



INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A New Hybrid Algorithm for Privacy Preserving Data Mining

S.Chidambaranathan

Department of MCA, St. Xavier's College (Autonomous), India

scharan2009@rediffmail.com

Abstract

The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge- and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. This approach alone may lead to excessive data distortion or insufficient protection. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Recently, PPDP has received considerable attention in research communities, and many approaches have been proposed for different data publishing scenarios. In privacy-preserving domain, the existing EA solutions are restricted to specific problems such as cost function evaluation. In this work, it is proposed to implement a Hybrid Evolutionary Algorithm using Genetic Algorithm (GA) and Ant colony Optimization (ACO). Both GA and ACO in the proposed system work with the same population. In the proposed framework, l-diversity is accomplished by Slicing approach of the original dataset. The hybrid optimization is used to search for optimal generalized feature set.

Keywords: Slicing, k-anonymity, Ant colony Optimization, Genetic Algorithm.

Introduction

Privacy-Preserving publishing of microdata has been studied extensively in recent years. Microdata contain records each of which contains information about an individual entity, such as a person, a household, or an organization. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties.

In this information age, data are increasingly being collected by various organizations and government agencies for the purpose of data analysis. To facilitate data analysis [1], it is often necessary to publish the data which, however, poses privacy risks to the individuals. A typical solution is to anonymize the data and release an anonymized version of the data. The goal of data anonymization is to provide

privacy protection for the individuals while allowing adhoc queries and analysis on the anonymized data.

Privacy preserving of data must safeguard from divulging sensitive data during publication of individual data. To maintain privacy, a number of techniques have been presented for modifying or transforming the data. Many data mining techniques are modified to ensure privacy.

The techniques for PPDM are based on cryptography, data mining and information hiding [1]. 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 [2]. In the crypto-based PPDM approach, data owners have to cooperatively implement specially designed data mining algorithms [3]. Though these algorithms achieve verifiable privacy protection and better data mining performance, it suffers from performance and scalability issues [4].

In recent research methods such as generalization [5], [6] for k-anonymity [7] and bucketization [8], [9], [10] for l-diversity [11] are used mostly. In both approaches, attributes are partitioned into three categories: 1) some attributes are identifiers that can uniquely identify an individual, such as Name or Social Security Number; 2) some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other publicly available databases) and which, when taken together, can potentially identify an individual, e.g., Birthdate, Sex, and Zipcode; 3) some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary.

In both generalization and bucketization, one first removes identifiers from the data and then partitions tuples into buckets. The two techniques differ in the next step. Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values.

The problem of discovering optimal l-diversity datasets using slicing or suppression has been proved to be NP-hard [11, 12]. Minimum data loss can be achieved by optimizing an aggregated value over all features and records. The Evolutionary Algorithms (EA) based on hybrid swarm intelligence used simple entities with limited memory evolving into increasingly better solutions.

The efficient swarm-based data mining approaches usually are some kind of hybrid approach; such as combining a swarm intelligence technique with some orthodox optimizing ACO-based clustering technique where the solution is obtained by k-means clustering [13] or combine several swarm-based approaches, such as the ACO technique [14].

Literature review

Bayardo, et al., [15] presented an optimization algorithm for k-anonymization. The presented method searches the space of possible anonymization and forms strategies to reduce computation. The census data was used for evaluation and experiments show that the presented method achieves optimal k-anonymizations using a wide range of k. The effects of different coding approaches and quality of anonymization and performance were also investigated. Real census data experiments demonstrated that the presented algorithm could locate optimal k-anonymizations under two representative cost measures and a wide range of k.

The organizations share their data with many other research communities for various uses. Today technologies are providing easy way of information sharing. However sharing the data with outsiders should not reveal the individual identification of a person[1]. Care must be taken to provide the privacy for the person specific data at the time of publishing personal information for research purposes.

The objective of privacy preserving mining is that this data, when published should not link back to the individual. The notion of k-anonymity was presented in [17], and generalization was used to achieve k-anonymity in Datafly system [18] and μ -Argus system [19]. All these works considered a single data source; therefore, data integration is not an issue. In the case of multiple private databases, joining all databases and applying a single table method would violate the privacy constraint private databases.

Information integration has been an active area of database research. This literature typically assumes that all information in each database can be freely shared [20]. Secure multiparty computation (SMC), on the other hand, allows sharing of the computed result, but completely prohibits sharing of data [21].

Liang et al. [23] and Agrawal et al. [20] presented the notion of minimal information sharing for computing queries spanning private databases. They considered computing intersection, intersection size, equijoin and equijoin size. Their model still prohibits the sharing of databases themselves. The sharing of data has posed several threats leading to individual identification.

Owing to this, privacy preserving data publication has become an important research problem [24]. The main goals of this problem are to preserve privacy of individuals while revealing useful

information. An organization may implement and follow its privacy policy. But when two companies share information about a common set of individuals, and if their privacy policies differ, it is likely that there is privacy breach unless there is a common policy. One such solution was presented for such a scenario, based on k-anonymity and cut-tree method for 2-party data. This paper suggests a simple solution for integrating nparty data using dynamic programming on subsets. The solution is based on thresholds for privacy and informativeness based on k-anonymity.

Kumari, et al., [25] suggested a holistic approach to achieve maximum privacy without information loss and minimum overheads. Studies showed that l-diversity and t-closeness techniques increased computational effort to infeasible levels, while increasing privacy. A few techniques account for maximum information loss when achieving privacy.

The presented method addresses this problem using fuzzy set approach which is a total paradigm shift and a new way of looking at data publishing privacy problem. This method allows personalized privacy preservation being useful for both numerical and categorical attributes and only necessary tuples are transformed.

Shang, et al., [26] presented a novel scheme for selective content distribution encoded as documents, preserving user privacy based on an efficient and novel group key management scheme. The presented approach is based on access control policies that specify which user can access either documents or sub-documents.

On this basis, a broadcast document is divided into multiple subdocuments. Each subdocument is encrypted with a different key. Conforming to modern attribute-based access control, policies are specifically against user identity attributes. But this approach preserves privacy such that users get access to specific documents, or subdocument, based on policies without needing to provide information about identity attributes to the publisher.

Under this approach, the document publisher does not learn identity values of users, and also does not know what policy conditions are verified by users which in turn prevents inferences about identity attributes values being prevented. Also, the presented key management scheme on which the broadcasting approach is based is efficient as decryption keys need not be sent to users together with the encrypted document.

Users can reconstruct keys to decrypt the authorized document portions of a document based on subscription information from the document publisher. Another advantage is that the scheme efficiently handles user's new and revoked subscriptions.

Privacy preserving data publishing

A typical scenario for data collection and publishing is described in Figure 1. In the *data collection* phase, the *data publisher* collects data from *record owners* (e.g., Alice and Bob). In the *data publishing* phase, the data publisher releases the collected data to a data miner or to the public, called the *data recipient*, who will then conduct data mining on the published data.

Data mining has a broad sense, not necessarily restricted to pattern mining or model building. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data publisher, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis.

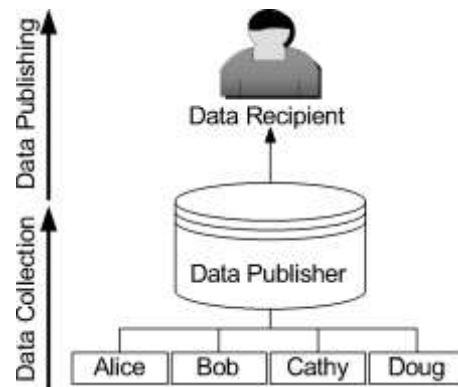


Fig. 1. Data collection and data publishing.

There are two models of data publishers [27]. In the untrusted model, the data publisher is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions [28]; anonymous communications [29]; and statistical methods were proposed to collect records anonymously from their owners without revealing the owners' identity.

In the trusted model, the data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not transitive to the data

recipient. The trusted model of data publishers and consider privacy issues in the data publishing phase.

In practice, every data publishing scenario has its own assumptions and requirements of the data publisher, the data recipients, and the data publishing purpose. The following are several desirable assumptions and properties in practical data publishing:

The nonexpert data publisher. The data publisher is not required to have the knowledge to perform data mining on behalf of the data recipient. Any data mining activities have to be performed by the data recipient after receiving the data from the data publisher.

Sometimes, the data publisher does not even know who the recipients are at the time of publication, or has no interest in data mining. For example, the hospitals in California publish patient records on the Web [Carlisle et al. 2007] [30]. The hospitals do not know who the recipients are and how the recipients will use the data.

The hospital publishes patient records because it is required by regulations [Carlisle et al. 2007] [30] or because it supports general medical research, not because the hospital needs the result of data mining. Therefore, it is not reasonable to expect the data publisher to do more than anonymize the data for publication in such a scenario.

In other scenarios, the data publisher is interested in the data mining result, but lacks the in-house expertise to conduct the analysis, and hence outsources the data mining activities to some external data miners.

In this case, the data mining task performed by the recipient is known in advance. In the effort to improve the quality of the data mining result, the data publisher could release a customized data set that preserves specific types of patterns for such a data mining task. Still, the actual data mining activities are performed by the data recipient, not by the data publisher.

The data recipient could be an attacker. In PPDP, one assumption is that the data recipient could also be an attacker. For example, the data recipient, says a drug research company, is a trustworthy entity; however, it is difficult to guarantee that all staff in the company is trustworthy as well. This assumption makes the PPDP problems and solutions very different from the encryption and cryptographic approaches, in which only authorized and trustworthy recipients are given the private key for accessing the cleartext.

A major challenge in PPDP is to simultaneously preserve both the privacy and information usefulness in the anonymous data.

Publish data, not the data mining result. PPDP emphasizes publishing data records about individuals (i.e., micro data). Clearly, this requirement is more stringent than publishing data mining results, such as classifiers, association rules, or statistics about groups of individuals.

For example, in the case of the Netflix data release, useful information may be some type of associations of movie ratings. However, Netflix decided to publish data records instead of such associations because the participants, with data records, have greater flexibility in performing the required analysis and data exploration, such as mining patterns in one partition but not in other partitions; visualizing the transactions containing a specific pattern; trying different modeling methods and parameters, and so forth.

The assumption for publishing data and not the data mining results, is also closely related to the assumption of a nonexpert data publisher. For example, Netflix does not know in advance how the interested parties might analyze the data. In this case, some basic “information nuggets” should be retained in the published data, but the nuggets cannot replace the data.

Truthfulness at the record level. In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Consider the example of patient records.

The pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug [Emam 2006] [31].

If a published record does not correspond to an existing patient in real life, it is difficult to deploy data mining results in the real world. Randomized and synthetic data do not meet this requirement. Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the patient represented.

Privacy threats

When publishing data, there are three types of privacy disclosure threats. The first type is membership disclosure. When the data set to be published is selected from a large population and the selection criteria are sensitive (e.g., only diabetes patients are selected), one needs to prevent adversaries from learning whether one’s record is included in the published data set.

The second type is identity disclosure, which occurs when an individual is linked to a particular record in the released table. In some situations, one wants to protect against identity disclosure when the adversary is uncertain of membership. In this case, protection against membership disclosure helps protect against identity disclosure. In other situations, some adversary may already know that an individual's record is in the published data set, in which case, membership disclosure protection either does not apply or is insufficient.

The third type is attribute disclosure, which occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the attributes of an individual more accurately than it would be possible before the release.

Similar to the case of identity disclosure, considering adversaries who already know the membership information. Identity disclosure leads to attribute disclosure. Once there is identity disclosure, an individual is reidentified and the corresponding sensitive value is revealed. Attribute disclosure can occur with or without identity disclosure, e.g., when the sensitive values of all matching tuples are the same.

In order to overcome the privacy threats l-diversity based slicing approaches are used and certain values are replaced with less specific but semantically consistent values. The problem of discovering optimal datasets using slicing approach has been proved to be NP-hard [12, 13]. The hybrid optimization algorithm is proposed for this purpose.

Methodology

L-diversity: vertices are partitioned into equivalence groups in all the vertices groups in public datasets. It yields privacy although the data publisher has no knowledge possessed by the adversary [20]. It insists all records that share the similar values of quasi identifiers to have l-diverse values for their sensitive attributes. It's too prone to adversary attacks but it ensures a low breach probability. Anatomy is the other l-diversity method. It does not violate the l-diversity property but it confirms that a prompt individual is involved in the data. Slicing approach is used as an anonymization technique. After the slicing approach in data publishing finding the optimal datasets is hard problem. The Evolutionary algorithm is used for this purpose in this research.

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 [32, 33]. Each chromosome is encoded using a sequence of 0s 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, fittest 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 [33].

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.

Ant Colony Optimization

Ant colony optimization is a meta-heuristic technique that uses artificial ants to find solutions to

combinatorial optimization problems. ACO is based on the behaviour of real ants and possesses enhanced abilities such as memory of past actions and knowledge about the distance to other locations. In nature, an individual ant is unable to communicate or effectively hunt for food, but as a group, ants possess the ability to solve complex problems and successfully find and collect food for their colony. Ants communicate using a chemical substance called pheromone.

As an ant travels, it deposits a constant amount of pheromone that other ants can follow. Each ant moves in a somewhat random fashion, but when an ant encounters a pheromone trail, it must decide whether to follow it. If it follows the trail, the ant's own pheromone reinforces the existing trail, and the increase in pheromone increases the probability of the next ant selecting the path.

Therefore, the more ants that travel on a path, the more attractive the path becomes for subsequent ants. Additionally, an ant using a short route to a food source will return to the nest sooner and therefore, mark its path twice, before other ants return. This directly influences the selection probability for the next ant leaving the nest.

This diagram shows the total amount of the enzymes level (call as pheromone). The large value of pheromone path is considered as a shortest path.

The ACO is used for finding the shortest path using the distance value assign to the each node. The host of the ant is considering as the source node and their food is representing as the destination. The current node is act as a ant in routing process for finding the next shortest nodes in the wireless mesh network.

In general the ACO assign two ants such as forward and backward ant. The forward ant is used while searching the food and the backward ant is used when the ant get back to host. But in transferring the information, only the forward ant can be used. There is no use of backward ant in the transferring process, but it can be used for the acknowledgement purpose.

Here the current node is assign as a forward ant during the transformation; it can be also act as backward ant during the acknowledgement. Now consider the general pseudo code for the ACO. If the data's are send from source to destination, then it follow the pseudo code of the forward ant. The steps of the forward ant as follow:

a. Get the next node based on the distance value. Which node have the less distance is consider.

b. Once it find the next node, update the data storage of the router (simply the routing table) and send the data packets to the certain node.

c. If there is no path or link or node are available then keep the record of the data packet as it is, and discard it. Find any other path.

In general the forward node (source to destination) use the stack (LIFO) order to store the data in the routing table. Similarly the backward node (destination to source) use the queue (FIFO) order to store the data in routing table [34].

If the distance of the path is different, then it is very easy to find the shortest distance using this algorithm. If the nodes have same distance then the ACO can't to find the optimal solution, to overcome using the GA for finding the fitness value for each and every node based on the cost value of the node. Even though the ACO can find the shortest distance, but it not be the optimal solution, for this reason only using both ACO and GA for produce the optimal solution.

Hybrid of ACO and GA:

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 Ant colony optimization. Both GA and ACO 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 pheromones in ACO. After initialization, new next generation individuals are created by enhancement, crossover, and mutation operations. The architecture of the proposed hybrid algorithm is given below.

For finding a optimal data using a ACO and to find the fitness, optimal path among these can found by GA. The reasons for using genetic algorithms are:

They are parallel in nature. They explore solution space in multiple directions at once. GA is well suited for solving problems where the solution space is huge and time taken to search exhaustively is very high.

They perform well in problems with complex fitness. If the function is discontinuous, noisy, changes over time or has many local optima, then GA gives better results.

Genetic algorithm has ability to solve problems with no previous knowledge (blind). For this reason hybrid the ACO with GA to find the shortest path is used. Following diagram can easily explain remaining steps in hybrid of ACO and GA.

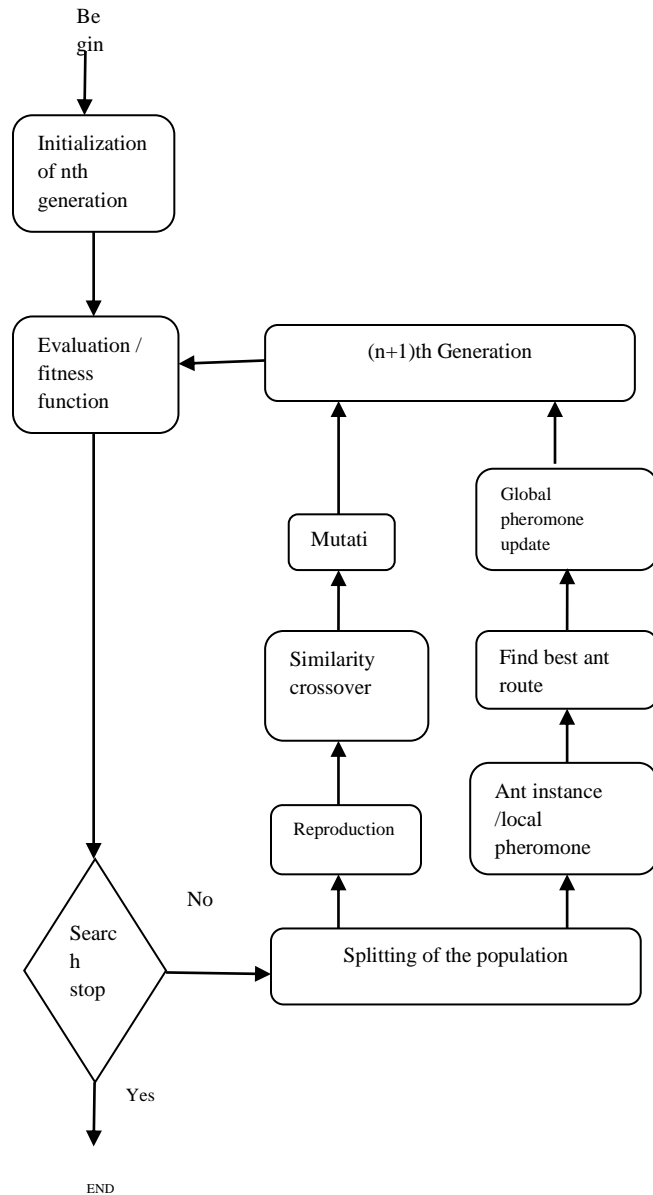


Figure 2 Processes of ACO and GA Hybrid

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 2. The Figure shows an example for generalization of continuous data used in this work.

Dataset

Using the Adult data set from the UC Irvine machine learning repository,1 which is comprised of data collected from the US census. The data set is described in Table 2. Tuples with missing values are eliminated and there are 45,222 valid tuples in total. The adult data set contains 15 attributes in total.

An experiment, obtaining two data sets from the Adult data set is explained. The first data set is the “OCC-7” data set, which includes seven attributes: QI ¼ fAge, Workclass, Education,Marital-Status, Race, Sexg and S ¼ Occupation.

The second data set is the “OCC-15” data set, which includes all 15 attributes and the sensitive attribute is S ¼ Occupation. Note that do not use Salary as the sensitive attribute because Salary has only two values f50K;<50Kg, which means that even 2-diversity is not achievable when the sensitive attribute is Salary. Also note that in membership disclosure protection, do not differentiate between QIs and SA.

In the “OCC-7” data set, the attribute that has the closest correlation with the sensitive attribute Occupation is Gender, with the next closest attribute being Education. In the “OCC- 15” data set, the closest attribute is also Gender but the next closest attribute is Salary.

	Attribute	Type	# of values
1.	Age	Continuous	74
2.	Workclass	Categorical	8
3.	Final-weight	Continuous	NA
4.	Education	Categorical	16
5.	Education Num	Continuous	16
6.	Martial status	Categorical	7
7.	Occupation	Categorical	14
8.	Relationship	Categorical	6
9.	Race	Categorical	5

10	Sex	Categorical	2
11	Capital Gain	Continuous	NA
12	Capital-loss	Continuous	NA
13	Hours per week	Continuous	NA
14	Country	Categorical	41
15	Salary	Categorical	2

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.

It is observed from Figure 3, that the classification accuracy decreases with the increase in k-anonymity level. Figure 4 and 5 show the precision and recall for class label income greater than 50k and less than or equal to 50k respectively.

Conclusion

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 Ant colony Optimization (ACO). Both GA and ACO complement each other to provide global optimization.

In the proposed framework, l-diversity 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.


References

1. Agrawal R., Srikant R. *Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.*
2. 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.*
3. 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.*
4. 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*
5. P. Samarati, "Protecting Respondent's Privacy in Microdata Release," *IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.*
6. L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.*
7. X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," *Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.*
8. D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy-Preserving Data Publishing," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.*
9. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.*
10. A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam, "'-Diversity:

- Privacy Beyond k-Anonymity,” Proc. Int’l Conf. Data Eng. (ICDE), p. 24, 2006.*
11. 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.
 12. P.Samarati, *Protecting respondents’ identities in microdata release*, *IEEE Transactions on Knowledge and Data Engineering* 13 (6) (2001) 1010–1027.
 13. 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.
 14. 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.
 15. Bayardo R. J., Agrawal R.: *Data Privacy through Optimal k-Anonymization*. *Proceedings of the ICDE Conference*, pp. 217–228, 2005.
 16. Dalenius, T.: *Finding a needle in a haystack - or identifying anonymous census record*. *Journal of Official Statistics*, 1986.
 17. Sweeney, L.: *Achieving k-anonymity privacy protection using generalization and suppression*. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, 2002.
 18. Hundepool, A., Willenborg, L.: *μ - and τ -argus: Software for statistical disclosure control*. In: *Third International Seminar on Statistical Confidentiality*, Bled, 1996.
 19. Agrawal, R., Evfimievski, A., Srikant, R.: *Information sharing across privatedatabases*. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, San Diego, California, 2003.
 20. Yao, A.C.: *Protocols for secure computations*. In: *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, 1982.
 21. Liang, G., Chawathe, S.S.: *Privacy-preserving interdatabase operations*. In: *Proceedings of the 2nd Symposium on Intelligence and Security Informatics*, 2004.
 22. S.Ram Prasad Reddy, KVSVN Raju, V.Valli Kumari “*A Dynamic Programming Approach for Privacy Preserving Collaborative Data Publishing*” *International Journal of Computer Applications* (0975 – 8887) Volume 22–No.4, May 2011.
 23. V. Valli Kumari, S.Srinivasa Rao, KVSVN Raju, KV Ramana and BVS Avadhani, *Fuzzy based approach for privacy preserving publication of data*, *IJCSNS International Journal of Computer Science and Network Security*, VOL.8 No.1, January 2008, pp:115-121.
 24. N. Shang, F. Paci, M. Nabeel, and E. Bertino. *A privacy-preserving approach to policy-based content dissemination*. *Technical Report 2009-14*, Purdue University Center for Education and Research in Information Assurance and Security (CERIAS), 2009.
 25. GEHRKE, J. 2006. *Models and methods for privacy-preserving data publishing and analysis*. *Tutorial at the 12th ACM SIGKDD*.
 26. YANG, Z., ZHONG, S., AND WRIGHT, R. N. 2005. *Anonymity-preserving data collection*. In *Proceedings of the 11th ACM SIGKDD Conference*. ACM, New York, 334–343.
 27. CHAUM, D. 1981. *Untraceable electronic mail, return addresses, and digital pseudonyms*. *Comm. ACM* 24, 2, 84–88.
 28. CARLISLE, D. M., RODRIAN, M. L., AND DIAMOND, C. L. 2007. *California inpatient data reporting manual, medical information reporting for California (5th Ed)*, Tech. rep., Office of Statewide Health Planning and Development.
 29. EMAM, K. E. 2006. *Data anonymization practices in clinical research: A descriptive*

- study. Tech. rep. Access to Information and Privacy Division of Health in Canada.
30. Machanavajjhala A., Gehrke J., Kifer D. , "l-diversity: privacy beyond k-anonymity". *Proceedings of the 22nd IEEE Intl. Conf. on Data Engineering, 2006*
 31. L. David, *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold. 1991.
 32. D.E. Goldberg, *Genetic Algorithms: in Search, Optimization, and Machine Learning*. New York: Addison-Wesley Publishing Co. Inc. 1989.
 33. ACO algorithm: Introduction and beyond, Anirundhu Shekwat pratik poddar, Dinesh Boswal, IIT Bombay, AI seminar 2009.
 34. K Vijayalakshmi and S Radhakrishnan, "Dynamic Routing to Multiple Destinations in IP Networkin using Hybrid Genetic Algorithm (DRHGA), *International Journal of Information Technology, Vol 4, No 1, PP 45-52.*

Author Bibliography

	<p>S.Chidambaranathan, He received his post graduate degree in Mathematics from Madurai Kamaraj University, Madurai. He also earned post graduate degree in Computer Application and M.Phil in computer Science from Manonmaniam Sundaranar University, Tirunelveli. Presently he is working as HoD in the Department of MCA, St. Xavier's College (Autonomous), Palayamkottai, Tamil Nadu. He is an author for many books including "PHP for beginners", "XML - An Practical approach" and "Everything HTML". He has published many research papers in National, International journals and conference proceedings. Email: scharan2009@rediffmail.com</p>
---	---