ISSN: 0976-3104
SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS)
Mohana et al.

A COMPARITIVE FRAMEWORK FOR FEATURE SELCTION IN PRIVACY PRESERVING DATA MINING TECHNIQUES USING PSO AND K-ANONUMIZATION

S Mohana*, SA Sahaaya, Arul Mary
M.I.E.T Engineering College, Jayaram College of Engineering &Tech, Tamil Nadu, INDIA

ABSTRACT

The trend of technological era leads to accumulate and utilization of enormous quantity of private details of individuals using internet, which eventually lead to disclose their personal identities. Privacy preserving of data must uphold from revealing sensitive data during the disclosure of the individual's data. Privacy preserving should be incorporated as mining of these datums and the domain deals with this known as Privacy Preserving Data Mining. In the proposed framework, an attribute suppression technique is employed using Particle swarm optimization algorithm and a generalization technique for anonymization is proposed. Also the same work is done using k anonymization and the results are compared for classification accuracy, Precision and recall. In the proposed system Genetic Algorithm and Particle Swarm Optimization takes the common population for evaluation and the results are compared. An optimal generalized feature set is acquired by the PSO and k anonymization technique and is which is used for classification task. The end results of classification are compared with average classification accuracy, average precision and average recall.

Published on: 2nd -December-2016

KEY WORDS

Privacy Preserving Data Mining (PPDM), K-Anonymity Generalization, Particle swarm optimization (PSO), Classification accuracy)

***Corresponding author: Email:** mohana.p3@gmail.com, samjessi@gmail.com

INTRODUCTION

Alzheimer's Data mining process is digging out useful information from the huge capacity of data. Data mining tasks have been classified as association rule mining, clustering, classification and prediction. Gathering data for mining, may leads to the collection of the private data which identifies personal details of individuals must be confined without affecting the data mining process. The main objective of PPDM is to aggregate the information available in the data by not leaking the individual information of the participants. There are two main privacy models. Privacy-preserving data mining (PPDM) algorithms are built in such a way that the private data on which data mining is employed should not publicize the client private information. PPDM is broadly classified in to non-interactive model and interactive model. In the non-interactive model, the database is sanitized and revealed to public. The interactive model deals with accessing the desired data by asking multiple questions to the database and getting the answers from the database. Many methods for privacy preserving have been proposed so far.Some of the methods to preserve privacy are, preserving privacy while publishing the data, changing the results of data mining for preserving privacy and changing or restricting the results of a query to preserve the privacy either online or offline. Perturbation, Randomization, k-anonymity, t-d closeness and l-diversity are some of the methods used for privacy preserving while publishing the data [1].

Anonymization of data converts a dataset in to a form that maintains privacy using k-anonymity, so that individually identifiable information is covered. K Anonymization converts data to equivalence classes where of the classes has a set of K- fields indistinguishable from one another.

Generalization technique is used in k-anonymity, where some values are replaced with less specific but retaining the meaning of the value and some of the values are suppressed. For example, country can be generalized to state, age is generalized to age-range such as 30-35 yrs or young, middle, old etc which leads to less identification of individual's data. Data loss can be minimized through optimization of an aggregated value in all

features and records.

Particle Swarm Optimization (PSO) is population based heuristic search technique, which is usually employed for solving the NP-hard problems. In Particle Swarm Optimization, a particle which can be idle because of the stagnant way and also suboptimal solutions were got because of early congregation. During optimization problem, reducing or maximizing an objective function is the difficulty encountered. Global optimization is the best set of permissible environment achievable for an objective in given constraints. Many fusions of algorithms integrating GA are proposed to overcome the limitations of PSO.

In this paper, classification accuracy is compared to investigate the effect of anonymization for the Adult dataset with respect to k anonymization and PSO. PSO is used to select the features of the data set. Performance of classifier is analyzed with PSO selection and K Anonymization for different levels. Sections 2 presents the Literature review of previous work; Section 3 describes the how PSO is used to select to select features from the anonymous data. Section 4 reports on evaluation of various aspects of the proposed work and concludes the paper.

RELATED WORK

There are two different approaches for privacy preserving in data mining suggested by Wu [2]. petrubaration is used as a first approach in which perturbing the data using a random process is employed. Cryptographic methods for multi party computing is used as Second approach used. After preserving the privacy association rule mining was applied to the data. The data flow in uni-direction in the processor boards was taken and the author compared the impact of privacy preserving in rule mining. The results proved that cryptographic methods were better than data perturbation method.

PPDM needs accurate models for data aggregation without accessing the precise information in the data record of individuals. Perturbation-based PPDM approach was widely used for preserving data before publishing the data. This approach was limited because it used trust only on data miners. Therefore, Li et al [3] suggested Multilevel Trust (MLT-PPDM). The trustful data miner was one, which access less the perturbed copy of data. Malicious miner could access the data more times in various means and jointly used them to infer the original data which was not published by the owner of the data. This attack was reduced by correlating the copies of data among different miners and trust level was created based on the correlation level. This trust level was used when miner requested data. Since many users are unwilling to share the personal data, many users give incorrect information. This may affect the results of data mining because of not having sufficient amount of correct information [4]. The dimensionality of the information is also large and selection of good privacy preserving is needed for the success of data mining.

Kadampur et al [5] proposed a strategy to protect the privacy of data during decision tree construction of data mining process. With the numeric attributes of data a specific noise is added and then given to second party to construct the decision tree by CART algorithm. The best split point was fond based on the info gain measures. The decision obtained from the original tree was compared with the decision constructed by the second party using the obfuscated data. The comparison proved that both the trees were similar. Therefore this method preserved the privacy.

Lindell et al [6] analyzed how multiparty secure communication is applied to privacy preserving in data mining. Authors implemented two different approaches for privacy preserving. In the first approach, original data was divided into many partitions and distributed to multiple parties. While performing mining, these partitions were united by not allowing each party to see the individual data stored in other parties. In this multiparty secure communication the goals were set by using the properties such as privacy, correctness, fairness, independence of inputs and guaranteed output delivery. In the second method, from the original data only statistical information is calculated and released for data mining. The mining results were compared for these two approaches. Results proved that multiparty secure communication was better than statistical information for data mining.

Sumana and Hareesh [7] proposed a k-anonymity model by generalization and suppression to protect the ' identities of individuals while releasing truthful information. This k-anonymity model protected against the identity disclosure, but it could not protect against the individual attribute disclosure. The ℓ-diversity method solved this problem by using equivalence classes and each class had at least ℓ well-defined values for each sensitive attribute. The author proposed a complete (α, k) model by using the distance between two distributions

with a threshold α. Results proved that complete model preserved the privacy is better than simple k-anonymity model.

Sharing the patient data is often needed for the purpose of research. But the identity of individuals must be protected. Most commonly used methods to hide the personal details are k-anonymity and l-diversity models. Tamas et al [8] analyzed these two models with the details of cancer patients which were published to health professionals. After implementing these models, discernibility was used to compare the performance. Result proved that l-diversity was better for single sensitive attribute and k-anonymity model was better for multiple sensitive attribute.

Most of the privacy preserving algorithms are based on various privacy and utility assumption. Bingchun et al [9] proposed creation of decision tree from the anonymated data directly. This method avoided the data preparation by the ID3 algorithm. Experimental results showed that proposed decision tree from the k-anonymated data performed efficiently for classification problems.

In k-anonymity model, generalization technique is used to swap a sensitive value with a less specific value and maintaining semantically consistent value, and suppression was used to hide a value at all. Generalization was commonly used, because suppression may lessen the quality of the data mining results if not used efficiently. But in generalization every quasi-identifier needs to consider the hierarchy of the domain. Therefore, Kisilevich et al [10] proposed multidimensional suppression for generating classification trees. Multidimensional suppression was used based on the attribute values without using the domain hierarchy trees. Experiments were conducted with 10 different data sets and the results proved that classification accuracy was improved up to 5.3% than manual classification and classification tree from generalized data.

Mandapati et al [11] proposed a Hybrid Evolutionary Algorithm using Particle Swarm Optimization (PSO) Genetic Algorithm (GA) and for PPDM. While preserving privacy, the entire existing EA algorithm produced solutions which were restricted to specific problems like the cost function evaluation. Authors proposed both the GA and PSO with the same population and, k-anonymity was used to generalize the actual dataset. The hybrid optimization found the optimal generalized feature set and the improved the success of mining.

Slava Kisilevich et al [12] proposed a method to achieve k-Anonymity of Classification Trees by Using Suppression (kACTUS) where kACTUS performs an efficient multi-dimensional suppression,in which , suppression is done only on certain records based on other attribute values, without manually-producing domain hierarchy trees. Results proved that kACTUS' predictive performance was good than the k-anonymity. Also, average the accuracies of TDS, TDR and kADET are lower than kACTUS in 3.5%, 3.3% and 1.9% correspondingly regardless of usage of manually defined domain trees.

Goryczka et al [13] proposed m-privacy in the anonymity model which preserves the privacy constraint against in any group of m number of colluding providers of data. Heuristic algorithms were used for data aware anonymization which provides the m-privacy efficiently. Experiments were conducted on the real data sets and the efficiency of the proposed algorithm was compared with base line algorithms which provided m-privacy.

MATERIALS AND METHOD

[Table- 1] The Adult dataset from UCI machine learning Repository is used for assessment. There are 48,842 rows, containing both categorical and integer attributes derived from Census information from the year 1994. There are about 32,000 rows with 4 numerical columns, and the column contains age {17 – 90}, fnlwgt {10000 – 1500000}, hrsweek {1 – 100} and edunum {1 – 16}. k- anonymization is employed in age column and native country . Original attributes of the Adult dataset is shown in **Table- 1**.

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

THE IIOAB JOURNAL
ISSN: 0976-3104

49	Jamaica	<=50K
52	United-States	>50K
31	United-States	>50K
42	United-States	>50K

K-ANONYMIZATION

[Figure- 1] In k-anonymity, the data is changed to equivalence classes , in which each class consist of a set of k- records that diverges from K others. suppression &Generalization techniques are employed to lessen the minute sign of the pseudo-identifiers. The features are generalized to a series so as to lessen the microscopic view, for example, street is generalized as city and it prevents the disclosure of individual's information. Suppression is used to remove the value of the attribute in order to reduce the identification risk with the records available publically and the example is shown. Because of its easiness the k-anonymity is a popularly used technique and also many techniques are existing to practice anonymization [14].

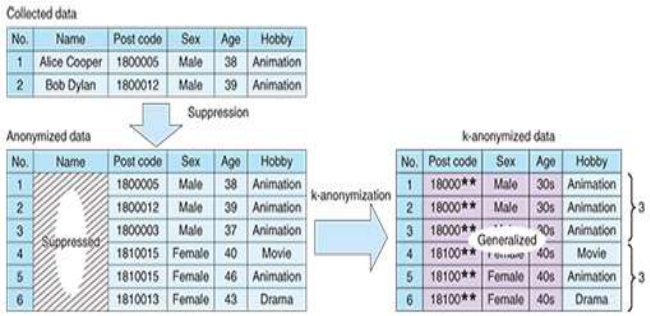

Fig: 1. 3-Anonymous Table

THE PARTICLE SWARM OPTIMIZATION

[Figure- 2] It is a well-built optimization technique based on the behavior and attitude of swarms. It shares the idea of group communication to solving of problem. It makes use of a amount of particles that characterize a swarm going around in the search space affording the best result. every particle is consider as a point in a K-dimensional space which alters its "flying" having its own flying knowledge as well as the flying knowledge of supplementary particles. [11] best balue can be obtained by keeping track of its coordinates in the result space which are associated with the finest solutionwhich is known as personal best , ***pbest***. ***hbest is the value*** obtained so far by any particle in the neighborhood of that particle.

The notion of PSO falls in moving each particle in the direction of its pbest and the hbest position.

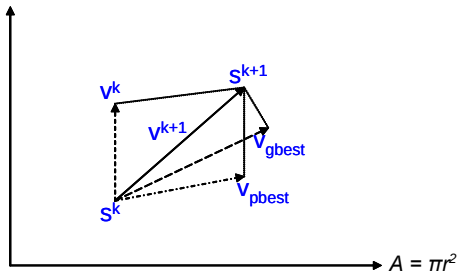

Fig: 2. modifying the searching point

S^k: current searching point
S^{K+1}: modified searching point
V^k: current velocity
V^{k+1}: modified velocity
V_{pbest}: velocity based on pbest
V_{gbest} Velocity based on hbest

Every particle aims to adjust its place using information from the existing positions, the existing velocities, the distance involving the present position and pbest,the distance involving the present position and hbest. The modification of the particle's position can be mathematically modeled according the following equation :

$$V_{IK+1} = wV_{ik} + c_1 rand1(\ldots..) \times (p_{besti} - S_{ik}) + C_2 + rand2(\ldots.) \times (g_{best} - S_i^k) \text{-------- (1)}$$

Where

V_{ik}:velocity of agent i at iteration k,
W:weighting function,

THE IIOAB JOURNAL
ISSN: 0976-3104

C^i:weighting factor,

$rand$:uniformly distributed random number between 0 and 1,

S_i^k:current position of agent I at iteration k,

p_{besti}: p_{besti} t of agent i,

g_{best}t: g_{best} of the group

The following weighting function is usually utilized in (1)

$$w = w_{max} - \frac{[(W_{max} - W_{min}) \times iter]}{maxIter} \quad \text{..... (2)}$$

where w_{max}= initial weight,

W_{min}= final weight,

$maxIter$ = maximum iteration number,

$iter$ = current iteration number.

$$S_i^{k+1} = S_i^k + V_i^{k+1} \quad \text{---------- (3)}$$

In the present paper, the 'Adult' dataset available in the UCI machine learning repository is used. Adult Dataset provides the 1994 Census information. The dataset contains 48842 instances, with both categorical and integer attributes. There are about 32,000 rows and 4 numerical columns present in the Adult Data set. The columns and their ranges are: age[17 - 90], fnlwgt [10000 - 1500000], hrsweek[1 - 100] and edunum[1 - 16]. The age column and the native country were aggregated using the principles of K anonymization. Table I and II show the original data and the modified attribute data. Using 10 fold cross validation the original and the K anonymized dataset are classified.

The Naïve Bayes classifier is popular because it is more efficienct. It also has the benefit of having fine classification accuracy and is employed in a number of area. Using Bayes theorem the classifier model is formulated as:

$$p\left(C|F_1,....,F_n\right) = \frac{p(C)\, p\left(F_1,....,F_n|C\right)}{p\left(F_1,....,F_n\right)}$$

Table: 2. The K-Anonymous Dataset

Age	Job	Class
Old	Employed	Good
Young	Employed	Bad
middle age	Employed	Good
middle age	Employed	Good
middle age	Employed	Bad
Young	Employed	Good
middle age	Employed	Good
Young	Employed	Good
Old	Employed	Good
Young	Employed	Bad

Table: 3.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

Anonymization is achieved using attribute suppression technique and generalization using Particle swarm optimization algorithm. Also the same work is done using k anonymization for different levels of K and the results are compared for classification accuracy, Precision and recall using Bayesian classifier. Both K Anonymization and PSO work with the same population in the

THE IIOAB JOURNAL
ISSN: 0976-3104

system proposed and the results are compared. An optimal generalized feature set is acquired by the PSO and k anonymization technique which is used for classification task. The end result of classification are compared with average classification accuracy, average precision and average recall.In this paper it is proposed to compare the classification accuracy of Navie Bayes anonymized dataset for PSO and k anonymization . As the anonymization complexity increases it is observed that the classification accuracy of K-Anonymization outrages the classification accuracy of PSO.

RESULTS

[Figure- 3] The classification accuracy obtained from Naïve Bayes is shown. It is shown that the classification accuracy of k Anonymization outrages the classification accuracy of PSO.

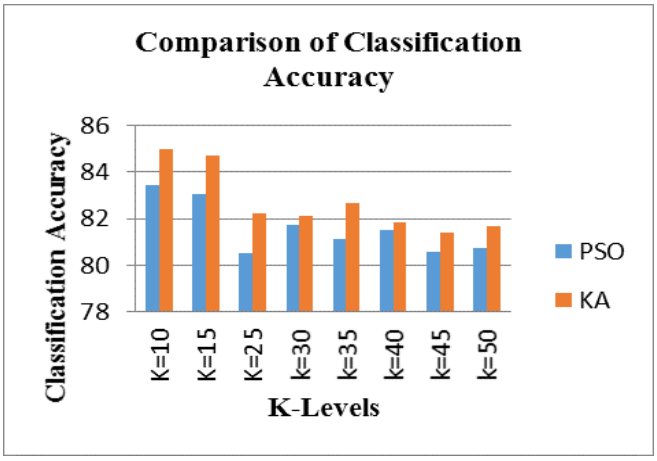

Fig: 3. Classification accuracy of Naïve Bayes for PSO and K Anonymization

[Figure- 4] The Precision value of the classification accuracy is shown in **[Figure- 4]**. It is shown that the precision of k anonymization outrages the Precision value of

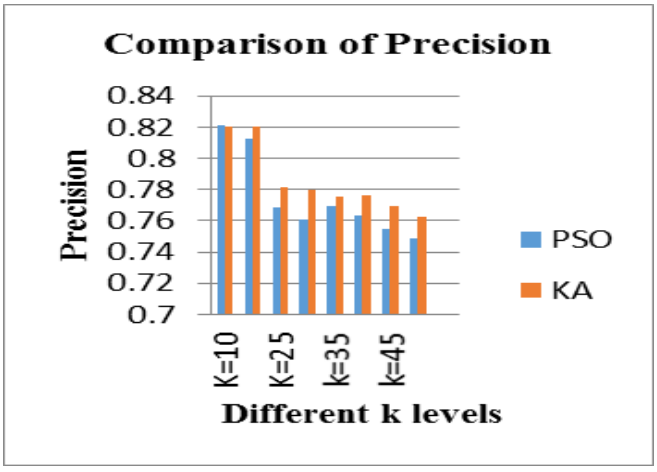

Fig: 4. Comparison of Precision for PSO and K Anonymization

[Figure- 5] The Recall value of the classification accuracy is shown in **[Fgure- 5]**. It is shown that the Recall value of k Anonymization outrages the Recall value of PSO

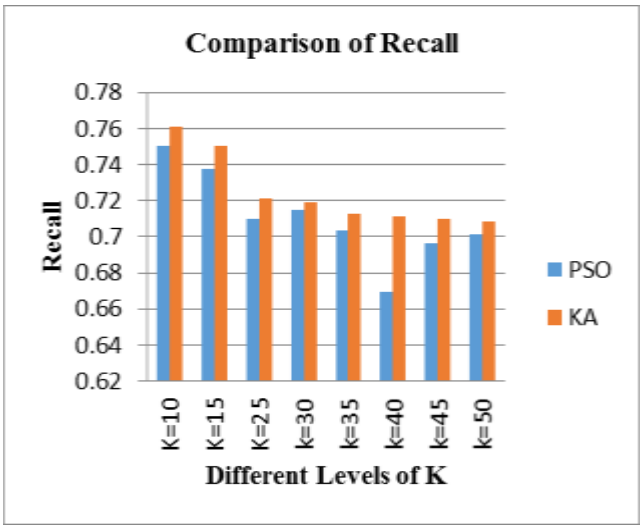

Fig: 5. Comparison of Recall for PSO and K Anonymization

CONCLUSION

In this work, it is proposed to implement feature selection by using PSO and K anonymization technique. To validate the results classification accuracy of Particle Swarm Optimization (PSO) and k anonymization technique is compared. K-anonymization outrages PSO feature selection which is evaluated in terms of Classification accuracy, Precision and Recall. K-anonymity is accomplished by generalization and suppression of the original dataset. For different levels of k-anonymity experiments were performed and the results achieved are evaluated.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

ACKNOWLEDGEMENT

None

FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

REFERENCES

[1] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. [2007]. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1): 3.

[2] Wu CW. [2005] Privacy preserving data mining with unidirectional interaction. In *2005 IEEE International Symposium on Circuits and Systems* (pp. 5521-5524). *IEEE*.

[3] Li Y, Chen M, Li Q, Zhang W. [2012] Enabling multilevel trust in privacy preserving data mining. *IEEE Transactions on Knowledge and Data Engineering*, 24(9):1598-1612.

[4] Wang J, Luo Y, Zhao Y, Le J. [2009] A survey on privacy preserving data mining. In *2009 First International Workshop on Database Technology and Applications* (pp. 111-114). IEEE.

[5] Kadampur MA. [2010] A noise Addition scheme in Decision tree for privacy preserving data mining. *arXiv preprint arXiv:1001.3504*.

[6] Lindell, Y, & Pinkas, B. [2009]. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality* 1(1):5.

[7] Sumana M, Hareesha KS. [2010]. Anonymity: An Assessment and Perspective in Privacy Preserving Data Mining. *International Journal of Computer Applications*, 6(10).

[8] Gal TS, Chen Z, Gangopadhyay A. [2008] A privacy protection model for patient data with multiple sensitive attributes. *IGI Global*, 28-44.

[9] Bingchun L, Guohua, L. [2011] The Classification of k-anonymity Data. In *Computational Intelligence and Security (CIS), 2011 Seventh*

International Conference on (pp. 1374-1378). IEEE.

[10] Kisilevich, S, Rokach, L, Elovici, Y, & Shapira, B. [2010]. Efficient multidimensional suppression for k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, *22*(3): 334-347.

[11] Mandapati S, Bhogapathi RB, Chekka RB. [2013] A Hybrid Algorithm for Privacy Preserving in Data Mining. *International Journal of Intelligent Systems and Applications 5*(8):47.

[12] Kisilevich, S, Elovici, Y, Shapira, B, & Rokach, L. [2009]. kACTUS 2: privacy preserving in classification tasks using k-anonymity. In *Protecting Persons While Protecting the People* (pp. 63-81). Springer Berlin Heidelberg.

[13] Goryczka S, Xiong L, Fung BC. [2014] -Privacy for Collaborative Data Publishing. *Ieee Transactions On Knowledge And Data Engineering 26*(10):2520-2533.

[14] El Emam, K, & Dankar, F. K. [2008]. Protecting privacy using k-anonymity.*Journal of the American Medical Informatics Association 15*(5): 627-637.