Protecting User's Information Based on Clustering Method in Data Mining

Heba Adnan Raheem

Computer Department, College of Science, University of Kerbala, Iraq

Received Date: 12/Aug/2015

Accepted Date: 18/Oct/2015

الخلاصة

الحفاظ على خصوصية تنقيب البيانات هو أحدث مجال بحوث التنقيب عن البيانات. وتعرف بأنها "حماية معلومات المستخدم". أصبحت حماية الخصوصية ذات أهمية في مجال البحوث وتنقيب البيانات بسبب زيادة القدرة على تخزين بيانات شخصية عن المستخدمين، وتطوير خوارزميات التنقيب عن البيانات للاستدلال على هذه المعلومات. الهدف الرئيس في الحفاظ على خصوصية تنقيب البيانات هو تطوير نظام لتعديل البيانات الأصلية بطريقة ما، بحيث أن البيانات الخاصة والمعرفة تبقى سرية حتى بعد انتهاء عملية التعدين.في هذا البحث اقترحنا نظاما يستخدم خوارزمية التجمع PAM في مجموعات بيانات طبية لغرض توليد مجموعة من العناقيد ، ثم أقترحنا حماية المعلومات الحساسة في كل كتله لغرض زيادة سرية معلومات المستخدمين.أن حماية المعلومات الحساسه تتم باستعمال تقنيات السرية ومن خلال تعديل قيم البيانات (الصفات) في قاعدة البيانات. ثم أقترحنا أستخدام تقنيات البعثرة العشوائية نسخ البيانات (وهي طريقة جديدة مقترحة في هذا العمل) لمنع المهاجمين من أستنتاج معلومات الأفراد. بعد التعديل نفس خوارزمية التجمع تطبق على قاعدة البيانات المحدثة للتحقق من أن المعلومات الحساسة مخفية أم لا. النتائج التجريبية على هذه التقنيات المقترحة أثبتت أن الخوارزمية PAM فعالة للتجميع في جميع مجموعات البيانات وأن الكتلة المحددة تم حمايتها بكفاءة باستخدام تقنيات (نسخ البيانات). هذه التقنيات تم تطبيقها على بيانات سرطان الثدي، مجموعة بيانات السكري. أخيرا نتائج النظام المقترح أثبتت أن تشويه البيانات يمكن أن يخفض عندما نسبة الخصوصية تزداد. هذه القضايا مهمه في عملية حفظ الخصوصية (السرية) في تعدين البيانات، لذا فأن النظام المقترح ناجح جدا في تحقيق حماية السرية.

الكلمات المفتاحية

التكتل، خوارزميةالتجمع، الخصوصية، نسخ البيانات.

ABSTRACT

Privacy preserving data mining is a latest research area in the field of data mining. It is defined as "protecting user's information". Protection of privacy has become important in data mining research because of the increasing ability to store personal data about users and the development of data mining algorithms to infer this information. The main goal in privacy preserving data mining is to develop a system for modifying the original data in some way, so that the private data and knowledge remain private even after the mining process. In this paper we propos a system that used *PAM (partitioning around medoid)* clustering algorithm in health datasets in order to generate set of clusters, then we suggest protecting the sensitive attributes in each cluster in order to increasing the privacy of users information. Protecting the sensitive attributes is done by using privacy techniques through modifying the data values (attributes) in the dataset. We suggest using randomization technique

Data copying (which is a new suggested technique in this paper) to prevent attacker from concluding users privacy information. After modification, the same clustering algorithm is applied to modified data set to verify whether the sensitive attributes are hidden or not. Experimental results on these proposed techniques prove that the PAM algorithm is efficient for clustering in all data sets and the selected clusters are protected efficiently by using **Data Copying** technique. This technique is applied to Wisconsin breast cancer and diabetes data set. Finally the results of the proposed system prove that the distortion of data can be reduced when the privacy ratio was increased. These are important issues in PPDM, therefore the proposed system is highly successful in achieving the protection of privacy.

Keywords

Clustering, PAM, Privacy, Data Copying

1. INTRODUCTION

"Data mining is the extraction of hidden predictive information from large databases and also a powerful new technology with great potential to analyze important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Every user need to collect and use the tremendous amounts of information is growing in a very large manner. Initially, with the advent of computers and means for mass digital storage, users has started collecting and storing all sorts of data, counting on the power of computers to help sort through this combination of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial confusion has led to the creation of structured databases and database management systems"[1]. Today users can handle more information from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Privacy is defined as "protecting individual's information". Protection of privacy has become an important issue in data mining research. A standard dictionary definition of privacy as it pertains to data is "freedom from unauthorized intrusion"[2]. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information [2]. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and knowledge remain private even after the mining process [2]. A number of techniques such as randomization and k-anonymity have been suggested in recent years in order to perform privacy-preserving data mining [3].

2. RELATED WORKS

In [3] proposed a novel clustering method for conducting the k-anonymity model effectively. The similarity between this method and our proposal method is in the reducing of the information distortion. The difference is in clustering algorithm that is used and privacy technique.

In [4] discussed a condensation approach for data mining. This approach uses a methodology which condenses the data into multiple groups of predefined size. For each group, certain statistics are maintained. A greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. They use the statistics from each group in order to generate the corresponding pseudo-data. The results show that our proposal method is the best in reducing amount of information loss.

In [1] they have used four clustering algorithms (PAM, CLARA, CLARANS and ECLARANS) to detect outliers and also proposed a new privacy technique GAUSSIAN PERTURBATION RANDOM METHOD to protect the sensitive outliers in health data sets. The similarity between this method and our proposal method is in using the PAM clustering algorithm and some of data set that is used. The difference is in the privacy technique.

In [5] presented a framework for adding noise to all attributes (both numerical and categorical). the similarity between this method and our proposal method is in using the additive noise privacy technique, but we applied it only to the selected cluster and only to the numerical attributes because of the nature of datasets that is used which is numerical.

3. The Main Objective AND METHODOLOGY

The main objective of this research work is, applying the privacy preserving data mining by using clustering algorithm. *Protecting the sensitive attributes in each cluster by using a privacy techniques (Data Copying) in the form of modifying the data items in the dataset.* After modification the same clustering algorithm is applied for modified data set. Now, verify whether the clusters are hided or not. The performance of the clustering algorithm and the privacy technique are analyzed. The system Architecture is summarized in Fig. (1):

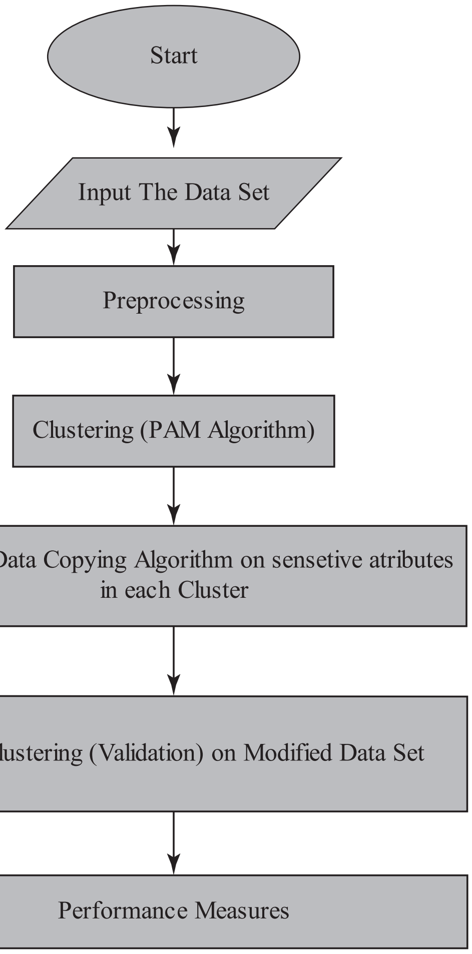

Start
Input The Data Set
Preprocessing
Clustering (PAM Algorithm)
Applying Data Copying Algorithm on sensetive atributes in each Cluster
PAM Clustering (Validation) on Modified Data Set
Performance Measures

Fig. (1): A Flow Chart of the Proposed System

3. 1. Dataset as Input

Breast Cancer Wisconsin, Diabetes and heart stat log data sets are used for clustering and cluster selected protection. These datasets are collected from http://archive.ics.uci.edu/ml/datasets.html.

3.1. 1. Breast Cancer Wisconsin Dataset

This dataset consists of 699 instances and 10 attribute. The dataset characteristics are multivariate. The attribute characteristics are integer.

3.1. 2. Diabetes Data Set

This dataset consists of 768 instances and 9 attributes. The dataset characteristics are multivariate.The attribute characteristics are real.

3.1. 3. heart stat log Data Set

This dataset consists of 270 instances and 14 attributes. The dataset characteristics are multivariate.The attribute characteristics are real.

3. 4. Pre-Processing

The dataset is modified by dealing with the missing values.To do so it is replaced with more repeating value of that attribute over the whole dataset.

3. 5. An Approach for clustering

The following clustering algorithms are used in our research:

3.3. 1. PAM (Partitioning Around Medoid)

PAM uses a *k*-medoid method for clustering. It is very robust when compared with k-means in the presence of noise and outliers. The most common realisation of *k*-medoid clustering is the Partitioning Around Medoids (PAM) algorithm and is as follows [6]**:**

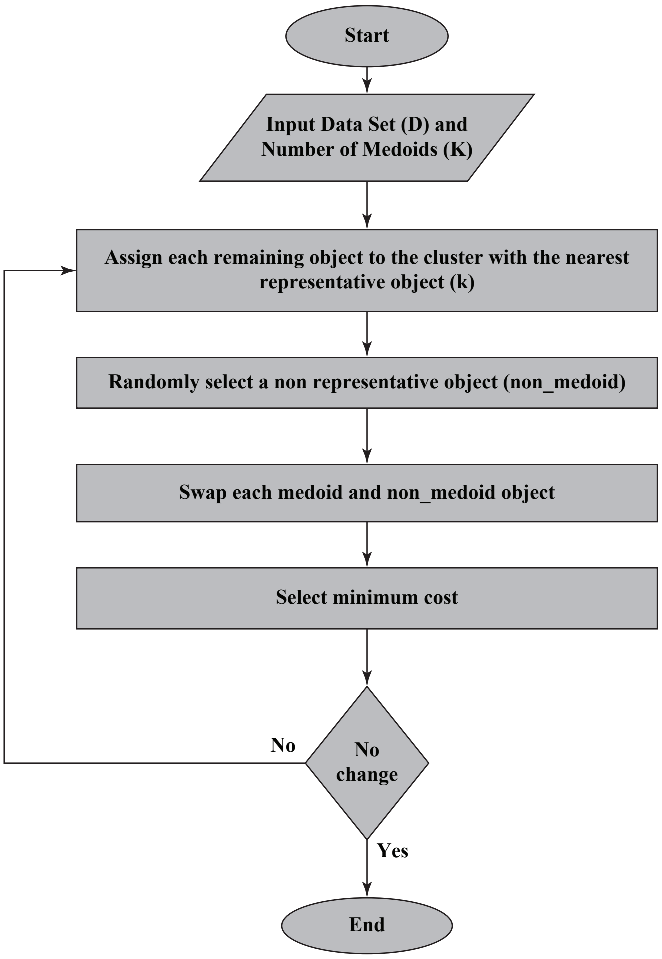

Start
Input Data Set (D) and Number of Medoids (K)
Assign each remaining object to the cluster with the nearest representative object (k)
Randomly select a non representative object (non_medoid)
Swap each medoid and non_medoid object
Select minimum cost
No change
No
Yes
End

Fig. (2): A Flow Chart of Partitioning Around Medoids (PAM) clustering algorithm

3. 4. Applying A Privacy Preserving Data Mining Techniques (PPDM)

Our work in this research is based on copying techniques, but first the most three sensitive attributes SAR from each record in each cluster that have minimum of sum square error value (min SSE error) between the record and the medoid(representative point of the selected cluster) are found by Equation 1.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2 (m_i, x) \quad \text{....(1)}$$

x is a data point in cluster C_i and m_i is the representative point (medoid) for cluster C_i, then the PPDM technique is applied for every data set [2].

3.4. 1. Data Copying Technique

This is a new perturbative technique that is suggested in this research for protecting the sensitive numerical attributes in each cluster. It is very similar to Data Swapping technique because it is simple and can be used only on sensitive data without disturbing non sensitive data. A Flow Chart of Data Copying algorithm is summarized in Fig. (3).

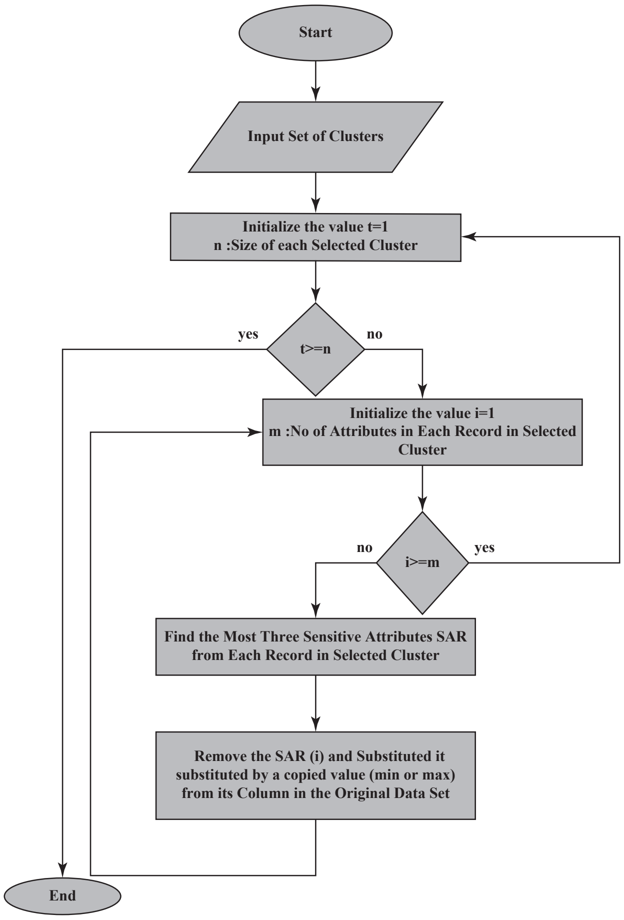

Start
Input Set of Clusters
Initialize the value t=1
n :Size of each Selected Cluster
yes
t>=n
no
Initialize the value i=1
m :No of Attributes in Each Record in Selected Cluster
no
i>=m
yes
Find the Most Three Sensitive Attributes SAR from Each Record in Selected Cluster
Remove the SAR (i) and Substituted it substituted by a copied value (min or max) from its Column in the Original Data Set
End

Fig. (3): A Flow Chart of Data Copying Algorithm

Example 1:

Consider some of the clustering results after applying PAM algorithm on Breast Cancer Wisconsin data set as shown in Table (1):

Table (1): Clustering before applying Data Copying algorithm

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	clu no
3	5	4	4	5	7	10	3	2	1	0
5	6	8	8	1	3	4	3	7	1	0
2	5	1	1	1	2	1	3	1	1	1
4	3	1	1	1	2	2	3	1	1	1
8	1	1	1	1	2	10	3	1	1	2
9	2	1	2	1	2	1	3	1	1	2

Note: *The values which have red color represent the sensitive attributes.* By applying Data Copying algorithm, the result of privacy is as shown in Table (2):

Table (2): Clustering after applying Data Copying algorithm

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	clu no
3	5	4	4	5	7	10	3	2	1	0
5	6	8	8	1	3	4	3	7	1	0
2	5	1	1	1	2	1	3	1	1	1
4	3	1	1	1	2	2	3	1	1	1
8	10	1	1	1	9	10	8	1	1	1
9	2	10	2	10	9	1	3	1	1	1

Note: *The values which have red color represent the copied values* (min or max) from attributes column in the original data set *which substituted with sensitive attributes which has been removed by Copying process*.

3. 5. Applying PAM Clustering for Validation

Notice that by applying the PPDM techniques for each one of health data set in this work, it could be able to protect the sensitive cluster information. Later the dataset is modified based on the privacy technique. Now, after modification the PAM algorithm is applied to the modified dataset in order to verify whether the cluster is hidden or not. All the sensitive attributes in the requirement cluster are protected by using this technique according to the results of the evaluation measures.

3. 6. Performance Measures.

This research work has implemented in C# language and executed in the processor Intel(R) Core (TM) 2 Duo CPU 2.00 GHZ processor and 2.GB main memory under the Windows 7 Ultimate operating system..The experimental results are analyzed based on the following performance factors.

3.6. 1. Privacy Ratio.

The privacy ratiois measured by the percentage between the number of records that remained near to the original cluster after privacy and the number of records in the original cluster before privacy **[*8*].** The privacy ratio is calculated by Equation 2 :

$$\boldsymbol{PR} = \left(\mathbf{1} - \frac{\boldsymbol{R}(\boldsymbol{c'})}{\boldsymbol{R}(\boldsymbol{c})}\right) * \mathbf{100} \quad \ldots(2)$$

3.6. 2. Information Loss Ratio

This performance factor is used to measure the percentage of distortion of the information of all data set after applying the privacy technique **[*9*].** The information loss ratio is calculated by Equation 3.

$$\boldsymbol{ILR} = \frac{\sum|\boldsymbol{original\ value} - \boldsymbol{new\ value}|}{\sum|\boldsymbol{original\ values}\ |} *100 \ \ldots.(3)$$

3.6. 3. Covering of Data Ratio

This performance factor is used to measure the percentage of average number of clusters covered in hidden cluster **[*8*].** it is calculated by Equation 4.

$$COD = \frac{C}{K} * 100 \quad \ldots.(4)$$

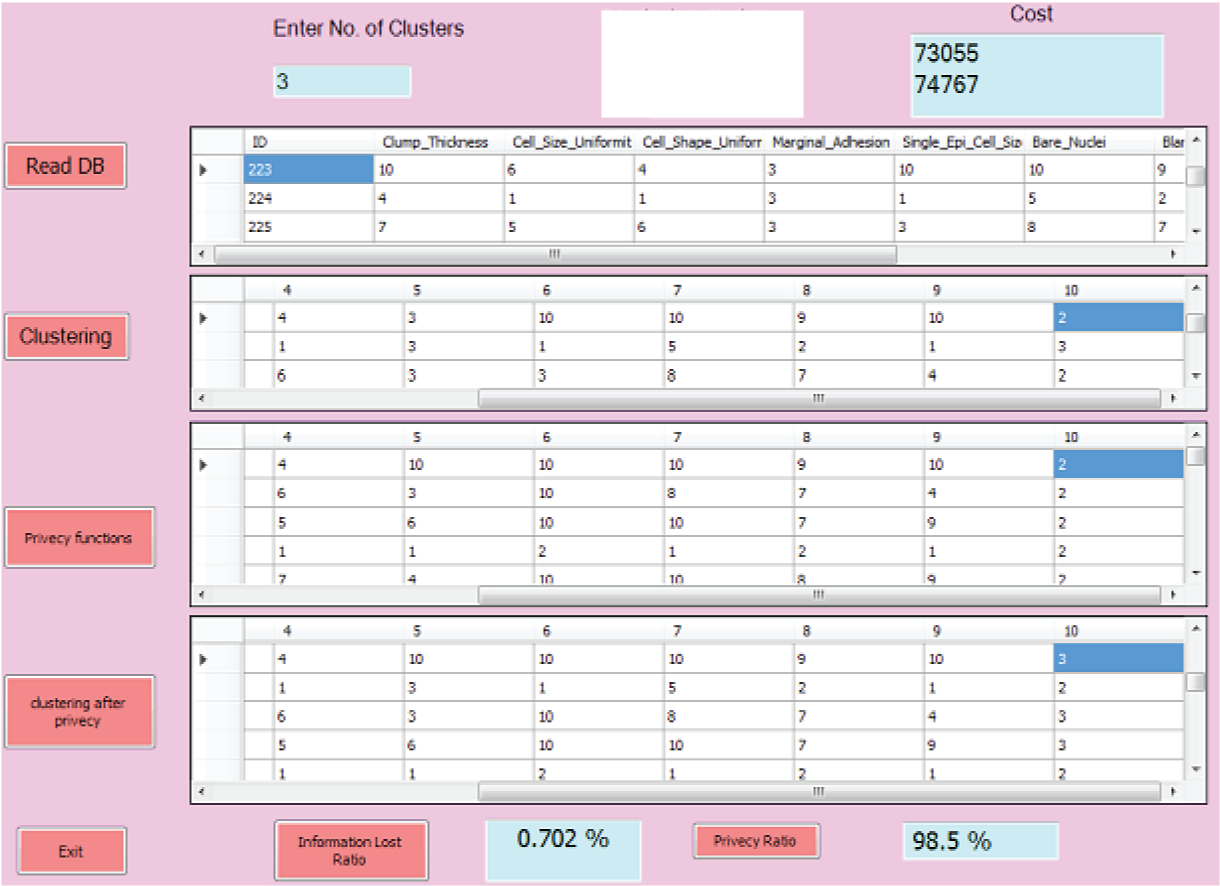
Where **C** corresponds to the number of clusters that contained the records of the selected cluster after privacy. K to the value of the original clusters number before privacy.

3.6. 4. Running Time

In this work, the efficiency (time requirements) is calculated by using the CPU time. These measures have been applied to evaluate and test the results that are obtained by applying the PPDM techniques.

3. 7. Experiments and Results for Applying PPDM Techniques and The Evaluation Measures.

The SAR in each cluster are protected according to the following results:



Enter No. of Clusters: 3

Cost: 73055 74767

ID	Clump_Thickness	Cell_Size_Uniformit	Cell_Shape_Uniforr	Marginal_Adhesion	Single_Epi_Cell_Siz	Bare_Nuclei	Blar
223	10	6	4	3	10	10	9
224	4	1	1	3	1	5	2
225	7	5	6	3	3	8	7

4	5	6	7	8	9	10
4	3	10	10	9	10	2
1	3	1	5	2	1	3
6	3	3	8	7	4	2

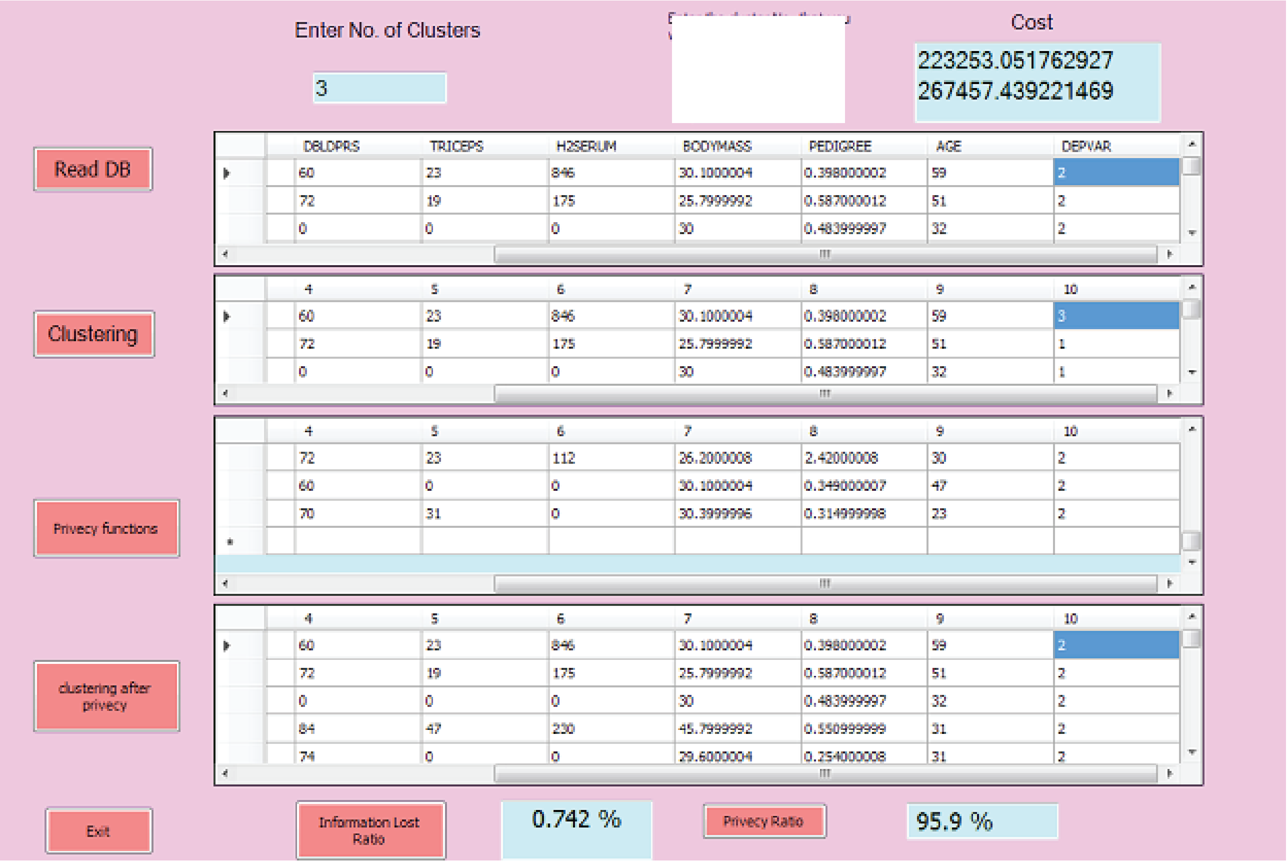
4	5	6	7	8	9	10
4	10	10	10	9	10	2
6	3	10	8	7	4	2
5	6	10	10	7	9	2
1	1	2	1	2	1	2
7	4	10	10	8	9	2

4	5	6	7	8	9	10
4	10	10	10	9	10	3
1	3	1	5	2	1	2
6	3	10	8	7	4	3
5	6	10	10	7	9	3
1	1	2	1	2	1	2

Information Lost Ratio: 0.702 %

Privecy Ratio: 98.5 %

Fig. (4): The results of running breast cancer dataset



Enter No. of Clusters: 3

Cost:
223253.051762927
267457.439221469

DBLDPRS	TRICEPS	H2SERUM	BODYMASS	PEDIGREE	AGE	DEPVAR
60	23	846	30.1000004	0.398000002	59	2
72	19	175	25.7999992	0.587000012	51	2
0	0	0	30	0.483999997	32	2

4	5	6	7	8	9	10
60	23	846	30.1000004	0.398000002	59	3
72	19	175	25.7999992	0.587000012	51	1
0	0	0	30	0.483999997	32	1

4	5	6	7	8	9	10
72	23	112	26.2000008	2.42000008	30	2
60	0	0	30.1000004	0.349000007	47	2
70	31	0	30.3999996	0.314999998	23	2

4	5	6	7	8	9	10
60	23	846	30.1000004	0.398000002	59	2
72	19	175	25.7999992	0.587000012	51	2
0	0	0	30	0.483999997	32	2
84	47	230	45.7999992	0.550999999	31	2
74	0	0	29.6000004	0.254000008	31	2

Information Lost Ratio: 0.742 %

Privecy Ratio: 95.9 %

Fig. (5): The results of running diabetes dataset

4. Conclusions

In the present study, the following facts can be concluded:

1. The results showed that by implementation PAM clustering algorithm there is no empty clusters because medoids are less influenced by outliers and noise than k_means and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.

2. Distortion of data in all the data sets has been minimum values and the records of each cluster are protected with good format because the sensitive attribute is removed and substituted by a copied value (min or max) from its column in the original data set which is the farthest from it in range.

3. Experimental results show that the PAM algorithm is efficient for clustering in all data sets and each cluster is protected efficiently by the Data copying techniques in Wisconsin breast cancer and diabetes data set

5. Suggestions and Future Works

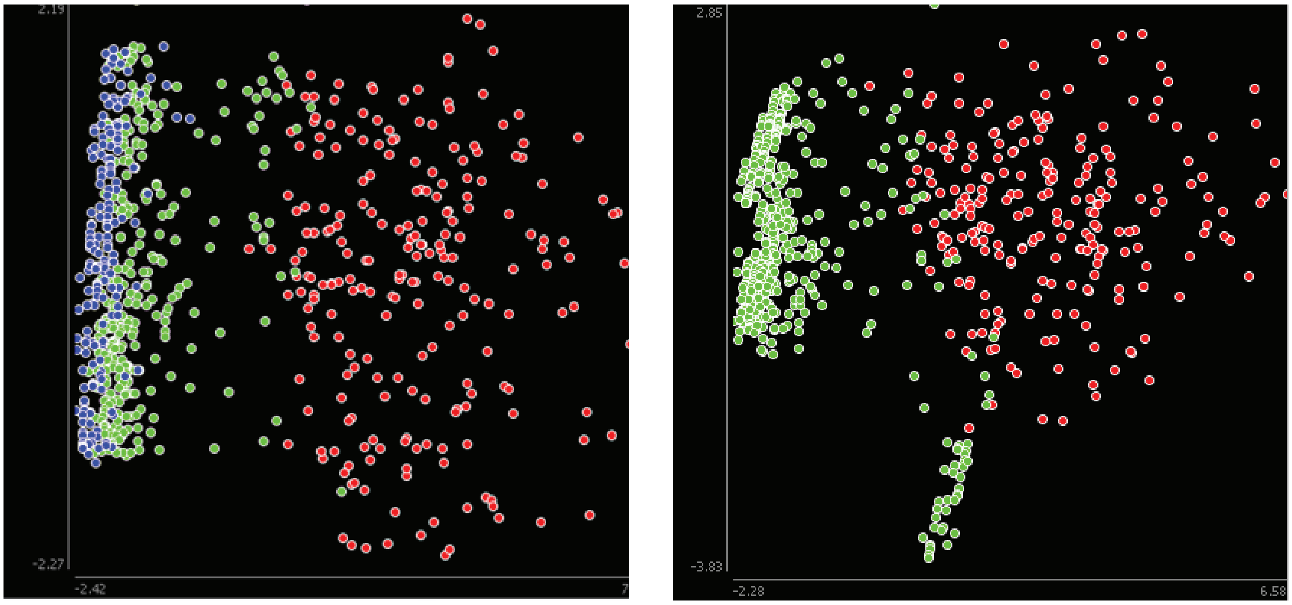
1. Developing the suggested privacy techniques by choosing the sensitive cluster number automatically instead of manually and according to specific conditions (measures).

2. Applying another clustering technique such as DBSCAN algorithm and comparing the results with our method.

3. Applying PPDM with Artificial Bee Colony (ABC) algorithm by using one of the better clustering algorithm.

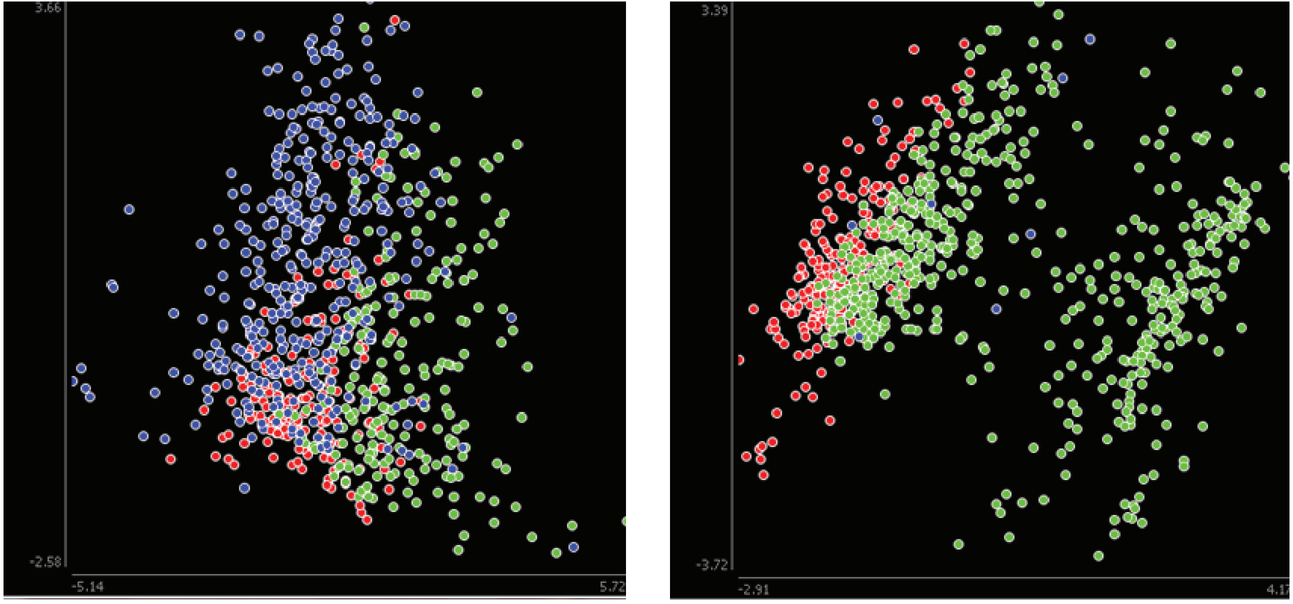
APPENDICIES

In this section the results of implementation the privacy ratio measure are presented by using Weka data mining toolset as in following':



(a) (b)

Fig. (6): The Privacy Ratio Results for Breast Cancer Wisconsin Dataset Using Copy Technique. (a) Before privacy, (b) After privacy.



(a) (b)

Fig. (7): The Privacy Ratio Results for Diabetes Data set Using Copy Technique. (a) Before privacy , (b) After privacy.

REFERENCES

[1] S. Vijayarani and S.Nithya." Sensitive Outlier Protection in Privacy Preserving Data Mining", International Journal of Computer Applications (0975-

8887), Volume 33-No. 3, November(2011).

[2] Aggarwal C. C, Yu P. S. "Models and Algorithms: Privacy-Preserving Data Mining," Springer, ISBN: 0-387-70991-8. (2008).

[3] Chuang-Cheng Chiu and Chieh-YuanTsai," A *k*-Anonymity Clustering Method for Effective Data Privacy Preservation", Springer, (Eds.): ADMA 2007, LNAI 4632, pp. 89–99, (2007).

[4] Jian Wang, Yongcheng Luo, Yan Zhao, Jiajin Le." A Survey on Privacy Preserving Data Mining". First International Workshop on Database Technology and Applications, IEEE (2009).

[5] Md Zahidul Islam and Ljiljana Brankovic." Privacy preserving data mining: A noise addition framework using a novel clustering technique". Knowledge-Based Systems 24, 1214–1223, Elsevier (2011).

[6] Margaret H. Dunham.." Data Mining, Introductory and Advanced Topics", Prentice Hall, (2002).

[7] S. Vijayarani and Dr. A. Tamilarasi." An Efficient Masking Technique for Sensitive Data Protection". IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, MIT, Anna University, Chennai. June 3-5, (2011).

[8] T. Pietraszek, "Alert Classification to Reduce False Positives in Intrusion Detection". Ph. D. thesis, Institut f¨ ur Informatik, Albert-Ludwigs- Universit¨ at Freiburg,Germany, : 1–224 (2006).

[9] S. Vijayarani and M. Sathiya Prabha." Association Rule Hiding using Artificial Bee Colony Algorithm". International Journal of Computer Applications (097-8887) Volume 33– No.2, November(2011).