Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GENETIC OPTIMIZATION FOR PRIVACY PRESERVING IN DATAMINING

[1]J. PARANTHAMAN, [2] Dr. T ARULDOSS ALBERT VICTOIRE
[1]University College of Engineering, Department of Computer Science and Engineering, Rajamadam, INDIA
[2] Anna University, Department of Electrical and Electronics Engineering, Coimbatore, INDIA
E-mail: [1] paran_2013@rediffmail.com, [2] t.aruldoss@gmail.com

ABSTRACT

Data mining needs accurate input for meaningful results, but privacy issues could influence users into providing fictitious information. To preserve client privacy in data mining various Anonymization techniques is used, one of the most common being k-anonymity. This converts data into an equivalence classes set with each class having a set of K- records indistinguishable from others. In this paper, k-anonymity is used for preservation of privacy when data mining algorithms are applied. A mushroom data set anonymized to varied levels to preserve privacy and genetic algorithm (GA) for optimization is used in evaluation. Experiments prove that the new method achieves good results.

Keywords: *Privacy Preserving Data Mining, K-Anonymity, Genetic Algorithm*

1. INTRODUCTION

Data mining was developed to provide tools to automatically/ intelligently transform large data knowledge relevant to users. The extracted knowledge, expressed as association rules, decision trees or clusters, permits locating patterns/regularities buried in data but meant to facilitate decision making. This knowledge discovery process returns sensitive information about individuals, compromising their right to privacy. Data mining techniques also reveal critical information about business, compromising free competition, and so disclosures of confidential/personal information should be prevented in addition to knowledge considered sensitive in a given context.

Hence, research was devoted to addressing privacy preservation in data mining resulting in many data mining techniques which included privacy protection mechanisms based on various approaches. Various sanitization techniques were proposed to hide sensitive items/patterns based on removing reserved information/inserting noise in data. Privacy preserving classification methods prevent miners from constructing classifiers capable of predicting sensitive data. Also, recently proposed privacy preserving clustering techniques distort sensitive numerical attributes but preserve general features for cluster analysis [1].

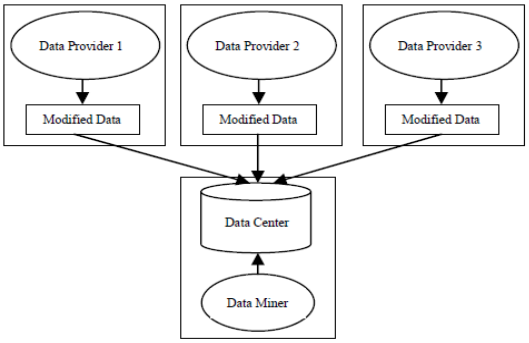

Figure 1: PPDM Based On Data Publishing Scenario

Data mining needs correct input for meaningful results, but privacy concerns influence users to provide wrong information. To preserve client privacy in data mining procedures, various random perturbation of data records based techniques were proposed. Randomization/Distortion are two methods that preserve privacy. Randomization modifies transactions through replacing some items with non-existing items and also through the addition of fake items to ensure privacy preservation. Distortion operates on a transaction database through probabilistically changing items in every transaction [2].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1.1 Models of PPDM

1.1.1 Trust Third Party Model

The security standard assumes we have a trusted third party to which all data is given. The third party performs computation and delivers results and except for this party, nobody learns anything inferable from own input/ results. Secure protocols aim to reach this privacy preservation level without finding a third party everyone trusts.

1.1.2. Semi-honest Model

In this, all parties follow protocol rules using correct input, but when the protocol is free it uses anything it sees during protocol execution to compromise security.

1.1.3. Malicious Model

In malicious model, participants have no restrictions. Any party is free to indulge in any action. Usually, it is difficult to develop efficient protocols valid under a malicious model.

1.1.4. Other Models - Incentive Compatibility

Though semi-honest and malicious models are well researched, other models outside purview of cryptography are also possible. An example is incentive compatibility. A protocol is incentive compatible when a cheating party is either caught/suffers an economic loss. Under the rational economics model, this ensures that parties have no advantage by cheating. Of course, this fails in an irrational model. [3].

1.2 K-anonymity

K-anonymity focuses on two techniques specifically: *generalization* and *suppression*, which, unlike existing techniques like scrambling/swapping, preserve information truthfulness.

Generalization substitutes a given attribute's values with general values. For this, the idea of domain captures the generalization process through the assumption of existence of a *generalized domains set*. The original domain set with generalizations is called Dom. Every generalized domain has generalized values and mapping between each domain and its generalizations exists.

Mapping is stated by a *generalization relationship* $\leq_D$. Given two domains *Di* and *Dj* $\in$Dom, *Di* $\cdot\leq$ *Dj* states values in domain *Dj* $\leq D$are generalizations of values in *Di*. The generalization relationship $\leq_D$ defines a partial order on set Dom of domains, requiring satisfaction of the following conditions equations (1) and (2):

$$c1: \forall D_{i,} D_{j,} D_z \in D_{om} \quad (1)$$

$$D_i \leq_D D_{j,} D_i \leq_D D_z \Rightarrow D_j \leq_D D_{z,} D_z \leq_D D_j \quad (2)$$

C2: all maximal elements of Dom are singleton. Condition C1 states that for each domain *Di*, the 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 every domain is always generalized to single value. The generalization relationship definition implies existence for each domain *D* $\in$Dom, a totally ordered hierarchy, called *domain generalization hierarchy*, denoted DGH*D* [4].

2. LITERATURE REVIEW

Matatov, et al., [6] suggested data mining privacy through decomposition (DMPD) algorithm which used genetic algorithm to locate optimal feature set partitioning. DPMD evaluated ten separate datasets to compare classification performance with other k-anonymity-based methods. DMPD performs better than existing k-anonymity-based algorithms according to results. There was also no need to apply domain dependent knowledge. Using multi objective optimization methods, author examined trade-off between 2 conflicting PPDM objectives: privacy/predictive performance.

A protocol for 2 parties each with a private data partition to apply genetic algorithms securely to discover a decision rules set for private data partitions without compromising individual data privacy was proposed by Han and Ng [7]. As GA is iterative, it is challenging to preserves data privacy at every iteration and also to ensure that intermediate results at each iteration did not compromise participating party's data privacy. The proposed protocol satisfied both data privacy requirements.

Meints and Möllera et al [8] briefly overviewed state-of-the-art in PPDM and some current suggestions to proceed towards PPDM standardization. They were summarized by considering how PPDM can improve based on the European Directive 95/46/EC, taking into account procedural/process-related considerations. Scoring practice in financial sector is an example to illustrate such considerations. Though this does not

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

demonstrate aspects relevant to data mining, it was analyzed from a data protection developments perspective. In addition to process chains having basic data providers, service providers having to calculate scoring values and banks using mining results, the paper analyses requirements which data controllers have to meet

Han and Ng presented a Privacy-Preserving GA for Rule Discovery [9]. The whole data set was partitioned between two parties with GA finding the best rules set without publishing private data. Two parties developed fitness function jointly to evaluate results using each party's private data without compromising data privacy by Secure Fitness Evaluation Protocol. To meet privacy related challenges, GA generated results did not compromise privacy of both parties with partitioned data. Creation of initial population and ranking individuals for reproduction was undertaken jointly by both parties.

Combined Simulated Annealing and Genetic Algorithm to Solve Optimization Problems were presented by Elhaddad et al [10]. Various evolutionary algorithms were used to optimize results. To improve methods and to ensure quality results in less time, hybrid techniques were used. GA and Simulated Annealing (SA) combined to solve optimization issues. Both searched a solution space in iterative manner till convergence. Both algorithms were different. GA's mechanism was parallel on solutions set exchanging information using crossover operation. SA works on one solution at a time. SA and GA combined to minimize both algorithms disadvantages.

The issue of security violations when malicious parties provide false data was studied by Han and Ng [11]. The author identified secure scalar productprotocols,4 privacy vulnerabilities in many PPDM algorithms, proposing a general model of 2-party interaction. Its applicability to securely compute (x1+y1)(x2 + y2) and (x + y) log2(x + y) where xi and yi are private values held by each party respectively was demonstrated and it showed how the model could securely compute 4 commonly used kernel functions and other functions. The author also proposed 2 necessary conditions and 2 basic measures for adoption in the current malicious model.

An efficient algorithm to mine privacy preserving high utility item sets by considering sensitive item sets was presented by Saravana bhavan and Parvathi et al [12]. The algorithm which has 3 steps to attain the research aim includes, 1) Data sanitization, 2) Construction of sensitive utility FP-tree and, 3) Mining of sensitive utility item sets. Experiments were carried out with real and synthetic dataset with its performance being evaluated with evaluation metrics like Miss cost and Database difference ratio.

3. METHODOLOGY

The proposed work uses k-anonymity's granularity reduction technique for privacy preserving during data mining. GA optimizes feature selection.

3.1 Genetic Algorithms

GAs is a family of evolution inspired computational models. They encode a potential solution for a specific problem on chromosome-like data structure applying recombination operators to these structures to preserve critical information. GAs are viewed as function optimizers, though problem range to which GA is applicable is broad [13].

3.2 GA Operators

A simple GA involves 3 operators: selection, crossover (single point), and mutation.

3.2.1 ***Selection:*** This operator chooses chromosomes for reproduction in a population. The fitter the chromosome, the more times it will be chosen to reproduce in equation (3) [14]

$$P_s(i) = f(i) / \sum_{j=1}^{n} f(j) \qquad (3)$$

where $ps(i)$ *and* $f(i)$ are selection and fitness value probabilities for ith chromosome respectively. Roulette wheel selection is implemented as follows:

1 Evaluate fitness, *fi* of each individual in a population

2 Compute probability (slot size), *pi*, of selecting each population member as in equation (4):

$$P_i = f_i / \sum_{j=1}^{n} f_j \qquad (4)$$

where *n* is population size.

3 Calculate cumulative probability, *qi*, for each individual as in equation (5):

$$qi = \sum_{j=1}^{i} p_j \qquad (5)$$

4 Generate uniform random number, $r \in (0, 1]$.

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5 If r <$q1$ then select first chromosome, $x1$, else select individual xi such that

$$qi-1< r \leq qi$$

6 Repeat steps 4–5 n times to create n candidates in mating pool [15].

3.2.2 Crossover operates individually. A *crossover point* is randomly chosen for 2 randomly chosen individuals (parents). The point is between 2 bits dividing each individual into left and right sections. Crossover swaps left (or right) section of both individuals. A crossover example: consider two parents:

Parent 1: 1010101010
Parent 2: **1000010000**
If crossover point randomly occurs after fifth bit. Then every new child receives one half of the parent's bits:

Child 1: 10101**10000**
Child 2: **10000**01010 [16]

3.2.3 Mutations are global searches. A mutation probability is predetermined before starting the algorithm and applied to every individual bit of each offspring chromosome for determining if it is to be inverted [17].

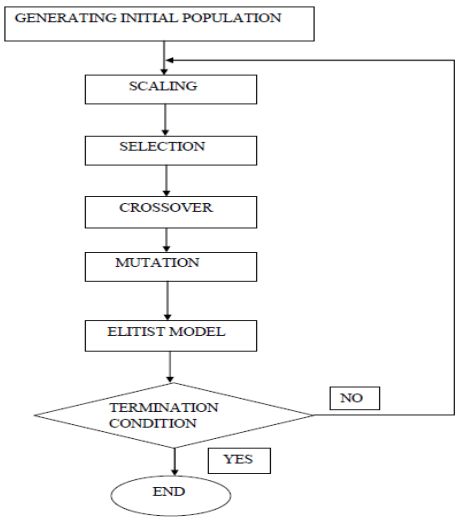

Figure 2 *Flowchart Of Genetic Algorithm*

4. RESULTS AND DISCUSSION

Mushroom data set is used, and proposed algorithm tested. Results reveal the algorithm being capable of finding an optimal/near optimal solution for varying k-anonymity model levels. Performance metrics used include average accuracy, precision and recall. Figures 3 to 5 depict this experiment's results.

Table1 Shows the Results Obtained by the Genetic Algorithm Method

Anonymization	Classification accuracy
No Anonymization	0.958271787
k=5	0.954455933
k=10	0.947193501
k=20	0.937715411
k=25	0.93525357
k=30	0.929837518
k=35	0.921221073
k=40	0.916789759
k=45	0.882912899
k=50	0.91211226

Figure 3 defines the plot between classification accuracy to anonymization.

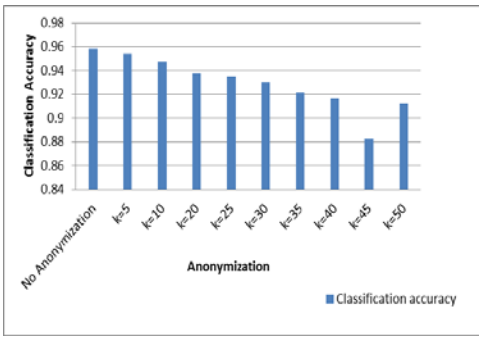

Figure 3: Classification Accuracy

It is seen from Fig 3 that classification accuracy declines with increase in k-levels. Between no anonymization and k==30 it decreases by 2.96% and between k=5 and k=50 it decreases by 4.43%.

Table 2: Precision and Recall Achieved

Anonymization	Precision	Recall
No Anonymization	0.961365139	0.956999557
k=5	0.957571805	0.953156209
k=10	0.950324439	0.945861951

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

k=20	0.94034287	0.936437867
k=25	0.937621654	0.934028091
k=30	0.930809935	0.928840408
k=35	0.92196315	0.920388942
k=40	0.917708964	0.915892426
k=45	0.886578431	0.882385574
k=50	0.912653555	0.911294989

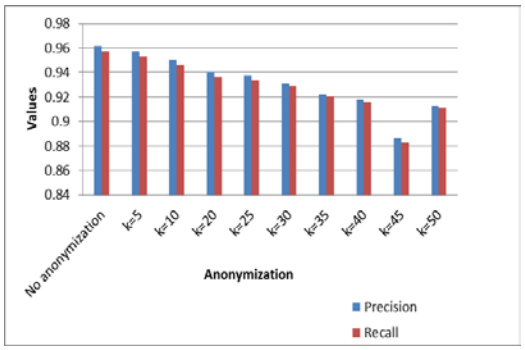

Figure 4: Average recall

It is seen from Fig 4 that average precision declines with increase in k-levels. Between no anonymization and k==30 it decreases by 3.17% and between k=5 and k=50 it decreases by 4.69%.It is also observed that average recall declines with increase in k-levels. Between no anonymization and k==30 it decreases by 2.94% and between k=5 and k=50 it decreases by 4.39%.

5. CONCLUSION

This work suggested a GA for data mining privacy preservation. K-anonymity method with differing k-levels was used. When mining large data set, evolutionary algorithms like GA find optimal data sets. Mushroom data sets evaluated the experiment and performance parameters like accuracy, precision, and recall, and were represented graphically with differing k-levels for granularity reduction. Experiments demonstrate that increase in k-anonymity levels ensures a decrease in classifier performance within acceptable levels.

REFRENCES:

[1]. Bertino, E., Fovino, I. N., & Provenza, L. P.. A framework for evaluating privacy preserving data mining algorithms*. *Data Mining and Knowledge Discovery*, *11*(2), 2005, 121-154.

[2]. Shrivastava, R., Awasthy, R., & Solanki, B.. New Improved Algorithm for Mining Privacy Preserving Frequent Itemsets. *International Journal of Computer Science & Informatics*, *1*. 2011

[3]. Ge, X., Yan, L., Zhu, J., & Shi, W. Privacy-preserving distributed association rule mining based on the secret sharing technique. In *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on* 2010, June, pp. 345-350. IEEE.

[4]. Ciriani, V., di Vimercati, S. D. C., Foresti, S., &Samarati, P.. κ-anonymity. In *Secure Data Management in Decentralized Systems* 2007, pp. 323-353. Springer US.

[5]. Yang, C., Alomair, B., & Poovendran, R. Optimized Relay-Route Assignment for Anonymity in Wireless Networks..

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

[7]. Han, S., & Ng, W. K. Privacy-preserving genetic algorithms for rule discovery. In *Data Warehousing and Knowledge Discovery*, 2007. pp. 407-417. Springer Berlin Heidelberg.

[8]. Meints, M., & Möller, J. Privacy Preserving Data Mining.

[9]. Han, S., & Ng, W. K. Privacy-preserving genetic algorithms for rule discovery. In *Data Warehousing and Knowledge Discovery* 2007, pp. 407-417. Springer Berlin Heidelberg.

[10]. Elhaddad, Y. R. Combined Simulated Annealing and Genetic Algorithm to Solve Optimization Problems. *World Academy of Science, Engineering and Technology* 2012.

[11]. Han, S., & Ng, W. K. Preemptive measures against malicious party in privacy-preserving data mining. In *SIAM International Conference on Data Mining* 2008. pp. 375-386.

[12]. C.Saravanabhavan And R.M.S.Parvathi,Privacy Preserving Sensitive Utility Pattern Mining Journal of Theoretical and Applied Information Technology 20th March 2013. Vol. 49 No.2, 2013

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[13]. Mathew, T. V. (1996). Genetic Algorithm. *Indian Institute of Technology Bombay, Mumbai*.

[14]. Melanie, M.. An introduction to genetic algorithms. *Cambridge, Massachusetts London, England, Fifth printing*, *3*. 1999.

[15]. Sastry, K., Goldberg, D., & Kendall, G.. Genetic algorithms. In *Search Methodologies* 2005, pp. 97-125. Springer US.

[16]. Spears, W. M., &Anand, V. *A study of crossover operators in genetic programming* 1991. pp. 409-418. Springer Berlin Heidelberg.

[17]. Mishra, B., & Patnaik, R. K.. *Genetic Algorithm and its Variants: Theory and Applications* (Doctoral dissertation) 2009.