*IOSR Journal of Computer Engineering (IOSRJCE)*
*ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 2 (Sep-Oct. 2012), PP 25-29*
*www.iosrjournals.org*

# Privacy Preserving Clustering on Distorted data

Thanveer Jahan[1], Dr.G.Narasimha[2], Dr.C.V.Guru Rao[3]
[1](M.Tech, (Ph d(CSE)), JNTU, Kukatpally ,Hyderabad, A.P,India)
[2](Assistant Professor (CSE), JNTU, Kukatpally, Hyderabad, A.P, India)
[3](HOD (CSE), SR Engineering College, Warangal, A.P, India)

***Abstract:*** *In designing various security and privacy related data mining applications, privacy preserving has become a major concern. Protecting sensitive or confidential information in data mining is an important long term goal. An increased data disclosure risks may encounter when it is released. Various data distortion techniques are widely used to protect sensitive data; these approaches protect data by adding noise or by different matrix decomposition methods. In this paper we primarily focus, data distortion methods such as singular value decomposition (SVD) and sparsified singular value decomposition (SSVD). Various privacy metrics have been proved to measure the difference between original dataset and distorted dataset and degree of privacy protection. The data mining utility k-means clustering is used on these distorted datasets. Our experimental results use a real world dataset. An efficient solution is achieved using sparsified singular value decomposition and singular value decomposition, meeting privacy requirements. The accuracy while using the distorted data is almost equal to that of the original dataset.*

***Keywords****- Privacy Preserving, Data Distortion, Singular Value Decomposition (SVD), Sparsified Singular Value Decomposition (SSVD), k--means clustering.*

## I. Introduction

The rapid increase of applications in data mining has raised a major concern in corporations. Private information of innocent people is collected and used for data mining purpose. In this modern technology, data of various kinds are collected and exchanged at an unprecedented speed and scale. A latest and fruitful direction for future research is to efficiently discover valuable information from large datasets and to develop techniques that incorporate privacy concerns. Now a day's data is an important asset of companies, governments, and research institutions [10] and is used for various public and private interests. Data is sensitive to privacy issues. Defense applications, financial transactions, healthcare records and network communication traffic are few of examples. Preserving Privacy in sensitive domains has become a major concern in data mining applications. Many data mining applications would not be acceptable, without an acceptable level of privacy of sensitive information. Data can be collected at centralized or distributed location. In centralized location, major concern is to shield the exact values of the attributes from the data analysts, where as in distributed locations data storage patterns are different i.e. they are horizontally distributed or vertically distributed [1, 8].There has been much research on privacy preserving data mining (PPDM) based on data perturbation or data distortion, randomization and secure multi party computations. The goal of privacy-preserving data mining techniques is used to hide sensitive or confidential data values from an unauthorized user and preserve data patterns. These patterns and semantics are used to build a valid decision model on distorted data sets. Different data mining techniques such as classification, clustering etc are proposed for privacy protection in data processing. The best scenario is to construct a data pattern model on distorted data equivalent to or better than an original data.

There are two approaches in this case i.e. to distort the data so that the analysts are unaware of original data and the second approach is to modify the data mining algorithms. In this paper we propose the first approach the analysts uses distorted dataset transformed into data matrix $\overline{D}$, not the original dataset $D$. The matrix $\overline{D}$ cannot be used to reconstruct the original matrix $D$, without knowing the error part $E = D - \overline{D}$. The analysts are unable to know attributes (columns) of original attributes and apply data mining algorithms on it. In this way data privacy preservation is premised on the maintenance of data analytical values. We transformed original dataset into distorted dataset to protect privacy. Among the widely used approach is Singular value decomposition (SVD) and its derivative Sparsified Singular value Decomposition (SSVD) are the one most popular techniques to address issues. SSVD was first introduced by Gao and Zhang in [4] to reduce cost and enhance performance of SVD in text retrieval applications. Xu et al. applied SVD and SSVD methods in terrorist analysis system [15,16]. SSVD was further studied in [5] in which structural partition strategies proposed to partition data into submatrices. In Ref. [11] privacy preserving clustering in singular value decomposition (SVD) was proposed and the results proved that accuracy of original and distorted dataset are equivalent. In our work, we take a closer look to perform data distortion by singular value decomposition and

sparsified singular value decomposition. Thus, data mining techniques k-means clustering is applied on the distorted dataset to attain inherent property of privacy protection.

The remaining part of the paper is organized as follows; Section 2 briefly introduces related work on data analysis system and data distortion methods: SVD and sparsified SVD and k-means clustering. Section 3 discusses the various data perturbation metrics. The experiments are carried out and the results are presented and discussed in Section 4. We finally sum up this paper and bring our future plans in Section 5.

## II. Related Work

### 2.1 Privacy preserving data mining

There has been a raising concern for disclosure of security and privacy, as the data mining techniques gain popularity and widely used in business and research. Two parties having private data wish to work in collaboration by to other party. Indeed, neither party shares their private data. In such cases privacy preserving data mining (PPDM) have major significance. PPDM develops algorithms for modifying the original data in a way that data and knowledge remain private even after mining process [12]. Common techniques include data perturbation, blocking feature values, swapping tuples etc. PPDM scheme should able to maximize the degree of data modification to retain the maximum data utility level.

### 2.2 Analysis system and data distortion

A simplified model of data analysis system consists of two parts, the data manipulation and the data analysis as illustrated in Fig 1.The original data is completely manipulated by the authorized user's or data owner using data distortion process i.e. matrix decomposition method .Data distortion is one of the important parts in many privacy preserving data mining tasks. The distorted methods must preserve data privacy and at the same time must keep the utility of the data after distortion. The data distorted or perturbated data is collected by analysts to perform all actions such as clustering etc. The protected data maintains privacy as analysts is unknown with actual data values. The classical data distortion methods are based on random value perturbation and are applied [8]. Singular value decomposition (SVD) is a popular method in data mining and information retrieval

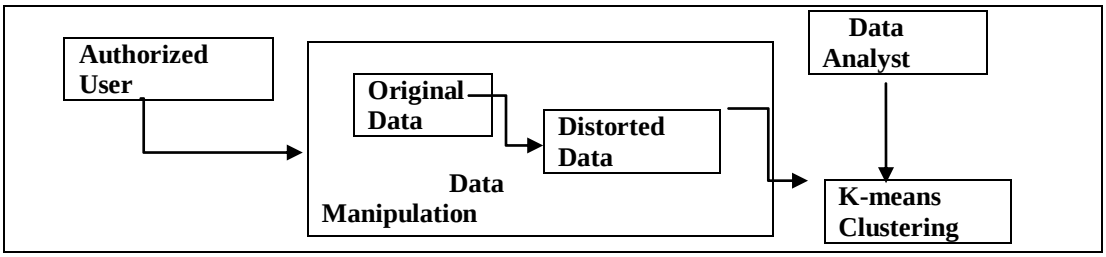


**Figure 1. A Data Analysis system for Clustering**

[9].SVD has numerous applications in data mining, information retrieval and image compression in which it is often used to approximate a given a matrix by a lower rank matrix with minimum distance between them. SVD is used to reduce dimensionality of the original dataset *D*. A sparse matrix *D* of dimension p×q represents the original dataset. The rows and columns are the data objects and attributes. The singular value decomposition of the matrix D is [3].

$$D = U S V^T$$

Where *U* is an p×p orthonormal matrix, $S$ = diag $[\sigma_1, \sigma_2, \ldots\ldots, \sigma_s]$ (s = min{p,q})is an p×q diagonal matrix whose nonnegative diagonal entries are in a descending order, and $V^T$ is an q×q orthonormal matrix. The number of nonzero diagonals of *S* is equal to the rank of the matrix *D*. The singular values in the matrix *S* are arranged in a descending order .The SVD transformation has property that the maximal variation among objects is captured in the first dimension, as $\sigma_1 \geq \sigma_i$ for $i \geq 2$.The remaining variations are captured similarly in the second dimension and so on. Thus, a transformed matrix with a lower dimension can be constructed to represent the original matrix i.e.

$$D_r = U_r S_r V_r^T$$

Where $U_r$ contains the first *r* columns of *U*, $S_r$ contains the first *r* nonzero diagonals of *S* and $V_r^T$ contains the first *r* rows of $V^T$ .The rank of the matrix $D_r$ is *r* and with *r* being small, the dimensionality of the dataset has been gradually reduced from min {p, q} to *r* (assuming all attributes are linearly dependent). $D_r$ is proved to be best *r* dimensional approximation of *D* in the sense of Frobenius norm. In data mining applications the use of $D_r$ to represent *D* has important function. The removed error part $E_r = D - D_r$ can be considered as the noise in the original dataset *D* [8]. Mining on reduced dataset $D_r$ yield better results than on original dataset *D*. Thus, the distorted data $D_r$ can provide effective protection for data privacy. Sparsified SVD is a data distortion method

better than a SVD in preserving privacy. After reducing rank of the SVD matrices, we set small size entries which are smaller than a certain threshold ϵ in $U_r$ and $V_r^T$ to zero. This operation is referred as a dropping operation [4]. Thus, drop $u_{ij}$ in $U_r$ , if $|u_{ij}| < \epsilon$ and $v_{ij}$ in $V_r^T$ , if $|v_{ij}| < \epsilon$. Let $U_r$ denote $U_r$ with dropped elements and $V_r^T$ denote $V_r^T$ with dropped elements, the distorted data matrix $D_r$ is represented as

$$D_r = U_r \, S \, V_r^T ,$$

The sparsified SVD method is equivalent to further distorting the dataset $D_r$. Denote

$$E_\epsilon = D_r - \overline{D_r} ,$$

$$D = \overline{D_r} + E_r + E_\epsilon ,$$

The data provided to analysts is $\overline{D_r}$ which is twice distorted in the sparsified SVD method. The sparsified SVD was proposed by Gao and Zhang in [4] for reducing storage cost and enhancing performance of SVD in text retrieval applications.

**2.3 K-means Clustering**

Clustering is a well-known problem in statistics and engineering, namely, how to arrange a set of vectors (measurements) into a number of groups (clusters). Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization [6]. The problem has been formulated in various ways in the machine learning, pattern recognition optimization and statistics literature. The fundamental clustering problem is that of grouping together (clustering) data items that are similar to each other. Given a set of data items, clustering algorithms group similar items together. Clustering has many applications, such as customer behavior analysis, targeted marketing, forensics, and bioinformatics.

## III. Data Perturbation Metrics

In literature privacy metrics have been proposed in [2, 10] .In Ref. [2] the metrics are incomplete and is proved in Ref. [8].It is important to know the density function of each attribute a priori, which may be difficult to obtain for the real world datasets. We propose some privacy measures which depend on the original matrix *D* and its distorted matrix *D*.

**3.1 Value difference (VD)**

The elements of data matrix change after distortion. The value difference (VD) of the datasets is represented by relative value difference in the Forbenius norm. VD is the ratio of the Forbenius norm of the difference of *D* and |*D*| to the Forbenius form of D.

$$VD = \| D - |D| \|_F \; / \; \| D \|_F .$$

**3.2 Position difference**

several metrics are used to measure position difference of the data elements. RP is used to denote average change of order of all attributes. The order of the element changes after distortion. Dataset *D* has n data objects and m attributes. $Ord^i_j$ denotes the ascending order of the jth element in attributes i, and $Ord^i_j$ denotes the ascending order of the distorted element $D_{ij}$Then, RP = $( \sum_{i=1}^{m} \sum_{j=1}^{n} | Ordij - Ordij | ) / (m * n)$

RK represents the percentage of elements that keep their orders of value in each column after the distortion.

$$RK = ( \sum_{i=1}^{m} \sum_{j=1}^{n} RK^i_j ) / ( m * n),$$

$$RK^i_j = \begin{cases} 1, \text{ if } Ord^i_j = \overline{Ord}^i_j , \\ 0, \text{ otherwise.} \end{cases}$$

The metric CP is used to define the change of order of the average value of attributes:

$$CP = ( \sum_{i=1}^{m} | (Ord DV_i - \overline{Ord} DV_i ) | \, m,$$

where $OrdDV_i$ is the ascending order the average value of attribute *i* while $OrdDV_i$ denotes ascending order after distortion. Like RK, we define CK to measure the percentage of the attribute that keeps their order of average value after distortion.

$$CK = ( \sum_{i=1}^{m} CK^i ) / m,$$

where $CK^i = \begin{cases} 1, \text{ if } OrdDV_i = \overline{Ord} DV_i , \\ 0, \text{ otherwise} \end{cases}$

The higher the value of RP and CP and the lower the value of RK and CK, the more privacy can be preserved. In order to be fair for a dataset, privacy metrics are calculated as shown in the Table 1.The value of VD,RP,CP is more in SSVD than in SVD .Among the four distortion methods SSVD is better than SVD to preserve privacy as shown in Fig 4.

**Table 1.Comparision of Privacy Metrics for distortion methods**

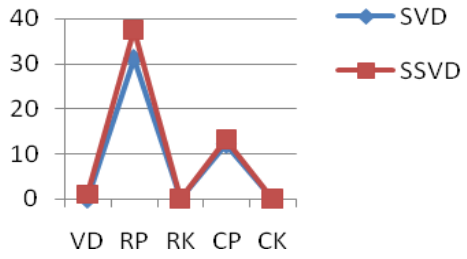


**Figure 2. Performance of privacy metrics**

| | | Privacy Metrics | | | |
|---|---|---|---|---|---|
| **DATA** | **VD** | **RP** | **RK** | **CP** | **CK** |
| **Org** | — | — | — | — | — |
| **SVD** | 0.0525 | 31.2 | 0.0251 | 12.2 | 0.12 |
| **SSVD** | 1.0422 | 37.5 | 0.0066 | 13.1 | 0.05 |

## IV. Experiments And Results

We conduct experiments on real data set having 100 data points. For a real world dataset, we downloaded information about 100 terrorists (q), 42 attributes (p) such as age, place, relationship etc. The original matrix is of dimension 100×42.

### 4.1 Proposed Algorithm

**Input:** Data matrix D, No of clusters k,
**Output:** Distorted Data matrix $\overline{D_r}$ , Clusters
**Step 1:** Finding sensitive or confidential attributes ($p_i$) i= 0,1,…….41 in *D.*
**Step 2:** Form the matrix *C.C* = [$p_0$,$p_1$,$p_2$,…$p_{41}$]
**Step 3:** Apply SVD to the matrix *C.*
SVD(*C*) = $U\,S\,V^T$ ,
Then distorted matrix $C_r = U_r\,S_r\,V_r^T$
**Step 4:** Then, apply SSVD to matrix $C_r$
Choosing the rank r and dropping
Threshold ϵ as $10^{-3}$
Then Distorted matrix SSVD ( $\overline{C_r}$) = $\overline{U_r}\,S_r\,\overline{V_r}^T$
**Step 4:** Update $\overline{C_r}$ in *D* , gives $\overline{D_r}$
**Step 5:** Generate Clusters for sensitive attribute in $\overline{D_r}$ .

**Table 2: Data objects and Clusters**

| **Clusters(k)** | **Original data** | | **SVD** | | | | **SSVD** | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 3 | 4 |
| **Data objects(points)** | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 1 |
| | 1 | 3 | 3 | 1 | 3 | 1 | 2 | 2 | 2 |
| | 1 | 3 | 3 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |
| | 1 | 3 | 1 | 1 | 3 | 1 | 2 | 3 | 2 |

The illustration of the above method is represented for 10 data objects is shown in the Table 2.we analyzed a specific number of clusters ranging from 2 to 4 clusters. The effectiveness is measured in terms of the proportion of the points that are grouped in the same clusters after we apply a transformation on data, such points as legitimate points. Considering the transformation attributes as relationship with terrorist group. k denotes the number of clusters to group the data objects. For the clusters k=2and k=3 the data objects grouped in original dataset and in SVD dataset are exactly same. In SSVD data objects are effectively grouped when, compared to original and SVD data set for k=2,3,4 .

**4.2 Measuring Accuracy**

The efficiency is measured on the number of data points those are legitimate and are grouped in the original and distorted datasets. k- means clustering do not consider noise.

A Misclassification Error is used to concentrate on a potential problem where the data point from a cluster migrates to a different cluster.

ME = 1/ N * $\sum_{i=1}^{K}(|Clusteri(D)| - |Clusteri(Dr)| - |Clusteri(D_r)|)$

**Table 3. Results of Misclassification Error**

| Data objects(points) | Original data set | | | Distorted dataset-SVD | | | Distorted dataset-SSVD | | |
|---|---|---|---|---|---|---|---|---|---|
| | K=2 | K=3 | K=4 | K=2 | K=3 | K=4 | K=2 | K=3 | K=4 |
| **10** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| **100** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |

Misclassification error must be 0% where N represents the number of point in the original dataset, k is the number of clusters under analysis and | Cluster$_i$(*D*)| represents the number of data points legitimate in the ith cluster in dataset *D*. The results are tabulated in the Table 3.The cluster analysis yields good results for the original and distorted datasets using SVD and SSVD distortion techniques. The results suggest that our techniques perform well to achieve feasible solution. The accuracy of distorted data is same as original data. Thus, a complete privacy can be obtained in k-means cluster analysis and is also proved in privacy metrics.

## V. Conclusion

We propose a better approach for a data analysis system to use data distortion techniques: singular value decomposition (SVD) and Sparsified SVD to preserve privacy. We have presented privacy preserving data mining application which distorts original dataset to meet privacy requirements. Experimental results show the effectiveness by measuring accuracy of original data and distorted data. It has proved that high degree of data distortion can maintain high level of data utility using k-means clustering. Future work may address other scenarios to protect data along with different data mining algorithms.

## References


[1] B.Gillburd, A. Schuster and R.Wolff. "K- TTP: A newPrivacymodelforlargescaledistributedenvironments". In *Proceedings of the 10th ACM SIGKDD* International Conference on Knowledge Discovery and Data Mining (kdd'04), Seattle, WA, USA, 2004.

[2] D.Agrawal and C .C.Aggarwal "On the and quantification of privacy –Preserving data minng algorithms". *In Proceedings of the 20th ACM SIGACT-SIGMOD.*

[3] G.H.Golub andC.F.van Loan. Matrix Computation, John Hopkins Univ., 3rd ED., 1996.

[4] J.Gao, J. Zhang. "Sparsification strategies in latent semantic indexing*", In Proceedings of the 2003 Text Mining Workshop, M.W. Berry and W.M.* Pottenger, (ed.), pp. 93-103, San Francisco, CA, May 3, 2003.

[5] J.Wang, W.Zhong,S. Xu,J.Zhang. "Selective Data Distortion via Structural Partition and SSVD for privacy Preseervation",In *Proceedings of the 2006 International Conference on Information & knowledge* Engineering,pp:114-120,CSREA Press,Las Vegas.

[6] J. Han, M. Kamber. Concepts and Techniques. Morgan Kaufmann Publishers, 2001

[7] N. Maheswari & K. Duraiswamy CLUST- SVD: Privacy preservinglusteringinsingularvaluedecomposition In World journal of modeling and simulation vol 4(2008) No4,pp250-256.

[8] R.Agrawal, A.Evfimieski and R.srikanth."Information sharing across private databases". *In proceedinds of the 2003 ACM SIGMOD* International Conference on management of data, pp.86-97,san Diego,CA,2003

[9] R.Agrawal and R.srikant"Privacy–Preserving data mining". *In proceedings of the 2000 ACM SIGMOD* International Conference on management of data, pp86-97, San Diego, CA, 2003.

[10] S.Deewester, S.Dumais, et al."*Indexing by latent semantic analysis"*, J Amer. Soc.Info.Sci, 41:391-407, 1990.

[11] V. S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis. "*State-of-the-art I privacy preserving data mining*", SIGMOD Record, 33(1):50-57, 2004

[12] V. S. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, andY. Theodoridis. "*State-of-the-art in privacy preserving data mining*",SIGMOD Record, 33(1):50-57, 2004

[13] W.Frankes and R.Baeza-Yates. *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Englewood Cliffs,NJ,19192

[14] V.Estvill–Castro, L Brankovic and D.L.Dowe. "Privacy *in data mining*".Australian Computer Society NSWBranch, Australia. Available at www.acs.org.au/nsw/articles/199082.html

[15] S. Xu, J. Zhang, D. Han, J.Wang. Data distortion for privacy protection in a terrorist analysis system. in: *Proceedings of the 2005 IEEE International Co*nference on Intelligence and Security Informatics, 2005, 459–464.

[16] S. Xu, J. Zhang, D. Han, J. Wang. Singular value decomposition based data distortion strategy for privacy protection.Knowledge and Information Systems, 2006, 10(3): 383–397.