k values.

Adult dataset: 'Adult' dataset used for evaluation from UCI Machine Learning Repository (Asuncion and Newman, 2007) contains 48,842 instances with both categorical and integer attributes from 1994 Census. The dataset contains 32,000 rows with 4 numerical columns whose ranges are: age (17-90), fnlwgt (10000-1500000), hrsweek (1-100) and edunum (1-16). The age column and native country are anonymized using k-anonymization principles. Table 1 shows original data.

K-anonymity: *k*-anonymization was introduced by Samarati (2001) and Sweeney (2002). A database is *k*-anonymous regarding quasi-identifier attributes (a set of attributes used with certain external information to identify specific individuals) if there are *k* transactions in a database having same values according to quasi-identifier attributes. In practice, to protect sensitive dataset *T*, before releasing *T* to public, it is converted to a new dataset *T* guaranteeing the *k*-anonymity property for a sensible attribute by performing value generalizations on quasi-identifier attributes. Hence, sensitive attributes degree of uncertainty is at least 1/*k*.

Table 1: The original attributes of 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

Table 2: The k-anonymous dataset

Age	Native-country	Class
Adult	United-States	< = 50K
Middle aged	United-States	< = 50K
Adult	United-States	< = 50K
Middle aged	United-States	< = 50K
Adult	US-oth	< = 50K
Adult	United-States	< = 50K
Middle aged	African	< = 50K
Middle aged	United-States	>50K
Adult	United-States	>50K
Adult	United-States	>50K

K-anonymity specifically focuses on two techniques: Generalization and suppression, which, unlike current techniques like scrambling/swapping, preserve information fidelity. Generalization substitutes a given attribute's values with general values. For this, the domain idea captures generalization assuming existence of generalized domains set. The original domain set with generalizations is D_{om}. Each generalized domain has generalized values and mapping between them and its generalizations does exist.

Mapping is stated by generalization relationship ≤D. Given two domains *Di* and *Dj*ϵ D_{om}, *Di*≤*Dj* states values in domain *Dj*≤ *D* are generalizations of values in *Di*. Generalization relationship ≤*D* defines partial order on set D_{om} of domains, requiring satisfaction of following conditions in Eq. (1) and (2):

$$C1: \forall D_{i,} D_{j,} D_z \in D_{om} \tag{1}$$

$$C2: D_i \leq_D D_{j,} D_i \leq_D D_z \Longrightarrow D_j \leq_D D_{z,} D_z \leq_D D_j \tag{2}$$

C2: All maximal elements of D_{om} are singleton. Condition C1 states that for every domain *Di*, domains set generalization of *Di* is totally ordered and, so each *Di* has at most one direct generalization domain *Dj*. It ensures determinism in generalization. Condition C2 ensures all values in each domain are generalized to single value. Generalization relationship definition implies existence of each domain *D*ϵD_{om}, a totally ordered hierarchy, called domain generalization hierarchy, denoted DGHD (Ciriani *et al*., 2008). Table 2 shows modified attribute data of k-anonymous dataset.

Naïve bayes classification: The Bayes classification proposed is based on Bayes rule of conditional probability. Bayes rule estimates likelihood of a property given, data as evidence or input Bayes rule or Bayes theorem in Eq. (3):

$$P(h_i \mid x_i) = \frac{P(x_i \mid h_i)P(h_i)}{P(x_i \mid h_i) + P(x_i \mid h_2)P(h_2)} \tag{3}$$

This approach is termed "naïve" as it assumes independence between various attribute values. Naive Bayes (Pandey and Pal, 2011) classifier is Bayes theorem based with strong (Naive) independence assumption and suits cases having high input dimensions. Naïve Bayes classification can be viewed as descriptive and predictive algorithms. Probabilities used to predict class membership for a target tuple are descriptive.

Proposed feature suppression technique: ABC algorithm optimizes feature suppression process as yielding best optimal features cannot be removed during anonymization without affecting classification accuracy. ABC algorithm is developed by inspecting real bee behavior when a food source is located. The source is called nectar and food sources information is shared with bees in the nest. In ABC, artificial agents are classified as employed bee, onlooker bee and scout (Karakos *et al*., 2011). Each plays a different role: employed bee stays at a food source and provides the neighbourhood of the source in its memory; the onlooker gets food source information from employed bees in the hive and selects one to gather nectar; the scout is responsible to find new food/new nectar, sources.

The process starts when bees leave hive to search for a food source (nectar). After locating it bees store it in their stomach. After returning to the hive, bees unload nectar and perform a waggle dance to share information about food source (nectar quantity, distance and direction from source to hive) and recruit new bees to explore rich food sources. Unlike optimization problems, where possible problem solutions can be represented by vectors with real values, candidate solutions to feature selection issues are represented by bit vectors (Tsai *et al*., 2009). Every food source is associated with a bit vector of size N, where N is total number of features. The position in vector corresponds to features number needing evaluation. If value at corresponding position is 1, it indicates that a feature is part of subset needing evaluation. ABC algorithm steps are (Schiezaro and Pedrini, 2013):

1. Initialize the food source positions
2. Evaluate the food sources
3. Produce new food sources (solutions) for the employed bees
4. Apply greedy selection
5. Calculate the fitness and probability values
6. Produce new food sources for onlookers
7. Apply greedy selection
8. Determine the food source to be abandoned and allocate its employed bee as a scout for searching the new food sources
9. Memorize the best food source found
10. Repeat steps 3-9 for a pre determined number of iterations

In new suppression technique binary encoding with 0 representing feature not selected is used and 1 represents feature selected.

Proposed generalization technique: Features are generalized using ABC. ABC steps are similar to steps explained in the earlier section, the difference being the initial population is taken from output of the new feature suppression technique. In the new generalization technique, features optimal generalization range is achieved.

RESULTS AND DISCUSSION

In this study, Adult dataset is used. Table 3 to 5 shows the result value for classification accuracy, precision and recall respectively. Figure 1 to 3 shows the same.

Table 3 and Fig. 1 reveal that classification accuracy decreases when k-anonymity level increases.

Table 3: Classification accuracy

	Without anonymization	ABC-only suppression	ABC-suppression and generalization
K = 1	0.9009	0.8979	0.8942
K = 10	0.8961	0.8894	0.8894
K = 20	0.8911	0.8881	0.8844
K = 30	0.8829	0.8759	0.8721
K = 40	0.8762	0.8733	0.8695
K = 50	0.8677	0.8648	0.8610
K = 60	0.8608	0.8559	0.8521
K = 70	0.8566	0.8536	0.8499
K = 80	0.8536	0.8402	0.8469
K = 90	0.8530	0.8500	0.8463

Table 4: Precision

	Without anonymization	ABC-only suppression	ABC-suppression and generalization
K = 1	0.881743	0.878962	0.875353
K = 10	0.879367	0.875088	0.872923
K = 20	0.876843	0.872907	0.870338
K = 30	0.872900	0.868330	0.865489
K = 40	0.868920	0.864548	0.861519
K = 50	0.864673	0.860423	0.857240
K = 60	0.860414	0.856000	0.852679
K = 70	0.856488	0.852150	0.848717
K = 80	0.852948	0.847647	0.845137
K = 90	0.849952	0.844769	0.842107

Table 5: Recall

	Without anonymization	ABC-only suppression	ABC-suppression and generalization
K = 1	0.835926	0.829808	0.821936
K = 10	0.832584	0.823317	0.818594
K = 20	0.829176	0.820696	0.815186
K = 30	0.825572	0.815967	0.810027
K = 40	0.821852	0.812745	0.806529
K = 50	0.817619	0.808863	0.802452
K = 60	0.813455	0.804535	0.797978
K = 70	0.809679	0.801003	0.794332
K = 80	0.806329	0.796168	0.791086
K = 90	0.803487	0.793605	0.788331

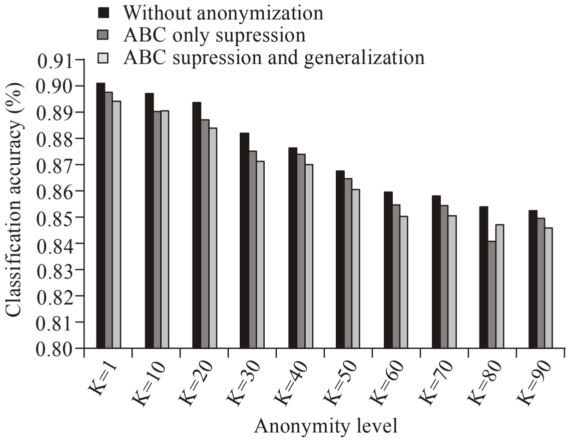

Fig. 1: Classification accuracy

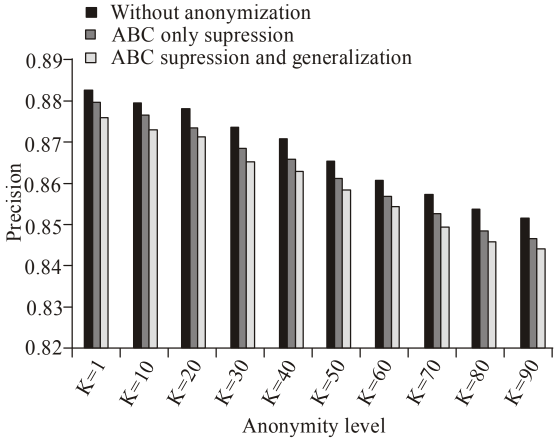

Fig. 2: Precision

The classification accuracy of proposed method reduces in the range of 0 to 1.2308% as the anonymity level increases from 1 to 90.

Table 4 and Fig. 2 reveal that precision decreases when k-anonymity level increases. The recall of proposed method reduces in the range of 0.2477 to 1.9273% as the anonymity level increases from 1 to 90.

Table 5 and Fig. 3 reveal that recall decreases when k-anonymity level increases. The recall of proposed method reduces in the range of 0.5753 to 1.9209% as the anonymity level increases from 1 to 90.

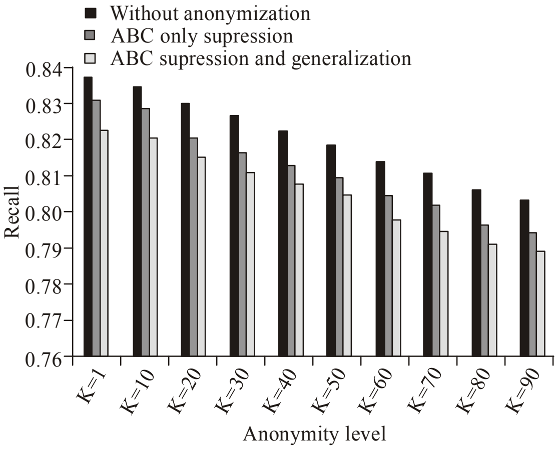

Fig. 3: Recall

CONCLUSION

Data mining technologies enable commercial and governmental organizations to extract knowledge from data for business/security related applications. While successful applications are encouraging, concerns increase about invasion of personal information privacy. Results show that classification accuracy decreases when k-anonymity level increase. Classification accuracy of the new method reduces in a 0 to 1.2308% range as anonymity increases from 1 to 90.

REFERENCES

Aggarwal, C.C. and P.S. Yu, 2008. On static and dynamic methods for condensation-based privacy-preserving data mining. ACM T. Database Syst., 33(1): 2.

Aggarwal, C.C. and S.Y. Philip, 2008. A General Survey of Privacy-Preserving Data Mining Models and Algorithms. Springer, US, pp: 11-52.

Agrawal, R. and R. Srikant, 2000. Privacy-preserving data mining. Proceeding of the ACM SIGMOD International Conference on Management of Data, pp: 439-450.

Agrawal, D. and C.C. Aggarwal, 2002. On the design and quantification of privacy preserving data mining algorithms. Proceeding of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'01), pp: 247-255.

Asuncion, A. and D. Newman, 2007. UCI Machine Learning Repository. Retrieved from: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Bertino, E., D. Lin and W. Jiang, 2008. A survey of quantification of privacy preserving data mining algorithms. Lect. Notes Comput. Sc., 34: 183-205.

Blanton, M., 2011. Achieving full security in privacy-preserving data mining. Proceeding of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE 3rd International Conference on Social Computing (SOCIALCOM), pp: 925-934.

Chen, K. and L. Liu, 2008. A Survey of Multiplicative Perturbation for Privacy-preserving Data Mining. Privacy-preserving Data Mining. Springer, US, pp: 157-181.

Ciriani, V., S.D.C. di Vimercati, S. Foresti and P. Samarati, 2008. k-anonymous data mining: A survey. Lect. Notes Comput. Sc., 34: 105-136.

Evfimievski, A. and T. Grandison, 2009. Privacy-preserving Data Mining. Handbook of Research on Innovations in Database Technologies and Applications: Current and Future Trends. IGI Global, pp: 1-8.

Farooqui, M.F., M. Muqeem and M.R. Beg, 2010. A comparative study of multi agent based and high-performance privacy preserving data mining. Int. J. Comput. Appl., 4(12): 23-26.

Ho, S.S., 2012. Preserving privacy for moving objects data mining. Proceeding of the IEEE International Conference on Intelligence and Security Informatics (ISI), pp: 135-137.

Hussein, M., A. El-Sisi and N. Ismail, 2008. Fast cryptographic privacy preserving association rules mining on distributed homogenous data base. Lect. Notes Comput. Sc., 5178: 607-616.

Jha, M.K.M. and M. Barot, 2014. Privacy preserving data mining. Int. J. Futurist. Trends Eng. Tech., 4(1).

Kadampur, M.A., 2010. A noise addition scheme in decision tree for privacy preserving data mining. J. Comput., 2(1): 137-144.

Karakos, D., M. Dredze, K. Church, A. Jansen and S. Khudanpur, 2011. Estimating document frequencies in a speech corpus. Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU, 2011), pp: 407-412.

Li, Y., M. Chen, Q. Li and W. Zhang, 2012. Enabling multilevel trust in privacy preserving data mining. IEEE T. Knowl. Data En., 24(9): 1598-1612.

Liu, L., M. Kantarcioglu and B. Thuraisingham, 2008. The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng., 65(1): 5-21.

Liu, L., M. Kantarcioglu and B. Thuraisingham, 2009. Privacy preserving decision tree mining from perturbed data. Proceeding of the 42nd Hawaii International Conference on System Sciences (HICSS'09), pp: 1-10.

Machanavajjhala, A., D. Kifer, J. Gehrke and M. Venkitasubramaniam, 2007. l-diversity: Privacy beyond k-anonymity. ACM T. Knowl. Discov. Data, 1(1): 3.

Magkos, E., M. Maragoudakis, V. Chrissikopoulos and S. Gritzalis, 2009. Accurate and large-scale privacy-preserving data mining using the election paradigm. Data Knowl. Eng., 68(11): 1224-1236.

Mandapati, S., R.B. Bhogapathi, M.C.S. Rao and V. Vjiet, 2013. Swarm optimization algorithm for privacy preserving in data mining. Int. J. Comput. Sci. Issues, 10(2).

Matatov, N., L. Rokach and O. Maimon, 2010. Privacy-preserving data mining: A feature set partitioning approach. Inform. Sciences, 180(14): 2696-2720.

Navarro-Arribas, G., V. Torra, A. Erola and J. Castellà-Roca, 2012. User< i> k</i>-anonymity for privacy preserving data mining of query logs. Inform. Process. Manag., 48(3): 476-487.

Nayak, G. and D. Swagatika, 2011. A survey on privacy preserving data mining: Approaches and techniques. Int. J. Eng. Sci. Technol., 3.3(2011): 2117-2133.

Pandey, U.K. and S. Pal, 2011. Data mining: A prediction of performer or underperformer using classification. Int. J. Comput. Sci. Inform. Technol., 2(2): 686-690.

Patel, M., P. Richariya and A. Shrivastava, 2013. A review paper on privacy preserving data mining. J. Eng. Technol., 2: 359-361.

Samarati, P., 2001. Protecting respondents identities in microdata release. IEEE T. Knowl. Data En., 13(6): 1010-1027.

Schiezaro, M. and H. Pedrini, 2013. Data feature selection based on artificial bee colony algorithm. EURASIP J. Image Video Process., 2013(1): 1-8.

Singh, U.K., K.P. Bhupendra and D. Keerti, 2011. An overview on privacy preserving data mining methodologies. Int. J. Eng. Trends Technol., 2(2).

Sweeney, L., 2002. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz., 10(5): 571-588.

Tsai, P.W., J.S. Pan, B.Y. Liao and S.C. Chu, 2009. Enhanced artificial bee colony optimization. Int. J. Innov. Comput. I., 5(12): 5081-5092.

Yang, B., H. Nakagawa, I. Sato and J. Sakuma, 2010. Collusion-resistant privacy-preserving data mining. Proceeding of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 483-492.

Zhan, J., 2008. Privacy-preserving collaborative data mining. IEEE Comput. Intell. M., 3(2): 31-41.