Journal of Computer Science 2012, 8 (12), 2042-2052
ISSN 1549-3636

doi:10.3844/jcssp.2012.2042.2052 Published Online 8 (12) 2012 (http://www.thescipub.com/jcs.toc)

Secured Disclosure of Sensitive Data in Data Mining Techniques

[1]Kirubhakar Gurusamy and [2]Venkatesh Chakrapani

[1]Department of Computer Science and Engineering, Faculty of Engineering,
Surya Engineering College, Erode, Tamilnadu, India
[2]Department of Electronics and Communication Engineering, Faculty of Engineering,
Erode Builder Educational Trusts Group of Institutions, Kangayam, Tamilnadu, India

Received 2012-07-13, Revised 2012-10-17; Accepted 2012-12-18

ABSTRACT

Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of securing sensitive data. People have become increasingly unwilling to share their data. This frequently results in individuals either refusing to share their data or providing incorrect data. In turn, such problems in data collection can affect the success of data mining, which relies on sufficient amounts of accurate data in order to produce meaningful results. Based on the analysis of shortcomings of earlier technologies this study proposes a new method for securing numerical and categorical data. In this method the categorical data is converted into Binary form and perturbation based noise is introduced as a security method based on the security level anticipated. Several types of noise addition methods were employed and generalized results were evaluated in terms of misclassification error and privacy level. An average of misclassification error was below 50% for 75-90% security level, which is better than earlier methods which didn't handle categorical data. The results obtained prove that the proposed method outperforms some of the currently existing methods thereby ensuring the possibility of securing sensitive data irrespective of its type being numerical or categorical.

Keywords: Security, Privacy, Data Dissemination, Clustering, Quantification

1. INTRODUCTION

With the development of data analysis and processing technique, organizations, industries and governments are increasingly publishing micro data (i.e., data that contain non aggregated information about individuals) for data mining purposes, studying disease outbreaks or economic patterns. While the released datasets provide valuable information to researchers, they also contain sensitive information about individuals whose privacy may be at risk (Samarati, 2001).

1.1. Problem Definition

Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from the point of view of privacy preservation. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. However, there are situations where the sharing of data can lead to mutual benefit.

Despite the potential gain, this is often not possible due to the confidentiality issues which arise. It is well documented that the unlimited explosion of new information through the Internet and other media has reached a point where threats against privacy are very common and deserve serious thinking.

Consider a scenario that there are several hospitals involved in a multi-site medical study. Each hospital has its own data set containing patient records. These hospitals would like to conduct data mining over the data sets from

Corresponding Author: Kirubhakar Gurusamy, Department of Computer Science and Engineering, Surya Engineering College, Erode, Tamilnadu, India

all the hospitals with the goal of obtaining more valuable information via mining the joint data set. Due to privacy laws, one hospital cannot disclose their patient records to other hospitals. How can these hospitals achieve their objective? Can privacy and collaborative data mining coexist? In other words, can the collaborative parties somehow conduct data mining computations and obtain the desired results without compromising their data privacy? We show that privacy and collaborative data mining can be achieved at the same time.

The problem of secured data mining has found considerable attention in recent years because of the recent concerns on the privacy of underlying data (Verykios *et al*., 2004).

Various secured data mining techniques fall under:

- K-Anonymity
- Cryptographic techniques
- Randomized Response techniques
- Data modification

Many recent papers on privacy have focused on the perturbation model and its variants. Methods for inference attacks in the context of the perturbation model have been discussed by Ackerman *et al*. (1999).

The goal of this study is to present technologies to solve security related data mining problems over large data sets with reasonable efficiency.

1.2. Literature Survey

A number of papers have also appeared on the *k*-anonymity model recently. Other related works discuss the method of top-down specialization for privacy preservation and workload-aware methods for anonymization (Du *et al*., 2004). A related topic is that of privacy-preserving data mining in vertically or horizontally partitioned data (Chen and Liu, 2005). In this case, we determine aggregate characteristics of the data which are distributed across multiple sites without exchanging explicit information about individual records. The key in many of these approaches is to reduce the communication costs as much as possible while retaining privacy, Chawla *et al*. (2005). (Liu *et al*., 2006) discusses transformation based methods to preserve the anonymity of the data. This is different from our technique which uses group-based pseudo-data generation in order to preserve anonymity.

1.3. Related Work: K Anonymity

When releasing microdata for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samarati and Sweeney (1998); Sweeney (2002) introduced the k anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the k anonymity requirement, they used both generalization and suppression for data anonymization.

In general, k anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/k. While k anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: The homogeneity attack and the background knowledge attack. The limitations of the k anonymity model stem from the two assumptions (Wong *et al*., 2006). First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k* anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods.

1.4. The Perturbation Approach

Agrawal and Aggarwal (2000) develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton (2002) and Rizvi and Haritsa (2002) develop methods for privacy-preserving association rule mining.

In the perturbation approach, the distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations.

1.5. Cryptographic Techniques

Another branch of privacy preserving data mining which uses cryptographic techniques was developed. This branch became hugely popular (Laur *et al*., 2006) for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it.

Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work (Wang *et al*., 2005) has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short in providing a complete answer to the problem of privacy preserving data mining.

1.6. Randomized Response Techniques

The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. Although information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items.

Randomized Response (RR) techniques were developed in the statistics community for the purpose of protecting the privacy.

Randomized Response technique was first introduced by Warner as a technique to solve the following survey problem: To estimate the percentage of people in a population that has attribute A, ueries are sent to a group of people (Polat and Du, 2005). Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models: Related-Question Model and Unrelated-Question Model have been proposed to solve this survey problem. In the Related-Question Model, instead of asking each respondent whether he/she has attribute A, the interviewer asks each respondent two related questions, the answers which are opposite to each other (Meregu and Ghosh, 2003).

1.7. The Condensation Approach

In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way ensure high privacy protection (Verykios *et al*., 2004). It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

Perturbation, which is done by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise):

- *Blocking*, which is the replacement of an existing attribute value with a "?"
- *Aggregation* or merging which is the combination of several values into a coarser category
- *Swapping* that refers to interchanging values of individual records and
- *Sampling* which refers to releasing data for only a sample of a population

Liu and Xu (2009) illustrated this issue conveniently, but for three numerical attributes: Age, Weight, Heart rate.

Raju *et al*. (2009) used multiply protocol based homomorphic encryption along with the existing concept of digital envelope technique to achieve collaborative data mining without sharing the private data among the collaborative parties.

1.8. Secured Disclosure Approach

This method has a number of advantages over the above models in terms of disclosing the sensitive attributes in an effective way.

2. MATERIALS AND METHODS

2.1. Cluster Analysis

Clustering is an important data mining problem. The goal of clustering, in general, is to discover dense and sparse regions in a dataset. Most previous work in clustering focused on numerical data whose inherent geometric properties (Verykios *et al*., 2004) can be exploited to naturally define distance functions between points. However, many datasets also consist of categorical attributes on which distance functions are not naturally defined. Recently, the problem of clustering categorical data started receiving interest.

2.2. Categorical Variables

Categorical variable (nominal variable) is a variable which can take more than two states and the domain of the categorical attribute is small. For example, marital status is a categorical variable that may have, say three states: Single, married, divorcee.

Let the number of states of the variable be M. The states can be denoted by letters or symbols. The dissimilarity between two objects i and j, defined by nominal variables can be computed using the simple matching approach in Equation 1 as:

$$d(i,j) = (p - m) / p \tag{1}$$

where, m is the number of matches (i.e., the number of variables for which i and j are in the same state) and p is the total number of variables.

2.3. Binary Variables

A binary variable has only two states: 0 or 1, where 0 means that the variable is absent and 1 means that it is present. A binary variable is symmetric if both of its states are equally valuable and carry the same weight; that is there is no preference on which outcome should be coded as 0 or 1. One such example could be the attribute gender having the states male and female. A binary variable is asymmetric if the outcomes of the states are not equally important, such as the positive and negative outcomes of a disease test. By convention, we shall code the most important outcome, which is usually the rarest one, by 1(e.g., HIV positive) and the other by 0(e.g., HIV negative).

2.4. Proposed System

The proposed system consists of two steps. In the first step, the categorical attribute is converted into binary attribute. Then, in the second step the geometric data transformation technique is used to transform the binary data and any of the conventional clustering algorithms like k-means can be used for clustering. This data transformation technique ensures the privacy of the original data.

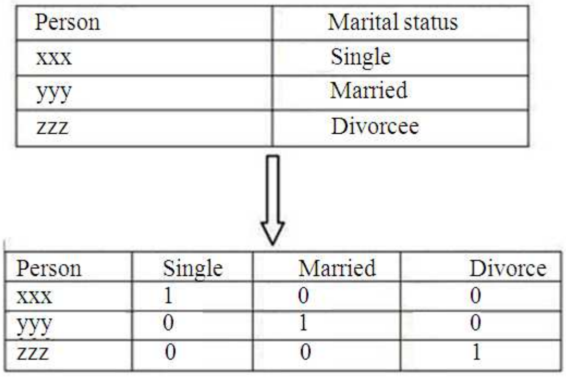
2.5. Converting Categorical Value into Binary Value

The Geometric data transformation methods cannot be applied for the categorical value. So, it is to be converted into binary value. Categorical variable can be converted into asymmetric binary variable by creating a new binary variable for each of the M nominal states. For an object with a given state value, the binary variable representing that state is set to 1 while the remaining binary variable are set to 0.

For example, to encode the nominal variable marital status, a binary variable can be created for each of the three values listed in **Fig. 1**. For a person having the marital status "*married*", the married variable is set to 1, while the remaining two variables are set to 0.

2.6. The Basics of Data Perturbation

In its simplest form, fixed-data perturbation methods involve perturbing a confidential attribute X by adding some noise term e to result in the perturbed attribute Y. When this method is used for multi attribute databases, each attribute in the database is perturbed independently of the others. In general, this method is described as Y = X + e, where e is drawn from some probability distribution (e.g., Uniform, Normal) with mean 0 and a known variance to the data. These methods are referred to as Additive Data Perturbation (ADP). Apart from ADP methods, Multiplicative Data Perturbation (MDP) can also be used to provide aggregate statistics, while protecting the privacy of individuals represented in a database. In such a method, for a single confidential attribute *X*, the perturbed attribute *Y* is described as *Y* = *Xe*, where *e* has a mean of 1.0 and a specified variance. Since the mean of *e* = 1.0, there is no bias in estimating the mean. When the MDP method is used to distort multiple confidential attributes, each attribute must be perturbed independently of other attributes.



Person	Marital status
xxx	Single
yyy	Married
zzz	Divorcee

Person	Single	Married	Divorce
xxx	1	0	0
yyy	0	1	0
zzz	0	0	1

Fig. 1. Mapping Categorical to Binary value

The Hybrid Data Perturbation Method, denoted by HDP, combines the strength of the previous methods (Oliveira and Zaïane, 2006): TDP, SDP and RDP. In this scheme, one operation is selected randomly for each confidential attribute that can take the values {Add, Mult, Rotate} in the set of operations *Di* (*OP*). Thus, each confidential attribute is perturbed using an additive, a multiplicative noise term, or a rotation.

This Hybrid data transformation approach can be used to transform the binary data to preserve the privacy of the original data. Conventional clustering algorithms like k-means can then be applied on the transformed data and again clustering results are analysed.

2.7. Measurements: Measuring Effectiveness

The effectiveness is measured in terms of the number of legitimate points grouped in the original and the distorted databases. After transforming the data, the clusters in the original databases should be equal to those ones in the distorted database. However, this is not always the case and there are some potential problems after data transformation: Either a noise data point end-up clustered, a point from a cluster becomes a noise point, or a point from a cluster migrates to a different cluster. Misclassification Error is measured in terms of the percentage of legitimate data points that are not well-classified in the distorted database. Ideally, the misclassification error should be 0%. The misclassification error, denoted by ME, is measured as denoted by Equation 2 as:

$$M_E = 1/N * \sum(|\text{Cluster}_i(D)| - |\text{Cluster}_i(D)|) \qquad (2)$$

2.8. Quantifying Privacy

Traditionally, the privacy provided by a perturbation technique has been measured as the variance between the actual and the perturbed values (Muralidhar and Sarathy, 2003). This measure is given by Var (X-Y) where X represents a single original attribute and Y the distorted attribute. This measure can be made scale invariant with respect to the variance of X by expressing security by Equation 3 as:

$$\text{Sec} = \text{V ar}(X - Y)\,\text{Var}(X) \qquad (3)$$

Clearly, the above measure to quantify privacy is based on how closely the original values of a modified attribute can be estimated.

The proposed system was coded in VB.Net using the clustering basics of WEKA tool. The various experimental results for 3000 records are shown in **Table 1-4** and **Fig. 2-5** which conclude that the results are likely to be fine at cluster levels less than 6.

3. RESULTS

- Data set taken: Census data set
- Sensitive item: Tax filer status, marital status
- Data mining task examined: Clustering
- Security methods: Additive, Multiplicative, Translation and Rotation Noises

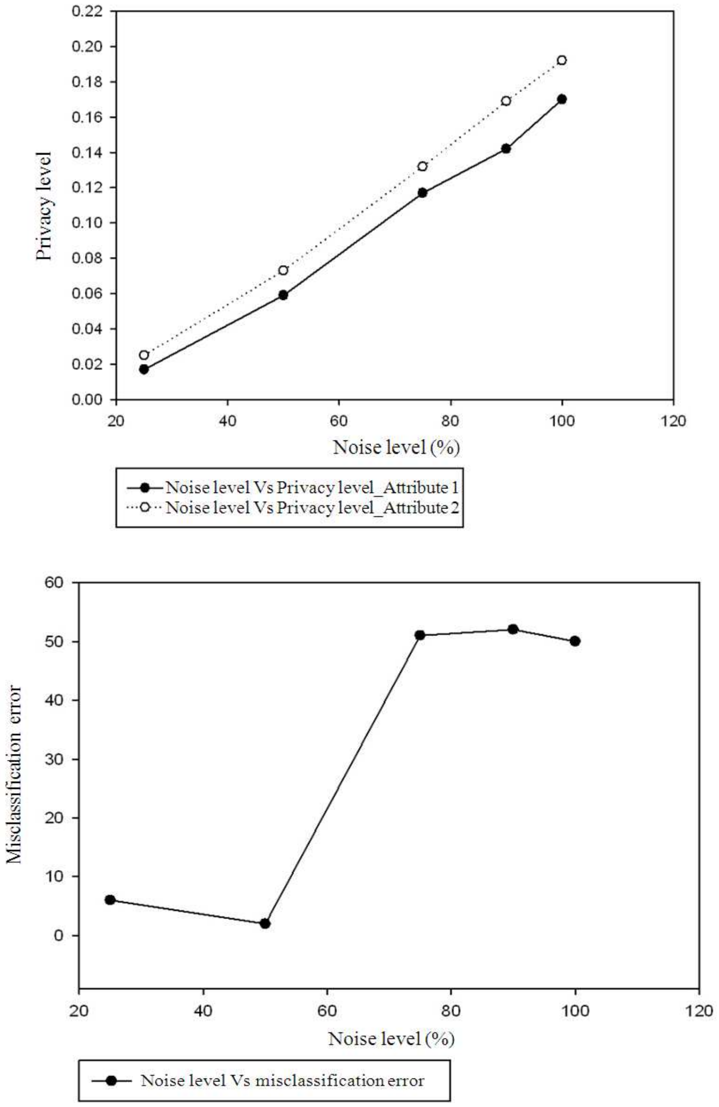


Fig. 2. Graphical results: Additive secutiy method

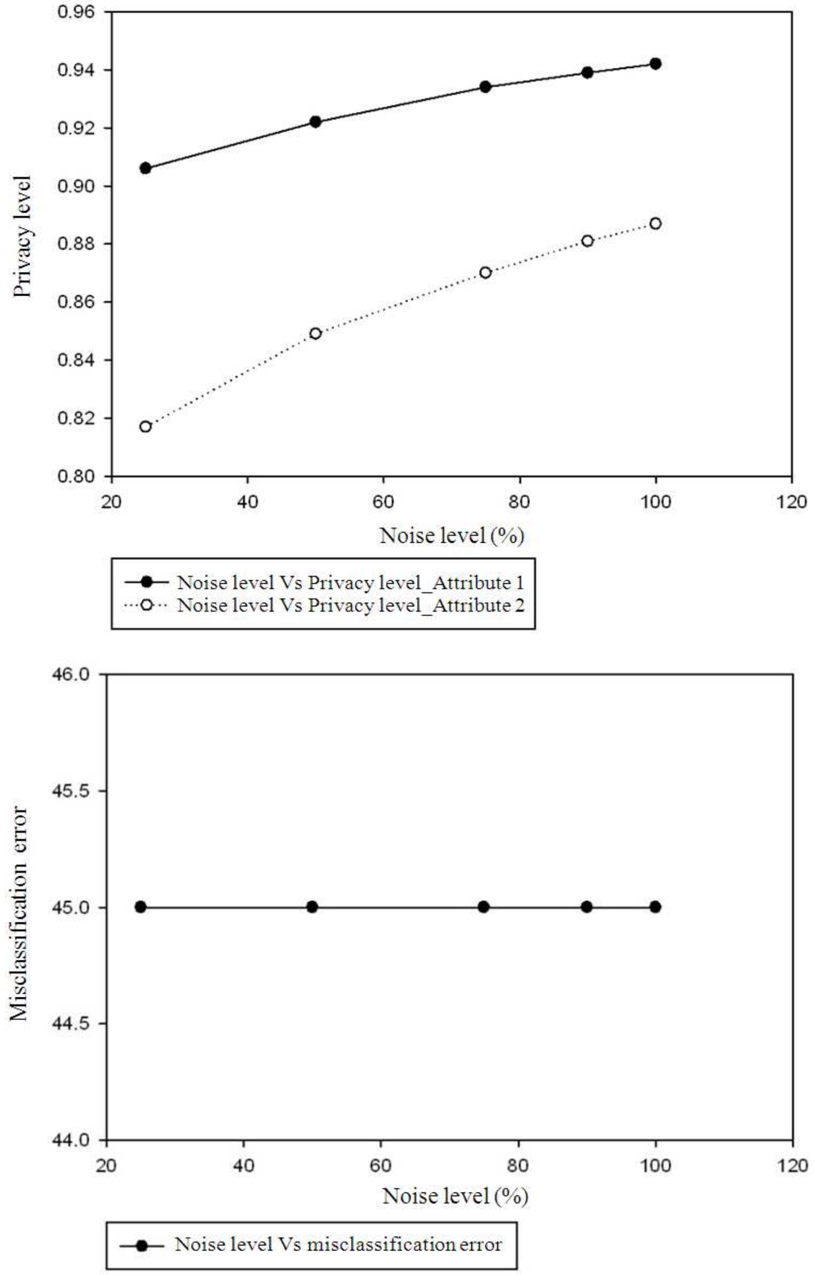


Fig. 3. Graphical results: Scaling security method

This evaluation is performed in two phases. In the first phase, the data mining task-clustering is performed without securing the sensitive details. In the second phase, the same data mining task-clustering is performed by securing the sensitive attributes.

For Census dataset, the data quality of the perturbed datasets is then compared with the data quality of the original dataset for estimating the effectiveness of secured disclosure in preserving the patterns.

The proposed secured data disclosure system was evaluated with census data set, for each type of the secured disclosure method (Additive, Scaling, Rotation and Hybrid) at various privacy levels (0, 0.25, 0.50, 0.75 and 1) and the misclassification errors levels are noted.

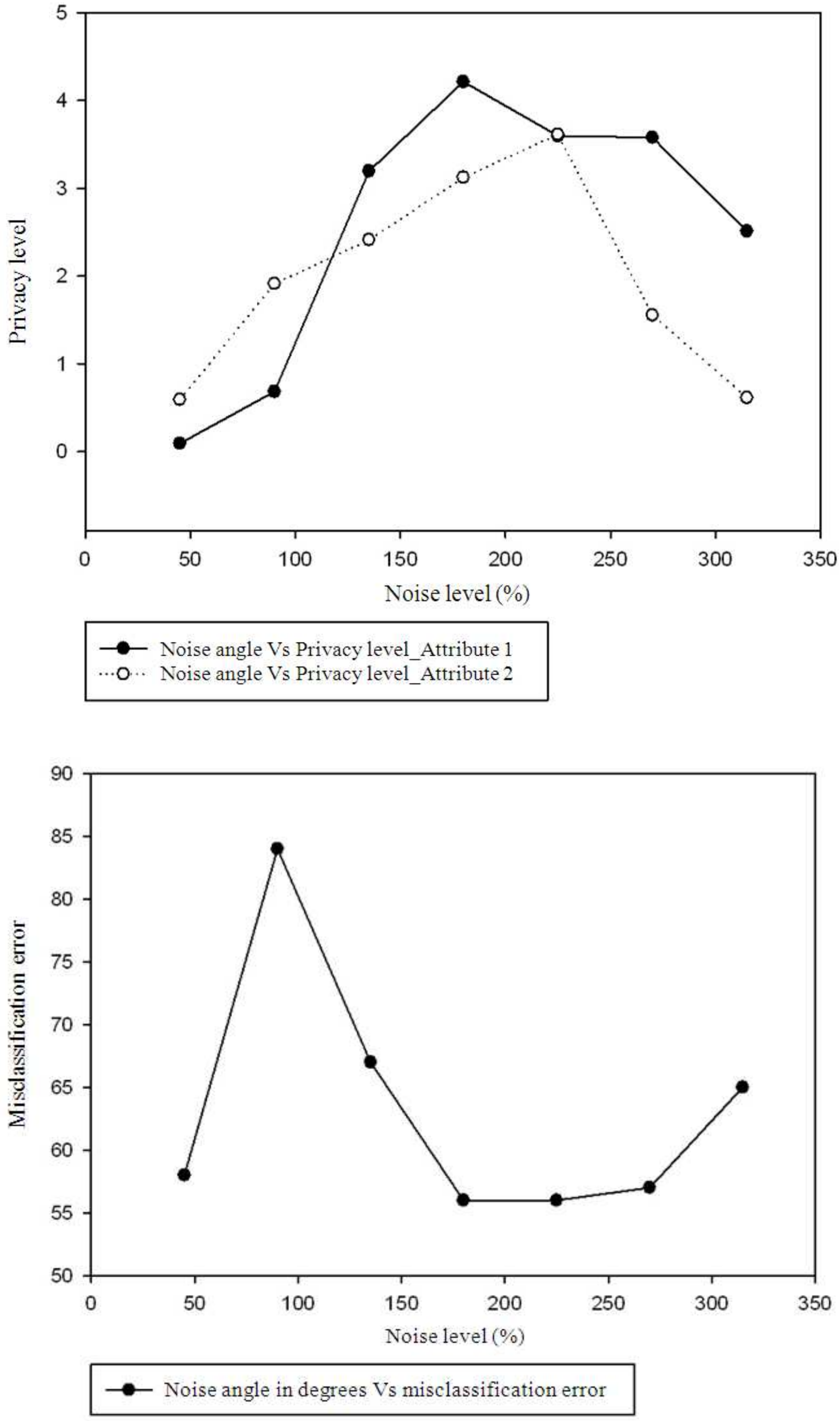


Fig. 4. Graphical results: Rotation security method

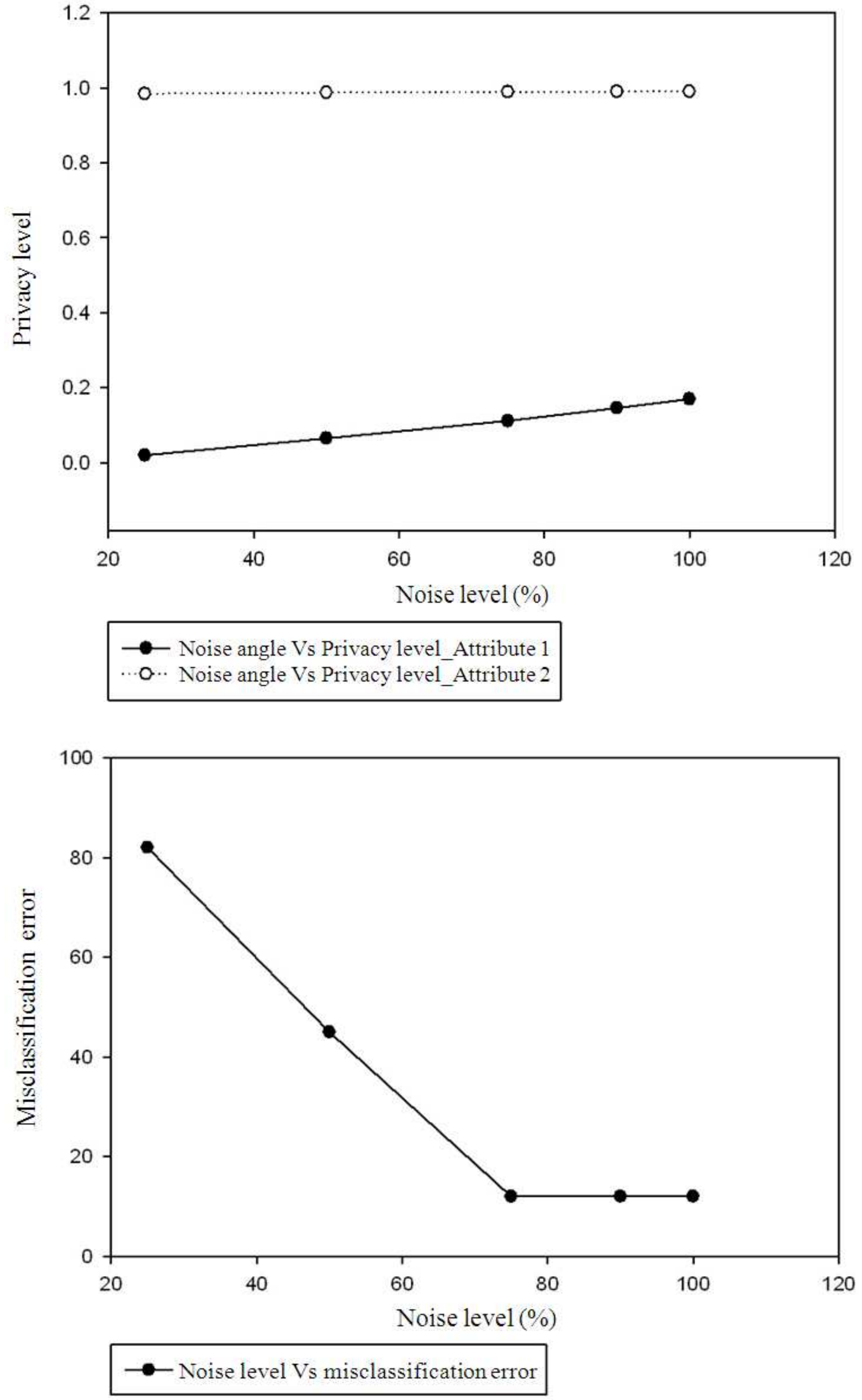


Fig. 5. Graphical results: Hybrid security method

Table 1. Tabular results: Additive security method

Noise %	Mis. Err %	PL SA1	PL SA2
25	6	0.0170	0.025
50	2	0.0590	0.073
75	51	0.1170	0.132
90	52	0.1420	0.169
100	50	0.1700	0.192

Table 2. Tabular results: Scaling security method

Noise %	Mis. Err %	PL SA1	PL SA2
25	45	0.906	0.817
50	45	0.922	0.849
75	45	0.934	0.870
90	45	0.939	0.881
100	45	0.942	0.887

Table 3. Tabular results: Rotation security method

Noise angle	Mis. Err %	PL SA1	PL SA2
45	58	0.091	0.592
90	84	0.682	1.914
135	67	3.197	2.413
180	56	4.214	3.124
225	56	3.599	3.613
270	57	3.579	1.553
315	65	2.512	0.613

Table 4. Tabular results: Hybrid noise

Noise %	Mis. Err %	PL SA1	PL SA2
25	82	0.020	0.984
50	45	0.065	0.987
75	12	0.112	0.989
90	12	0.146	0.990
100	12	0.170	0.990

3.1. Results and inference: Additive Security

For Census dataset, the proposed secured data disclosure system achieved a moderate privacy of 0.2 i.e., 20% was observed, for additive perturbation security level between 75%-100%, having misclassified records around 50% as in **Fig. 2** and **Table 1**.

3.2. Scaling Security

For Census dataset, the proposed secured data disclosure system achieved a good privacy of 0.9 i.e., 90% was observed, for scaling perturbation security level between 25%-100%, having misclassified records around 45% as in **Fig. 3** and **Table 2**.

3.3. Rotation Security

For Census dataset, the proposed secured data disclosure system achieved a moderate privacy of 4 i.e., 40% was observed, for rotation perturbation security levels between 180-270 degree, having misclassified records around 55% as in **Fig. 4** and **Table 3**.

3.4. Hybrid Security

For Census dataset, the proposed secured data disclosure system achieved a good privacy of 0.9 i.e., 90% was observed, for hybrid perturbation security levels between 75-100%, having misclassified records around 10% as in **Fig. 5** and **Table 4**.

3.5. Testing And Evaluation

The same dataset was tested with our own tool developed in VB.Net using the clustering basics of WEKA and also in WEKA. Experiments show that this method can greatly improve the privacy quality without sacrificing accuracy. Unlike the existing value randomization methods, where multiple columns are perturbed separately, this needs to perturb all columns together, where the privacy quality of all columns is correlated under one single transformation.

This method quantifies a privacy level on an average greater than 50% for which the misclassification error level is below 50% on an average. The values vary for different datasets and also for different runs as the random noise level also varies. But the noise level is kept within a range for quantifying the results.

There exists a growing body of literature on securing sensitive data mining. These algorithms can be divided into several different groups. One approach adopts a distributed framework. This approach supports computation of data mining models and extraction of “patterns” at a given node by exchanging only the minimal necessary information among the participating nodes without transmitting the raw data. Securing association rule mining from homogeneous and heterogeneous distributed data sets are a few examples. The second approach is based on data-swapping which works by swapping data values within same feature.

4. DISCUSSION

Our approach which works by adding random noise to the data in such a way that the data values are distorted, preserving the underlying distribution properties at a macroscopic level. The algorithms belonging to this group works by first perturbing the data using randomized techniques. The perturbed data is then used to extract the patterns and models.

In this approach, the distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations.

Perturbation techniques are often evaluated with two basic metrics, loss of privacy and loss of information. An ideal data perturbation algorithm should aim at minimizing both privacy loss and information loss. However, the two metrics are not well-balanced in many existing perturbation techniques (Agrawal and Srikant, 2000; Evfimievski *et al*., 2003; Aggarwal and Yu, 2004).

Loss of privacy and loss of information/accuracy are treated as two conflict factors in privacy preserving data classification. In this study, we propose a GDTM based perturbation technique that guarantees zero loss of accuracy, where the optimality is measured by a new multi-column privacy metric.

5. CONCLUSION

The family of Geometric Data Transformation Methods (GDTMs-Additive, Scaling, Rotation) ensures privacy preservation in clustering analysis, notably on categorical data. The proposed method distorts only confidential categorical attributes to meet privacy requirements, while preserving general features for clustering analysis. To best knowledge this is the first effort toward a building block solution for the problem of privacy preserving data clustering. This work can be summarized as follows: First, GDTMs are introduced and validated. The performance evaluation experiments demonstrated that the methods are effective and provide practically acceptable values for balancing privacy and accuracy.

5.1. Limitations

The transformed database is available for secondary use and must hold the following restrictions: (a) the distorted database must preserve the main features of the clusters mined from the original database; (b) an appropriate balance between clustering accuracy and privacy must be guaranteed. The results of the investigation clearly indicate that the methods achieved reasonable results and are promising.

5.2. Future Work

This study can be extended in two directions: (a) to investigate the impact of GDTMs on other clustering approaches; (b) designing new methods for privacy preserving clustering when considering the analysis of confidential categorical attributes, which requires further exploration.

6. REFERENCES

1. Ackerman, M.S., L.F. C and J. Reagle, 1999. Privacy in e-commerce: Examining user scenarios and privacy preferences. Proceedings of the 1st ACM Conference on Electronic Commerce, Nov. 03-05, ACM Press, USA., pp: 1-8. DOI: 10.1145/336992.336995
2. Aggarwal, C.C. and P.S. Yu, 2004. A condensation approach to privacy preserving data mining. EDBT.
3. Agrawal, D. and C.C. Aggarwal, 2002. On the design and quantification of privacy preserving data mining algorithms. Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, (PDS' 02), ACM Press, USA., pp: 247-255. DOI: 10.1145/375551.375602
4. Agrawal, R. and R. Srikant, 2000. Privacy-preserving data mining. J. Sigmod Record, 29: 439-450. DOI: 10.1145/342009.335438
5. Chawla, S., C. Dwork, F. McSherry, A. Smith and H. Wee, 2005. Toward privacy in public databases. Theory Cryptography, 3378: 363-385. DOI: 10.1007/978-3-540-30576-7_20
6. Chen, K. and L. Liu, 2005. Privacy preserving data classification with rotation perturbation. Proceedings of the 5th International Conference on Data Mining, Nov. 27-30, IEEE Xplore Press, pp: 589-592. DOI: 10.1109/ICDM.2005.121
7. Du, W., S. Chen and Y.S. Han, 2004. Privacy-preserving multivariate statistical analysis: Linear regression and classification. Proceedings of the 4th SIAM International Conference on Data Mining, (DM' 04), CiteSeerX, pp: 222-233. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.423
8. Evfimievski, A., J. Gehrke and R. Srikant, 2003. Limiting privacy breaches in privacy preserving data mining. Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Jun. 9-12, ACM Press, USA., pp: 211-222. DOI: 10.1145/773153.773174
9. Laur, S., H. Lipmaa and T. Mielikainen, 2006. Cryptographically private support vector machines. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug.

20-23, ACM Press, USA, pp: 618-624. DOI: 10.1145/1150402.1150477

10. Liu, J. and Y. Xu, 2009. Privacy preserving clustering by random response method of geometric transformation. Proceedings of the 4th International Conference on Internet Computing for Science and Engineering, Dec. 21-22, IEEE Xplore Press, Harbin, pp: 181-188. DOI: 10.1109/ICICSE.2009.31
11. Liu, K., H. Kargupta and J. Ryan, 2006. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowl. Data Eng., 18: 92-106. DOI: 10.1109/TKDE.2006.14
12. Meregu, S. and J. Ghosh, 2003. Privacy-preserving distributed clustering using generative models. Proceedings of the 3rd IEEE International Conference on Data Mining, Nov. 19-22, IEEE Computer Society, Melbourne, Florida, pp: 211-211. DOI: 10.1109/ICDM.2003.1250922
13. Muralidhar, K. and R. Sarathy, 2003. A theoretical basis for perturbation methods. Stat. Comput., 13: 329-335. DOI: 10.1023/A:1025610705286
14. Oliveira, S.R.M. and O.R. Zaiane, 2006. A unified framework for protecting sensitive association rules in business collaboration. Int. J. Bus. Intell. Data Mining, 1: 247-287. DOI: 10.1504/IJBIDM.2006.009135 http://dl.acm.org/citation.cfm?id=1356345
15. Polat, H. and W. Du, 2005. SVD-based collaborative filtering with privacy. Proceedings of the ACM Symposium on Applied Computing, Mar. 13-17, ACM Press, USA., pp: 791-795. DOI: 10.1145/1066677.1066860
16. Raju, R., R. Komalavalli and V. Kesavkumar, 2009. Privacy maintenance collaborative data mining-a practical approach. Proceedings of the 2nd International Conference on Emerging Trends in Engineering and Technology, Dec. 16-18, IEEE Xplore Press, Nagpur, pp: 307-311. DOI: 10.1109/ICETET.2009.184
17. Rizvi, S.J. and J.R. Haritsa, 2002. Maintaining data privacy in association rule mining. Proceedings of the 28th International Conference on Very Large Data Bases, (VLDB' 02), ACM Press, USA., pp: 682-693. http://dl.acm.org/citation.cfm?id=1287428
18. Samarati, P. and L. Sweeney, 1998. Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. The Pennsylvania State University. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5829
19. Samarati, P., 2001. Protecting respondents identities in microdata release. IEEE Trans. Knowl. Data Eng., 13: 1010-1027. DOI: 10.1109/69.971193
20. Sweeney, L., 2002. K anonymity: A model for protecting privacy. Int. J. Uncertainty, Fuzziness Knowl. Based Syst., 10: 557-570. http://arbor.ee.ntu.edu.tw/archive/ppdm/Anonymity/SweeneyKA02.pdf
21. Vaidya, J. and C. Clifton, 2002. Privacy preserving association rule mining in vertically partitioned data. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jul. 23-25, ACM Press, USA., pp: 639-644. DOI: 10.1145/775047.775142
22. Verykios, V.S., E. Bertino, I.N. Fovino, L.P. Provenza and Y. Saygin *et al*., 2004. State-of-the-art in privacy preserving data mining. ACM SIGMOD, 33: 50-57. DOI: 10.1145/974121.974131
23. Wang, K., B.C.M. Fung and P.S. Yu, 2005. Template-based privacy preservation in classification problems. Proceedings of the 5th IEEE International Conference on Data Mining, Nov. 27-30, IEEE Xplore Press, Canada. DOI: 10.1109/ICDM.2005.142
24. Wong, R.C., J. Li and A.W. Fu, 2006. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 20-23, ACM Press, USA., pp: 754-759. DOI: 10.1145/1150402.1150499