View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2007 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-2-2007

Privacy Preserving K-means Clustering with Chaotic Distortion

Jie Li

Yong Xu

Chao Chu

Yunfeng Wang

Follow this and additional works at: https://aisel.aisnet.org/iceb2007

This material is brought to you by the International Conference on Electronic Business (ICEB) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICEB 2007 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Proceedings of The Seventh International Conference on Electronic Business, Taipei, Taiwan, December 2-6, 2007, pp. 61-67.

PRIVACY PRESERVING K-MEANS CLUSTERING WITH CHAOTIC DISTORTION

Jie Li, Hebei University of Technology, China, lijie@hebut.edu.cn
Yong Xu, Hebei University of Technology, China, xuyong@hebut.edu.cn
Chao-Hsien Chu, Pennsylvania State University, USA, chu@ist.psu.edu
Yunfeng Wang, Hebei University of Technology, China, ywang@hebut.edu.cn

ABSTRACT

Randomized data distortion is a popular method used to mask the data for preserving the privacy. But the appropriateness of this method was questioned because of its possibility of disclosing original data. In this paper, the chaos system, with its unique characteristics of sensitivity on initial condition and unpredictability, is advocated to distort the original data with sensitive information for privacy preserving k-means clustering. The chaotic distortion procedure is proposed and three performance metrics specifically for k-means clustering are developed. We use a large scale experiment (with 4 real world data sets and corresponding reproduced 40 data sets) to evaluate its performance. Our study shows that the proposed approach is effective; it not only can protect individual privacy but also maintain original information of cluster centers.

Keywords: Privacy preserving, data mining, k-means clustering, chaotic distortion

INTRODUCTION

Data distortion methods have been popularly used in privacy preserving data mining to mask the original data from revealing. A large proportion of research uses randomized data distortion techniques for preserving the privacy [1]. Also known as data perturbation or data randomization, data distortion protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. Whereas, it is noted that random objects have "predictable" structures in the spectral domain and then offers a random matrix-based spectral filtering technique to recover original data from the data set distorted by adding random values [2]. Aiming at resolving this problem, the chaos system, with its unique characteristics of sensitivity on initial condition and unpredictability, is advocated to distort the original data with sensitive information for privacy preserving k-means clustering (PPKC).

K-means is a simple and widely used clustering technique to group items into k clusters, which results come in two forms: assignment of entities to clusters and the cluster centers themselves [3]. In this scenario, the cluster centers represent the total characteristics of a population, which should be kept for utilization. The fundamental objective of chaotic distortion is to modify the original dataset so that the original data points were changed to preserve individuals' privacy and the cluster centers still keep very closely to the original ones. This paper contributes to the literature in two aspects. First, the chaotic series is advocated for data distortion and the chaotic distortion method is introduced for PPKC. Second, we propose three specific performance metrics for PPKC, which includes an individual privacy preserving measure, a population information loss measure, and a comprehensive privacy vs. information loss index. Our systematic experiments show that the proposed method achieves outstanding performance in privacy preserving yet maintains good clustering results.

RELATED WORKS

The random distortion method attempts to preserve privacy of the data by modifying values of the sensitive attributes using a randomized process. Two commonly used methods are probability distortion and point distortion [4]. Among point distortions, value distortion is the most prominent one.

Oliveira and Zaiane [5] introduced a new method called dimensionality reduction-based transformation for privacy preserving clustering over centralized and vertically partitioned data. The method relies on the intuition behind random projection to protect the underlying attribute values subjected to cluster analysis. Inan et al. [6] proposed a method for constructing the dissimilarity matrix of objects from different sites in a privacy preserving manner which can be used for clustering over horizontally partitioned data. Jagannathan et al. [7] presented a simple I/O-efficient k-clustering algorithm that was designed with the goal of enabling a privacy-preserving version of the algorithm. Merugu and Ghosh [8] presented a general framework for distributed clustering that takes into account privacy requirements. It is based on building probabilities modes of the data at each local site, whose parameters are then transmitted to a central location. Jagannathan and Wright [9] introduced the concept of arbitrarily partitioned data and provided a privacy preserving protocol for k-means clustering using cryptographic techniques in the setting of arbitrarily partitioned data. Vaidya and Clifton [3] also proposed a method for k-means clustering over vertically partitioned data. Klusch et al. [10] proposed Kernel density estimation based clustering method over distributed data. In Table 1, we summarize the major works in terms of specific data mining tasks, proposed methods, and performance measures used.

Table 1: A summary of privacy preserving clustering works

Ref.	DM task	Method	Measures
[5]	Clustering centralized & vertically partitioned data	DRBT	• Overall quality: F-measure • Communication cost • Accuracy • Stress error
[6]	Clustering horizontally partitioned data	Constructing dissimilarity matrix	• Communication cost • Computation cost
[7]	*K*-Clustering over horizontally partitioned data	*K*-clustering algorithm	• Privacy • Communication complexity • Computation complexity
[8]	Clustering horizontally distributed data	Building probabilities modes	• Quality cost: KL-divergence • Privacy cost
[9]	*K*-means over arbitrarily partitioned data	Cryptographic techniques	• Communication complexity • Computation complexity • Privacy
[3]	*K*-means over vertically partitioned data	Cryptographic techniques	• Security • Communication cost
[10]	Clustering over distributed data	KDEC	• Privacy • Communication cost

As shown in Table 1, most researchers emphasized clustering task over distributed data. Correspondingly, communication cost is the most concerned measure. These works seldom used the specific metric for clustering except for F-measures [5]. Specific data mining techniques are selected for specific tasking, and specific performance metrics should be used according to the characteristics of certain data mining task.

CHAOTIC DISTORTION

Since both probability and value distortions have non-negligible drawbacks for privacy preserving data mining, effort has been made to explore a different method for better protection. Undoubtedly, it is a reasonable idea to modify the individual sensitive information with aggregate distribution unchanged. The key point is how to choose suitable noise to distort the original data. The so-called noise should meet two conditions:

1) It should be noisy enough in order to protect the individual privacy.
2) It does not change the data distribution significantly in statistics.

Clearly, we need a kind of noise, which cannot be predicted as white noise of certain distributions. A chaos system looks like but not a random system, which has the important characteristics of erotic and sensitive dependence on initial conditions. If the initial condition and chaos equation are unknown, the chaotic data series cannot be predicted. These features make it suitable for data distortion. Suitable chaotic data series within a limited range is a feasible candidate for data distortion for privacy preserving data mining.

Unpredictability of Chaos System

A chaotic system is a deterministic system that shows an irregular oscillatory process. It is difficult to distinguish chaos from random behavior [12]. There are three fundamental characteristics of chaos which makes it suitable for data distortion: 1) irregular periodicity, 2) sensitivity to initial conditions, and 3) a lack of predictability. Because of these characteristics, chaos function has been used by modern cryptographic schemes [13]. Logistic differential equation or logistic map, though simple, displays the major chaotic concepts. The logistic model can be expressed as:

$$x_{k+1} = rx_k(1-x_k),\ k=1, 2, 3..,\ r \in [3.57, 4] \quad (1)$$

Sensitivity to initial conditions leads to the unpredictability. This is the key reason of using a chaos system for data distortion.

Chaotic Distortion Procedure

Chaotic distortion is a method that adds a chaotic time series to the original data as the perturbed data such that:

$$Y = X + f\{X, L(X_0, r)\} \quad (2)$$

Where $L(X_0, r)$ is the logistic chaos matrix with the initial conditions matrix X_0 and parameter r. The $f\{X, L(x_0, r)\}$ represents the chaotic series matrix used to perturb the original data set X, which is the function of original dataset X and the logistic chaos matrix. In order to obtain a greater distortion, the chaotic matrix must be designed to fit the distribution of original dataset X. The perturbed data set Y will be displayed, replacing the original data. The privacy of individual information is thus well protected because of the unpredictability of the chaos series.

The proposed procedure is as follows:

1) *Select the chaos equation to produce the chaotic time series*. Here we use a logistic map with the parameter r = 4 for chaotic noise generation.
2) *Randomly generate the initial condition matrix of the chaos system*. The randomization of initial condition matrix makes it more difficult to predict the chaotic matrix and thus the privacy information can be protected.
3) *Produce chaotic matrix, analyze the data characters, and modify the chaotic series if necessary*. The chaotic time series is used as noise, which cannot be too prominent to mine the aggregate distribution of the original data. A common method is to modify chaotic data series based on the value ranges of every attribute in original dataset.
4) *Distort the original data by adding the chaotic data to produce the perturbed data set according to equation 2.*
5) *Analyze the aggregate characteristics of the perturbed data set to verify that the aggregate information loss is at an acceptable level.*
6) *Publish the perturbed data if the aggregate distribution has not been changed significantly.*

PERFORMANCE METRICS

We propose three measures based on the characteristics of PPKC. For k-means clustering, the cluster centers indicate the population information used for decision making. Therefore the cluster centers after data distortion should be kept as close as possible to the original centers to minimize the population information loss. We use centers distance to measure information loss. On the other hand, the individual objects should be distorted to be as different as possible from the original individuals in order to protect the individuals' privacy. We use the average distance between the original and correspondent perturbed individuals to measure the privacy protection performance. Whereas, these measures may be varying greatly for different datasets because of the different data ranges. With this consideration, a comprehensive index to measure the overall performance is also proposed.

Information Loss Measure

Our evaluation focuses on the overall quality of the generated clusters after data distortion. We compare how closely each cluster center of the distorted data matches its corresponding cluster center in the original data set.

Definition 1: K cluster centers $C_i = (c_{i1}, c_{i2}, \cdots c_{in})$, where $i = 1,2,\cdots,k$, are obtained based on the original data set with *k*-means clustering, $C_i^d = (c_{i1}^d, c_{i2}^d, \cdots c_{in}^d)$, where $i = 1,2,\cdots,k$, are corresponding cluster centers based on the distorted data set. The information loss of distortion (ILD) of cluster center C_i can be measured by the Euclidean distance between the corresponding cluster center pairs:

$$ILD_i = \sqrt{\sum_{j=1}^{n} (c_{ij} - c_{ij}^d)^2} \tag{3}$$

The average information loss ILD is the mean value of ILD_i ($i = 1,2,\cdots,k$).

$$ILD = \frac{1}{k} \sum_{i=1}^{k} ILD_i \tag{4}$$

Clearly, smaller ILD values indicate less information loss.

Privacy Protection Measure

The privacy protection measure should indicate how closely the original value of an item can be estimated from the distorted data [11]. We define a privacy protection measure as the average Euclidean distance between each data and its corresponding distorted data.

Definition 2: Consider a dataset matrix with *m* records and *m* attributes, $X = [x_{ij}]_{m \times n}$, and the distorted data set, $Y = [y_{ij}]_{m \times n}$. The privacy protection of the distortion (PPD) can be measured by the average Euclidean distance:

$$PPD = \frac{1}{m} \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} (x_{ij} - y_{ij})^2} \tag{5}$$

Clearly, a larger PPD value indicates a better privacy protection.

Comprehensive Index

A tradeoff between individual privacy protection and population information loss is needed. In order to protect individuals' privacy, a preferred high degree distortion of original data may cause large information loss. On the contrary, less distortion of original data may cause more risk of privacy disclosure. A comprehensive index of distortion (CID) is proposed to measure the overall performance of PPKC with the consideration of both privacy protection and population information loss.

$$CID = \frac{PPD}{ILD} \tag{6}$$

Clearly, a larger CID indicates a better overall performance of distortion. Specifically, if CID is greater than 1, the privacy protection effects overtake the degree of information loss. We may conclude that the distortion with CID being greater than 1 is a good performance, while the distortion with CID being less than 1 is an unacceptable performance.

PERFORMANCE EVALUATION

We empirically validate the chaotic distortion for privacy preserving k-means clustering using four real datasets. The experiments are conducted on a PC with AMD Athlon 64 processor, 1.81 GHZ clock speed and 512 MB RAM. The algorithm is programmed with Matlab clustering toolbox [14].

Data sets

We use four real-world datasets to evaluate the relative performance of the proposed method. Obviously, most attributes of these data sets are confidential business and financial information.

1) *Census data* [15]. There are 1080 individual records with 13 census related attributes, such as final weight, gross income, employer contribution for health insurance, net earnings, federal income tax liability, social security retirement payroll deduction, interest income, total person earnings, total other persons income, total person income, state income tax liability, taxable income, and total wage & salary.
2) *Housing data* [16]. There are 506 individuals records with 15 attributes, including housing price, age, number of rooms, property-tax rate, crime rate by town, , and so on.
3) *Tarragona data* [15]. There are 834 records with 13 financial accounting attributes including fixed assets, current assets, uncommitted funds, paid-up capital, short-term debt, sales, labor costs, depreciation, operating profit, financial outcome, gross profit, and net profit.
4) *Wages data* [17]. This data set consists of a random sample of 534 persons from the CPS with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership.

Experimental Settings

1) *Data sets duplications:* Every dataset is duplicated to produce 10 derived data sets with disordering the sequences of records. We name the original data set and the 10 corresponding disordered data sets as a data set family. For example, we refer *census* and the corresponding 10 derived data sets as census-family. In total, we have four dataset families and every family contains 11 data sets. We reproduce the data families for two purposes. First, we want to test the sensitivity of the proposed method to the record orders in a dataset. Second, the experiment scale is enlarged to make the analysis result more reliable.
2) *No. of clusters:* In the course of *k*-means clustering the number of clusters must be given by the user before the clustering, but it is rarely known apriori [14]. In this case it must be tested with different numbers of clusters with validity measures. Therefore we conduct the experiments on every data set with 7 cases of the number of clusters from 2 to 8.
3) *Evaluation metrics:* In addition to the three measures proposed in this paper, Dunn's Index (DI) [18] is used to measure the clustering validity for *k*-means method. The relationship of CID vs. DI is also analyzed to reveal the underlying rules. According to [19], the larger values of the index indicate the presence of compact and well-separated clusters.

Testing Procedure

Our performance evaluation was carried out through the following steps:

1) *Chaotic distortion*. The most important for chaotic distortion is to generate the suitable chaotic matrix for data distortion according to the characteristics of the original data set as shown in equation 5. The initial condition matrix X_0 of the logistic matrix is generated randomly by uniform distribution; the r is set as 4. In order to minimize the overall information loss, we subtract 0.5 from every data of logistic matrix, and produced the chaotic matrix L . The final chaotic distortion series for *ith* attribute of dataset, $f_i\{X_i, L_i\}$, is obtained based on the dataset distribution characteristics as follows:

$$f_i\{X_i, L_i\} = 0.1 \cdot L_i \cdot (\max(X_i) - \min(X_i)) \tag{7}$$

Where X_i is the *ith* attribute of the original dataset *X*, L_i is the *ith* column of the logistic matrix *L*.

2) *Dataset normalization*. We normalize the original and distorted data sets with min-max normalization.
3) *K-means clustering over original and distorted data sets.* For every data set, we run *k*-means clustering 10 times for a specific number of clusters to eliminate the influence of initial randomization of the algorithm. The numbers of clusters are assigned to 2 to 8 for every dataset. In total, every dataset is clustered for 70 times (7x10).
4) *Evaluation metrics calculation*. The metrics include privacy protection measure PPD, information loss measure ILD, Comprehensive index CID, clustering validation index DI, and CPU time.

Results and Discussion

We show the average metrics CID, DI and time in seconds for every data set family in Tables 2. Overall, the comprehensive performance of privacy protection vs. information loss is satisfying because all CID values are greater than 1. The computations are fairly efficient because all CPU time is less than one second.

Table 2: Summary of computational results

(a) Census-family

Index	No. of clusters						
	2	3	4	5	6	7	8
CID	24.70	17.09	6.18	8.68	3.06	2.44	1.78
DI	0.052	0.037	0.04	0.039	0.04	0.04	0.04
Time	0.610	0.270	0.47	0.630	0.40	0.50	0.57

(b) Housing-family

Index	No. of clusters						
	2	3	4	5	6	7	8
CID	20.28	8.96	2.58	0.64	1.54	1.38	1.53
DI	0.206	0.054	0.059	0.062	0.06	0.059	0.058
Time	0.060	0.150	0.170	0.210	0.12	0.130	0.150

(c) Tarragona-family

Index	No. of clusters						
	2	3	4	5	6	7	8
CID	1057	5.76	4.36	2.960	2.49	2.470	2.42
DI	0.128	0.043	0.022	0.014	0.011	0.009	0.008
Time	0.120	0.260	0.470	0.600	0.410	0.490	0.530

(d) Wages-family

Index	No. of clusters						
	2	3	4	5	6	7	8
CID	13.22	4.87	1.88	1.95	1.65	1.49	1.49
DI	0.378	0.347	0.262	0.238	0.201	0.176	0.148
time	0.050	0.080	0.110	0.140	0.090	0.090	0.090

IN-DEPTH ANALYSES AND RESULTS

Sensitivity to The Order of Records

We test the PPKC's sensitivity to the order of records to analyze whether the order of records will influence the performance of PPKC or not. In order to eliminate the influence of the number of clusters, we calculate the average CID with all cases of cluster numbers. Then ANOVA is conducted for every data set family as shown in Table 3. The large P-values indicate that they are not statistically different between data sets in a family. Therefore, we conclude that the proposed method is not sensitive to the order of records.

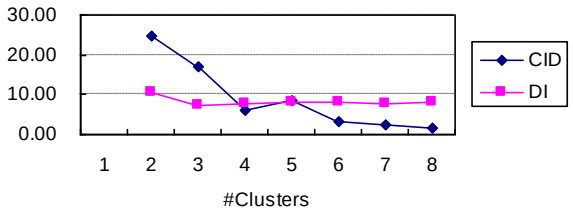
Table 3: ANOVA: CID sensitivity to object orders

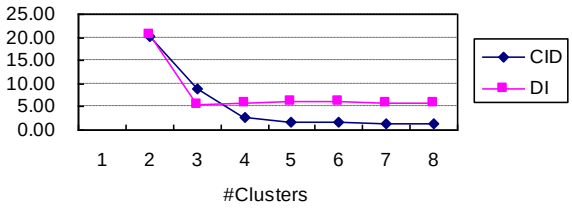
Index	Census	Housing	Tarragona	Wages
P-value	0.23	0.65	0.49	0.30
F	1.33	0.78	0.95	1.20

CID vs. DI

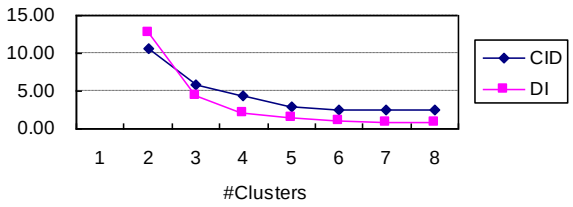
Naturally, we want to know whether there are relationships between CID and DI. In order to clearly show the relationship, we scale up the values of DI as compatible with CID. Figure 1 shows the relationships between CID and DI for the 4 data set families. The common rule underlying these figures is that CID and DI has same trends. The coefficient of correlation between CID and DI are 0.69, 0.91, 0.99, and 0.82 separately for census, housing, Tarragona, and wages data set families. This reveals that the CID measure is especially suitable for *k*-means clustering.

(a) Census-family (b) Housing-family





(c) Tarragona-family



(d) Wages-family

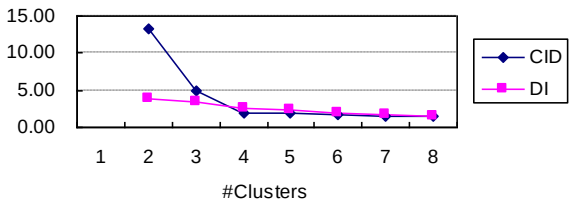


Figure 1: Comparisons of CID vs. DI

Privacy Protection vs. Information Loss

For any privacy preserving data mining task, we hope to have less population information loss but with more privacy protection. Although the values of comprehensive index CID for all data sets are greater than 1, we use statistical analysis to check if there are statistically significant between privacy protection and information loss. For every specified number of clusters, we conducted paired t-test between PPD and ILD with the results of all datasets. As shown in Table 4, the values of PPD are significantly greater than ILD for all cases of cluster numbers. This proved that the proposed method can protect privacy with relatively little information loss.

Table 4: Paired t-test: PPD vs. ILD

Index	No. of clusters						
	2	3	4	5	6	7	8
T stat	3.90	3.97	3.91	3.98	3.95	3.89	3.84
P value	0.000	0.000	0.000	0.000	0.000	0.000	0.000

CONCLUSIONS

Data distortion is a popular privacy preserving method. Whereas, the original data may be constructed with a suitable approach, and thus the privacy may be disclosed. In order to satisfy the conflicting objectives of providing high quality distorted dataset and at the same time preventing from exact or partial disclosure of individual information, a chaotic distortion technique is proposed for privacy preserving *k*-means clustering. A simple chaos system, logistic map, is used to produce the chaotic series as a noise replacing the white noise. It is sensitive to initial conditions and unpredictable, which makes the chaos system suitable for privacy preserving.

Three performance metrics specifically for *k*-means clustering are proposed including privacy protection index PPD, information loss measure ILD, and a comprehensive index CID for privacy preserving k-means clustering. A large-scale experimental analysis using four real world dataset families is conducted to verify the performance of the proposed approach. Our computational experience shows that the chaotic distortion method not only can protect individual privacy but also can maintain the original cluster centers information. It is also quite computationally efficient. Future studies can be extended to evaluate the relative performance of different chaotic distortion systems, as well as its appropriateness for other privacy preserving data mining methods.

REFERENCES

[1] Dutta, H., Kargupta, H. and Datta, S. (2003) "Analysis of privacy preserving random perturbation techniques: Further explorations," *Proceedings of the 2003 ACM workshop on Privacy in the Electronic Society*, pp. 31-38.

[2] Kargupta, H. Datta, S. Wang, Q. and Sivakumar, K. (2003) "On the privacy preserving properties of random data perturbation techniques," *Proceedings of the third IEEE International Conference on Data Mining*, pp. 99-106.

[3] Vaidya, J. and Clifton, C. (2003) "Privacy-preserving k-means clustering over vertically partitioned data," *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 206-215.

[4] Liew, C.K., Choi, U.J. and Liew, C.J. (1985) "A data distortion by probability distribution," *ACM Transactions on Database Systems*, Vol. 10, No. 3, pp. 395-411.

[5] Oliveira, S.R.M. and Zaïane, O.R. (2007) "A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration," *Computers and Security*, Vol. 26, pp. 81-93.
[6] Inan, A., Kaya, S., Saygin, Y., Savas, E., Hintoglu, A.A. and Levi, A. (2007) "Privacy preserving clustering on horizontally partitioned data," *Data & Knowledge Engineering*, Vol. 63, No. 3, pp. 646-666.

The complete reference is available upon request from the third author – Chao-Hsien Chu (chu@ist.psu.edu).