Journal of Kerbala University , Vol. 12 No.3 Scientific . 2014

Privacy Preserving in Data Mining

أبقاء السريه في تعدين البيانات

Assistant Lecturer/ Heba Adnan Raheem
Master Computer Sciences
Kerbala university/College of Sciences (Computer Department)

Assistant Professor /Safaa O. Al-Mamory
college of Information Technology, Babylon University, Iraq

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 an 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 proposed system that used PAM clustering algorithm in health datasets in order to generate set of clusters, then we suggested to select only one cluster to be hidden between another clusters in order to increasing the privacy of users information .The selected cluster are considered as sensitive cluster. Protecting the sensitive cluster is done by using privacy techniques through of modifying the data values(attributes) in the dataset. We suggest to use randomization techniques (Additive Noise , Data Swapping) and Data copying (which it is new suggested technique in this thesis) to prevent attacker from concluding users privacy information in the sensitive cluster. After modification the same clustering algorithm is applied for modified data set to verify whether the selected cluster are hidden or not. Experimental results on these proposed techniques proved that the PAM algorithm is efficient for clustering in all data sets and the selected cluster are protected efficiently by using (Additive Noise , Data Swapping, Data Copying) techniques. These techniques are applied on Wisconsin breast cancer, diabetes and heart stat log data set. The privacy ratio on heart stat log data set was 48%, 52.1739 % and 31.25% in Data Copying, Additive Noise and Data Swapping techniques, respectively, because these kinds of data sets have the special property that they are *extremely sparse*. Experimental results also proved that the Data copying technique is faster than the existing techniques (swapping and noise addition), finally the results of proposed system proved that the distortion of data can be reduced when the privacy ratio was increased. These are an important issues in PPDM, therefore the proposed system is highly successful in achieving the protection of privacy.

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 الرئيسي في الحفاظ على خصوصية تنقيب البيانات هو تطوير نظام لتعديل البيانات الأصلية بطريقة ما، بحيث أن البيانات الخاصة والمعرفة تبقى سريه حتى بعد انتهاء عملية التعدين.في هذا البحث اقترحنا نظاما يستخدم خوارزمية التجمع PAM في مجموعات بيانات طبيه لغرض توليد مجموعه من العناقيد ، ثم أقترحنا أختيار عنقود واحد فقط لكي يخفى بين العناقيد الأخرى لغرض زيادة سرية معلومات المستخدمين . أن العنقود المختار يعتبر كعنقود حساس.حماية العنقود الحساس تتم باستعمال تقنيات السريه ومن خلال تعديل قيم البيانات(الصفات) في قاعدة البيانات. ثم أقترحنا أستخدام تقنيات البعثره العشوائية (الضوضاء المضافة ، نسخ البيانات) ومبادلة البيانات (وهي طريقه جديده مقترحه في هذه الأطروحه) لمنع المهاجمين من أستنتاج معلومات الأفراد االسريه في التجمع الحساس. بعد التعديل نفس خوارزمية التجمع تطبق على قاعدة البيانات المحدثه للتحقق من أن العنقود الذي

تم أختياره مخفي 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(الضوضاء المضافة ، مبادلة البيانات ، نسخ البيانات).هذه التقنيات تم تطبيقها على بيانات سرطان الثدي , مجموعة بيانات السكري وبيانات سجل معلومات القلب.نسبة السريه لبيانات سجل معلومات القلب كانت 48%, 52.1739% ,31.25% في تقنيات مبادلة البيانات, الضوضاء المضافة , نسخ البيانات , على التوالي , لأن هذه الأنواع من مجموعات البيانات لديها مواصفات خاصة حيث تمتاز بأنها متناثره للغاية . أثبتت النتائج التجريبيه أيضا أن تقنية مبادلة البيانات أسرع من التقنيات الحاليه الموجوده(التبديل وأضافة الضوضاء), أخيرا نتائج النظام المقترح أثبتت أن تشويه البيانات يمكن أن يخفض عندما نسبة الخصوصية تزداد .هذه القضايا مهمه في عملية حفظ الخصوصيه (السريه) في تعدين البيانات، لذا فأن النظام المقترح ناجح جدا في تحقيق حماية السريه.

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. 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. 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 in the reducing the information distortion. The difference 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 shows that our proposal method is 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 in the using PAM clustering algorithm and some of data set that is used. The difference 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 in the using additive noise privacy

technique, but we are applied it only on the selected cluster and only on the numerical attributes because the nature of datasets that is used which it numerical .

3.PROBLEM FORMULATION AND METHODOLOGY

The main objective of this research work is, applying the privacy preserving data mining by using clustering algorithm. The cluster selected randomly to be hided are considered as sensitive cluster. Protecting the sensitive cluster by using a privacy techniques(Additive Noise, Data Swapping, 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 cluster are hided or not. The performance of the clustering algorithm and the privacy technique are analyzed. The System Architecture is summarized in Figure .1:

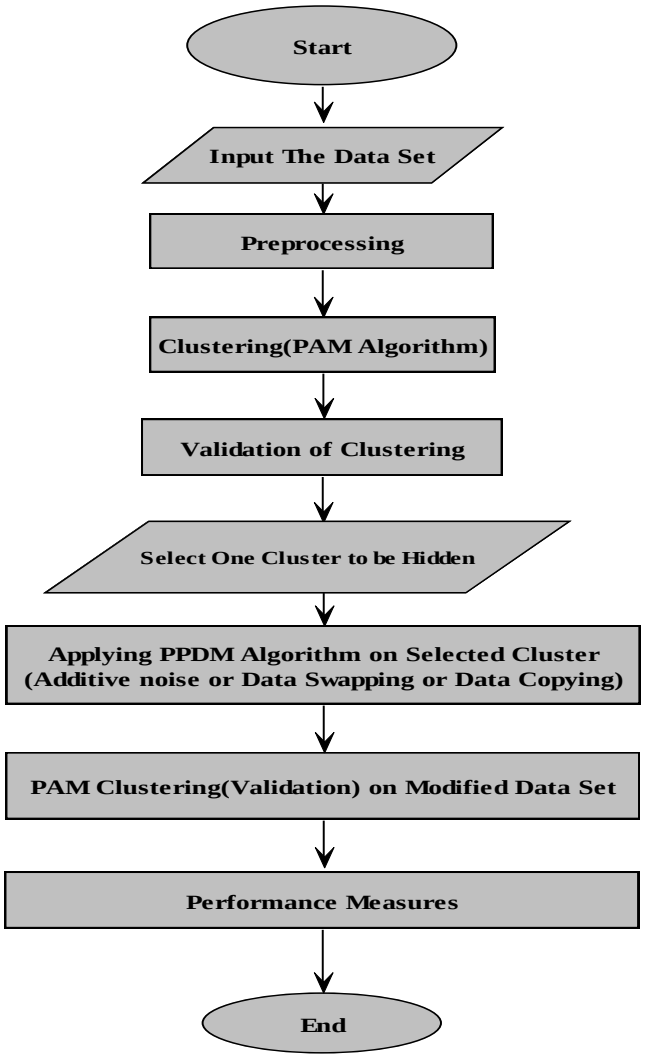


Fig .1 A Flow Chart of the Proposed System

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 attribute. 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 attribute. The dataset characteristics are Multivariate .The attribute characteristics are real.

3.2 Pre-Processing

Data cleansing is the approach of detecting and removing or correcting corrupt or inaccurate records from a record set, table or database, which is also called data scrubbing. 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.

*3.3*An Approach for clustering

The following clustering algorithm 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*]:**

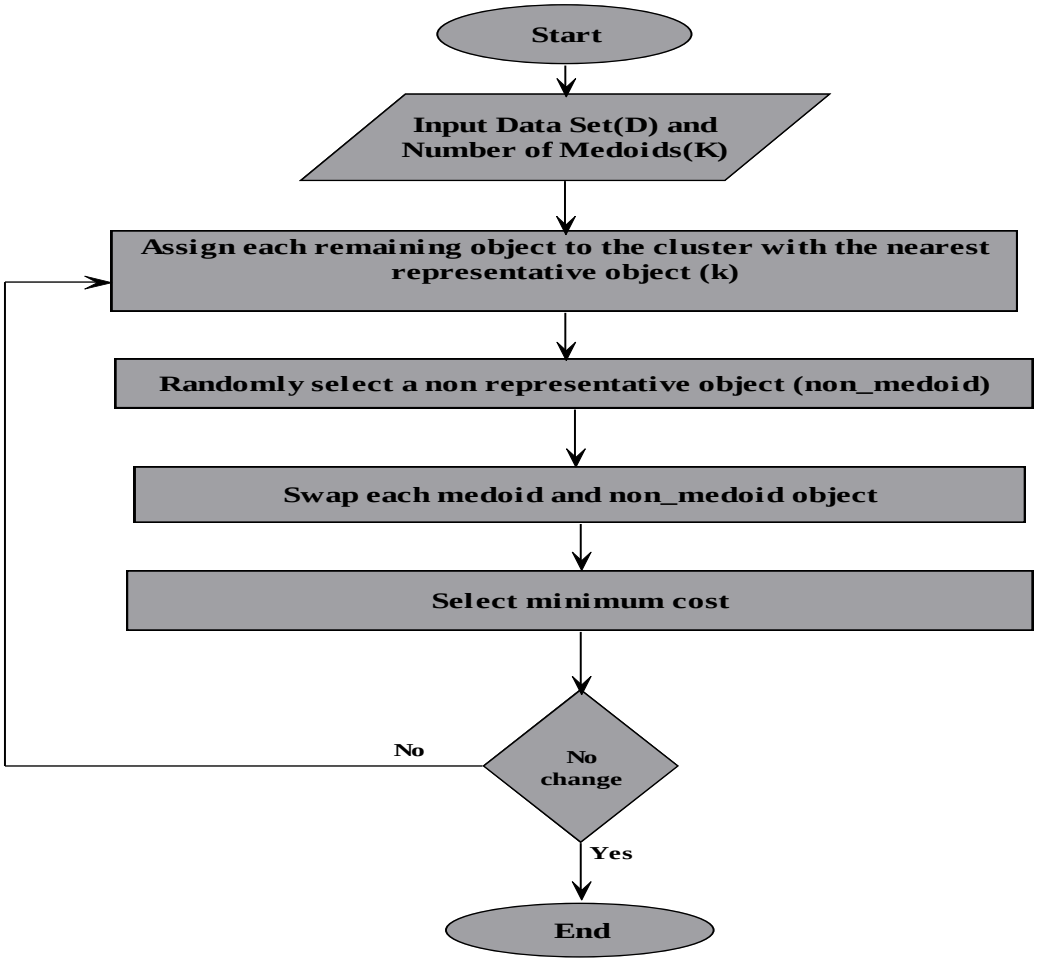


Fig. 2 A Flow Chart of PAM clustering algorithm

3.4 Validation of the Clustering Results

The optimal cluster number is determined by minimizing or maximizing the value of each index. In this research internal index validity measure(F-ratio or WB-index) is used to detect the optimal clustering number. The optimal cluster number is determined by minimizing the value of this index.

3.5 Select the Cluster to be Hidden

In this step the cluster selected to be hidden are considered as sensitive cluster. Protecting the sensitive cluster by using a privacy technique in the form of modifying the data items in the dataset .For example ,Suppose the requirement cluster is k=2.

3.6 Applying A Privacy Preserving Data Mining Techniques(PPDM) on the Selected Cluster.

Our work in this research is based on (Additive noise, Swapping, Copying) techniques, but first the most three sensitive attributes SAR from each record in the selected 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.

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

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.

3.6.1 Additive Noise Technique

The basic idea of the additive-noise-based perturbation technique is to add random noise to the actual data. The noise being added is typically continuous and with mean zero, which suits well continuous original data .A Flow Chart of Additive Noise algorithm **[7]** is summarized in Figure 3.

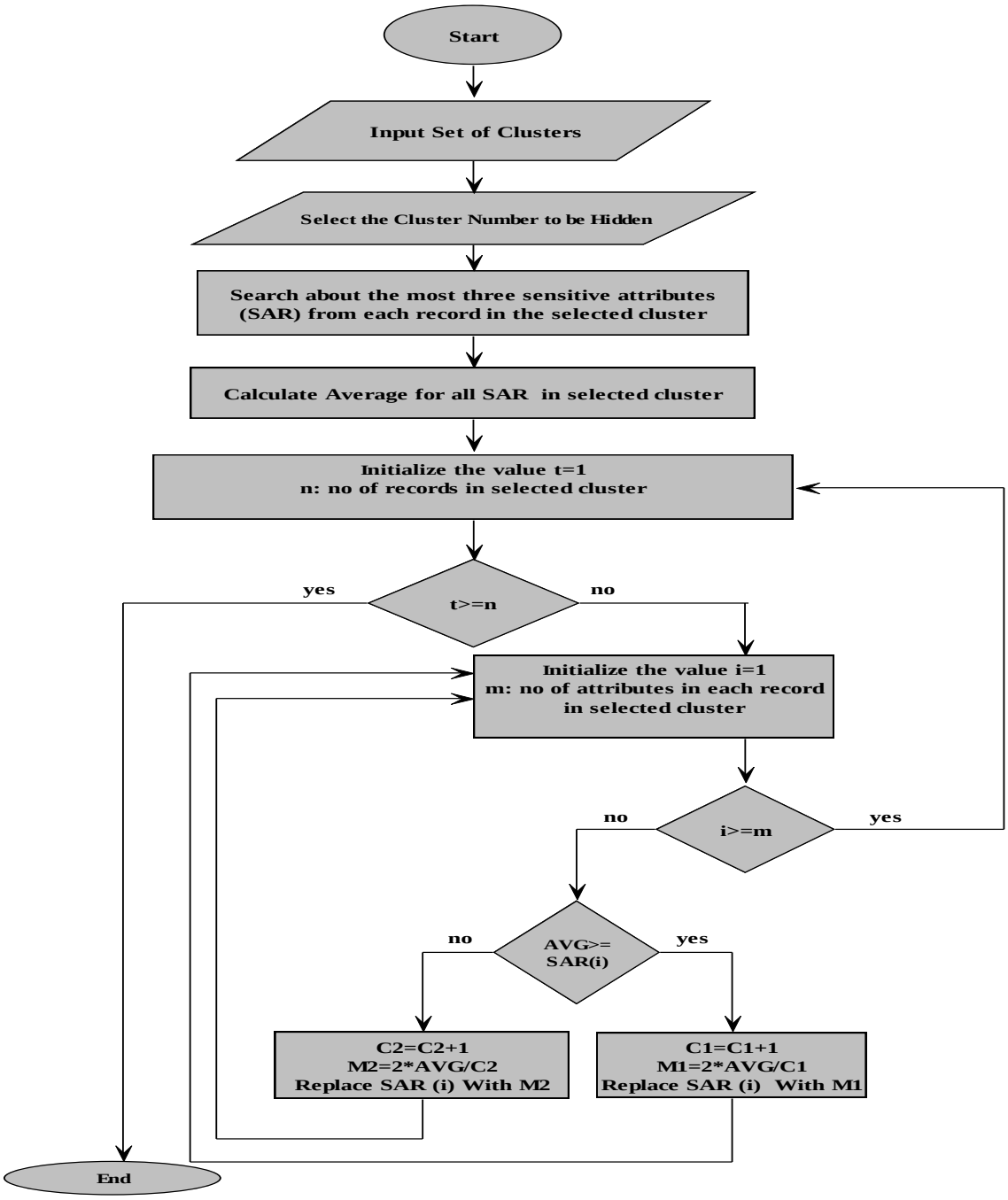


Fig. 3 A Flow Chart of Additive Noise algorithm

Where M1 and M2 are the noise values that they must add to the each *SAR(i),C1* and *C2* are counters.

3.6.2 Data Swapping Technique

Data swapping preserves the privacy of original sensitive information available at record level. If the records are picked at random for each swap then it is called random swaps. It is difficult for an intruder to recognize particular person or entity in database, because all the records are altered to the maximum level. In this research data swapping process is developed(updated) in order to increase the privacy result by choosing more than one value in the selected cluster (the values that are caused of founding this record in that cluster) to swap across different records in other clusters ,in order to shift this record to another cluster as shown in the following steps:

1-detect three of attributes(sensitive) in every record in the selected cluster as mentioned in Equation(3.7).
2- the first value is swapped with the symmetric value for any record from another cluster.
3-the second value is swapped with the symmetric value for any record from different cluster, and so on.

Note: *symmetric value is the value that has the same index in different record in another cluster* .A Flow Chart of Data Swapping algorithm is summarized in Figure 4. .

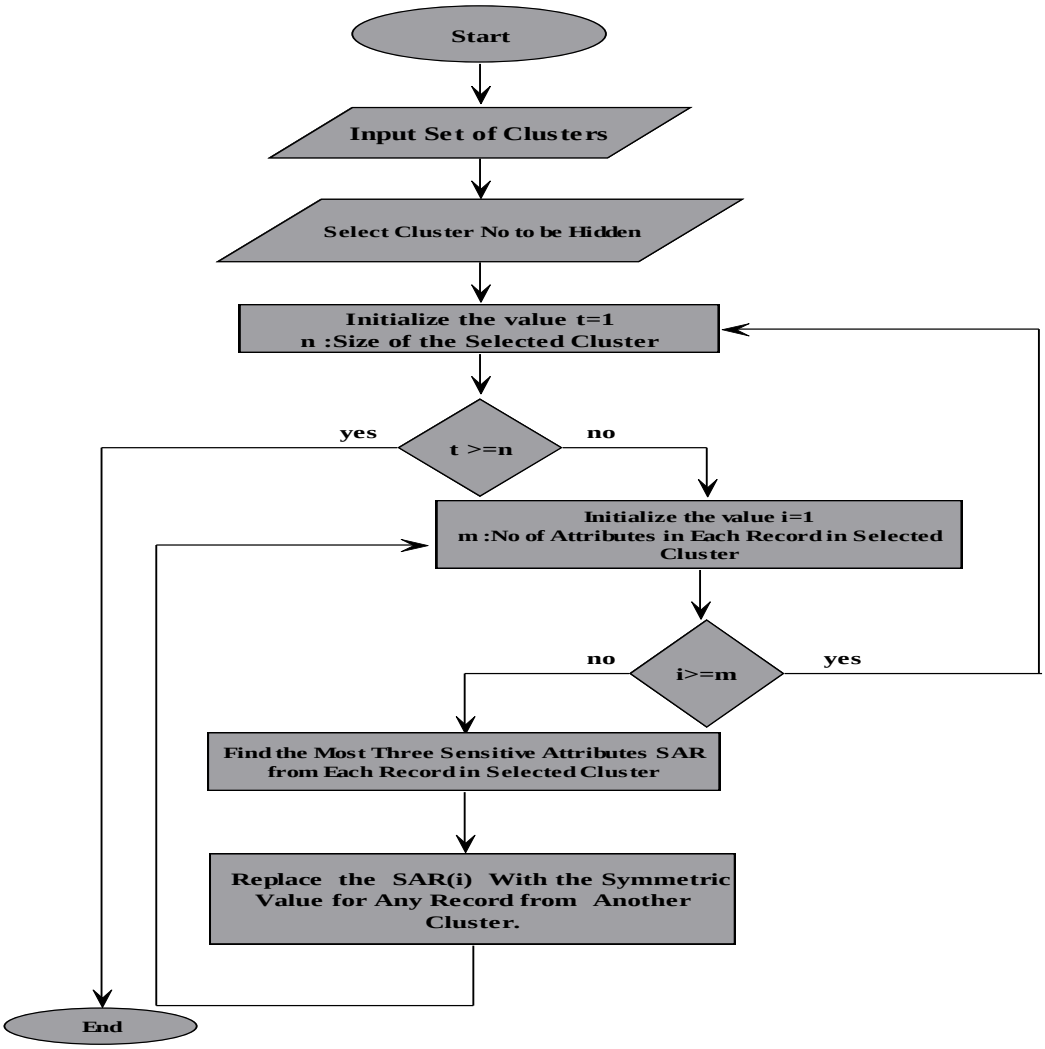


Fig .4 A Flow Chart of Data Swapping Algorithm

3.6.3 Data Copying Technique

This is a new perturbative technique that is suggested in this research for protecting the sensitive numerical attributes in the selected 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 Figure .5

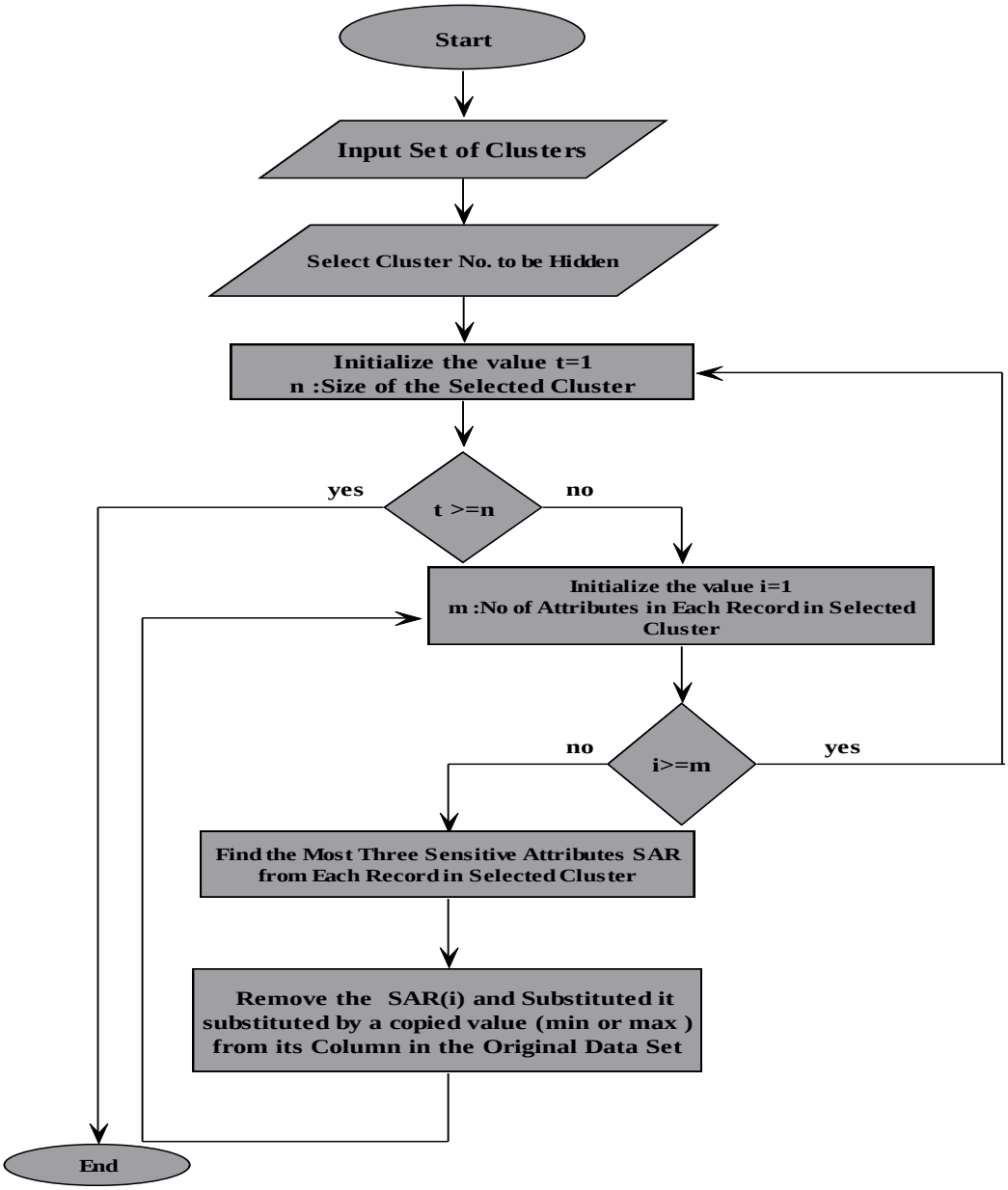


Fig.5 A Flow Chart of Data Copying Algorithm

3.7 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.

3.8 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.

3.8.1 Privacy Ratio.

The privacy ratio 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 :

$$\boldsymbol{PR = \left(1 - \frac{R(C')}{R(C)}\right) * 100} \qquad \textbf{…(2)}$$

3.8.2 information Loss Ratio

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

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

3.8.3 Covering of Data Ratio

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

$$COD = \frac{C}{K} * 100 \qquad \textbf{….(4)}$$

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.

3.8.4 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.

3.9 Experiments and Results for Applying PPDM Techniques and The Evaluation Measures.

The requirement cluster are protected according to the results of the following evaluation measures.

3.9.1 Privacy Ratio Results

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

Note: *The cluster to be hidden that has blue color.* Table. 1 describes the results that obtained by implementation all the PPDM techniques for all data sets:

Table. 1 Privacy Ratio Results Using Three Datasets

Data set	copy	Noise addition	swapping
wisconsin breast cancer	100٪	99.5121٪	92.3043٪
Diabetes	95.5752٪	97.8873٪	91.1764٪
heart stat log	48٪	52.1739٪	31.25٪

Figure 15 Shows the privacy ratio results for the different datasets and privacy techniques.

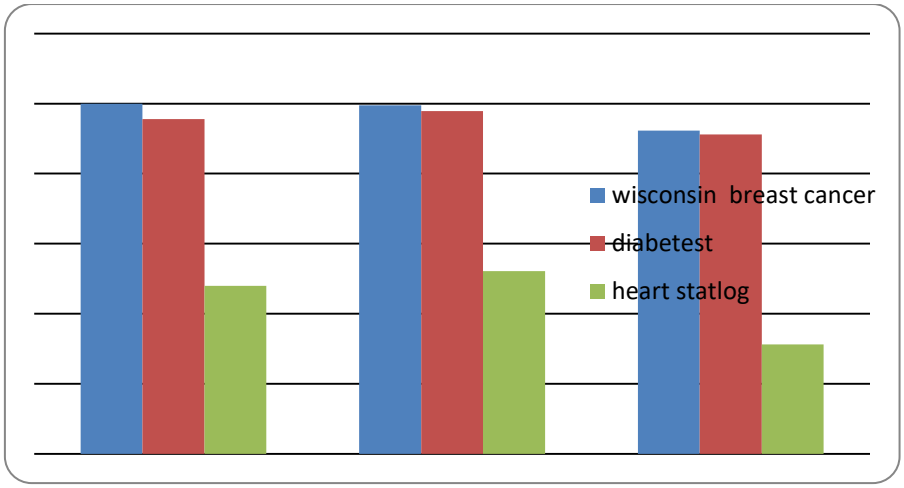


Fig. 15 Privacy Ratio

3.9.2 Information Loss Ratio Results

As mentioned in above, this factor is used to compute the percentage of distortion the information of all data set after applying the privacy technique . The results are presented as in Table.2 :

Table.2 Information Loss Ratio Results Using Three Datasets

Data Set	Copy	Noise Addition	Swapping
wisconsin breast cancer	٪1.2579	٪**0.3087**	٪0.4308
Diabetes	٪16.6339	٪0.4004	٪1.8505
heart stat log	٪0.7486	٪0.4306	٪0.3492

Figure 16 Shows information loss ratio results for the different datasets and privacy techniques.

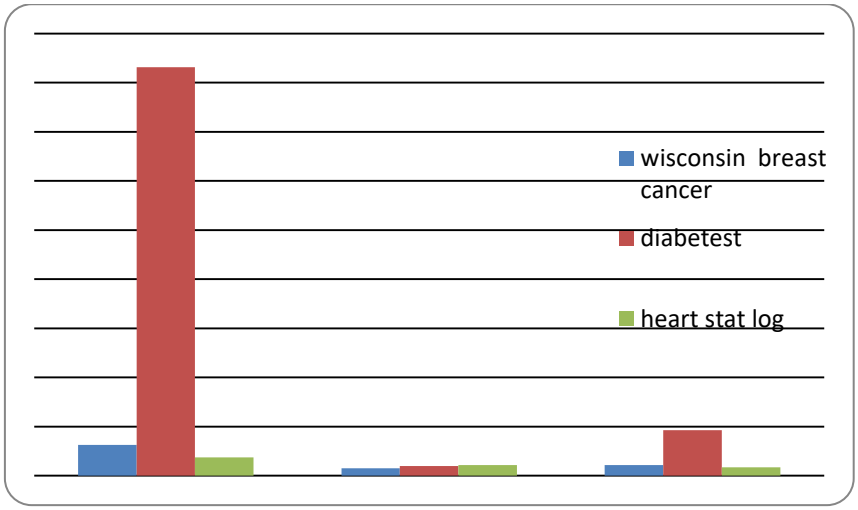


Fig. 16 Information Loss Ratio

3.9.3 Covering of Data Ratio Results

As mentioned in above, this performance factor is used to measure the percentage of average number of clusters covering in hidden cluster. The results are presented as in Table.3:

Table.3 Covering Data Ratio Results Using Three Dataset

Data set	copy	Noise addition	Swapping
wisconsin breast cancer	**66.66**٪	**66.66**٪	100٪
Diabetes	100٪	100٪	100٪
heart stat log	100٪	100٪	100٪

Figure 17 Shows covering of data ratio results for the different datasets and privacy techniques.

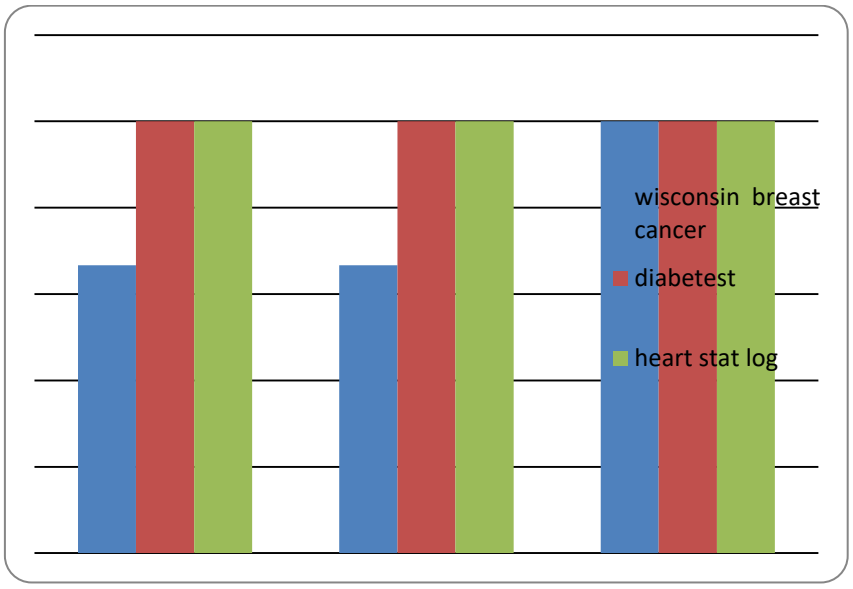


Fig. 17 Covering of Data Ratio

As shown in the above figure, Breast Cancer Wisconsin have the percentage of covering data equal to 66.66% (no of clusters that contain the records of requirement cluster is 2) in copy and noise addition because the privacy ratio for it was 100٪ and 99.5٪ respectively, that means the requirement cluster is completely hidden. While in swapping ,percentage of covering data was equal to 100٪(no of clusters that contain the records of requirement cluster is 3) because the privacy ratio for it was 92.30٪ ,therefore Some of few records are still near to its cluster.

3.9.4 Running Time Results

As mentioned in above, in this work, the efficiency(time requirements) is calculated by using the CPU time. The results are presented in seconds as in Table.4:

Table.4 The Execution Time Results Using Three Datasets

Data set	Copy	Noise addition	Swapping
wisconsin breast cancer	**1.529 sec**	10.047 sec	40.997 sec
Diabetes	1.825 sec	3.167 sec	4.01 sec
heart stat log	3.463 sec	31.34 sec	35.022 sec

Figure 18 Shows running time results in seconds for the different datasets and privacy techniques.

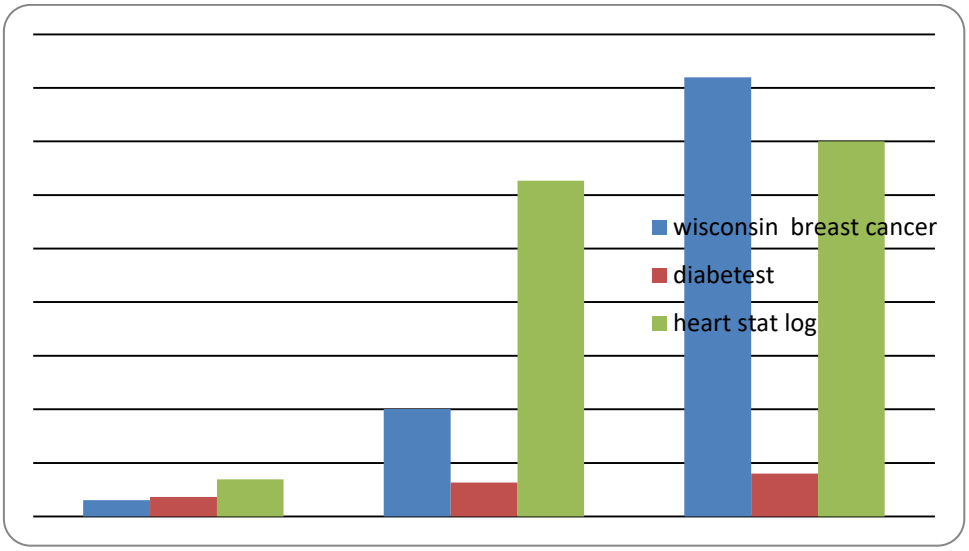


Fig.18 Running Time Results

4. Conclusions

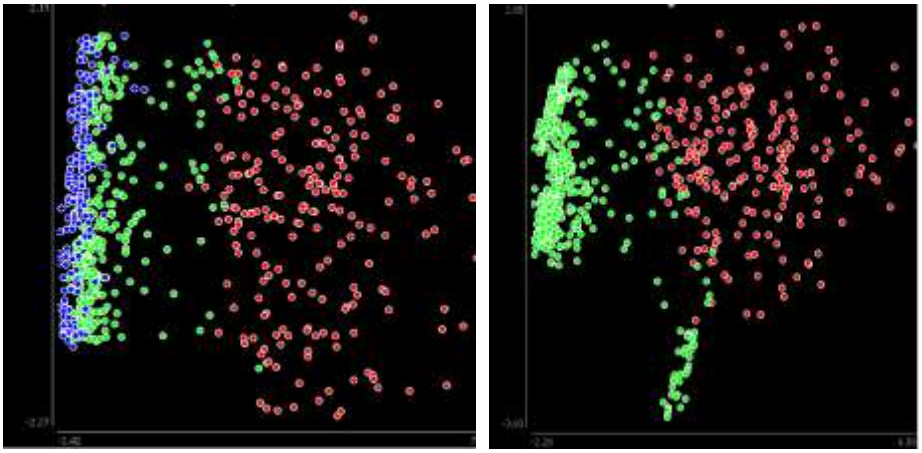
In the present study, the following facts can be concluded:

1. The privacy ratio results for heart stat log data set was 48%, 52.1739 % and 31.25% in copy, additive noise and swapping techniques respectively because these kinds of data sets have the special property that they are extremely *sparse*..
2. Distortion of data in the all data sets has been minimum values, except the Diabetes data set which the result of distortion on its data when applying Data Copying algorithm equal to 16.6339% because the sensitive attribute is removed and substituted by a copied value (min or max) from its column in the original data set which it is farthest from it in range.
3. Covering of data proved that the records of requirement cluster are distributed with good format between other clusters. The results showed that Breast Cancer Wisconsin data set have the percentage of covering data equal to 66.66% (no. of clusters that contain the records of requirement cluster is 2) in copy and noise addition because the privacy ratio for it was 100٪ and 99.5٪ respectively, that means the requirement cluster is completely hidden.

5. Suggestions and Future Works

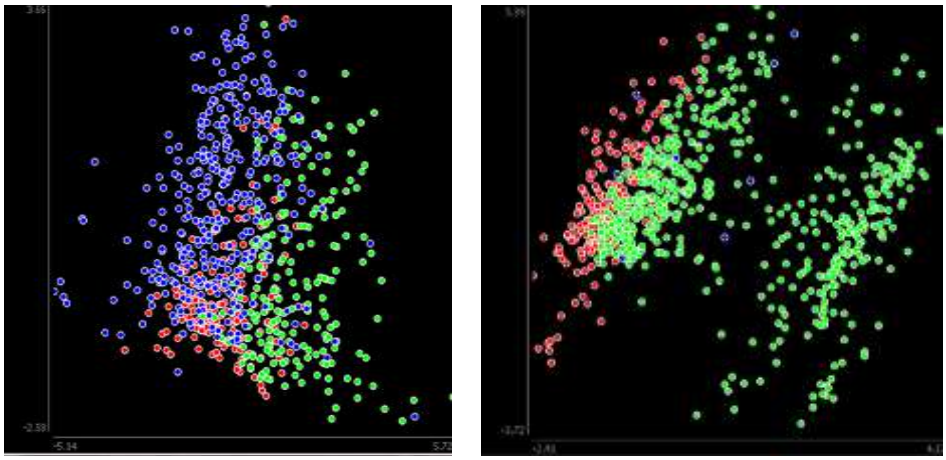
1. Developing the suggested privacy techniques by choosing the sensitive cluster number automatically instead of manually and according to specific conditions(measures).
2. Applying another clustering technique such as DBSCAN algorithm and comparing the results with our method.
3. Applying PPDM with Artificial Bee Colony(ABC) algorithm by using one of the better clustering algorithm.

APPENDICIES



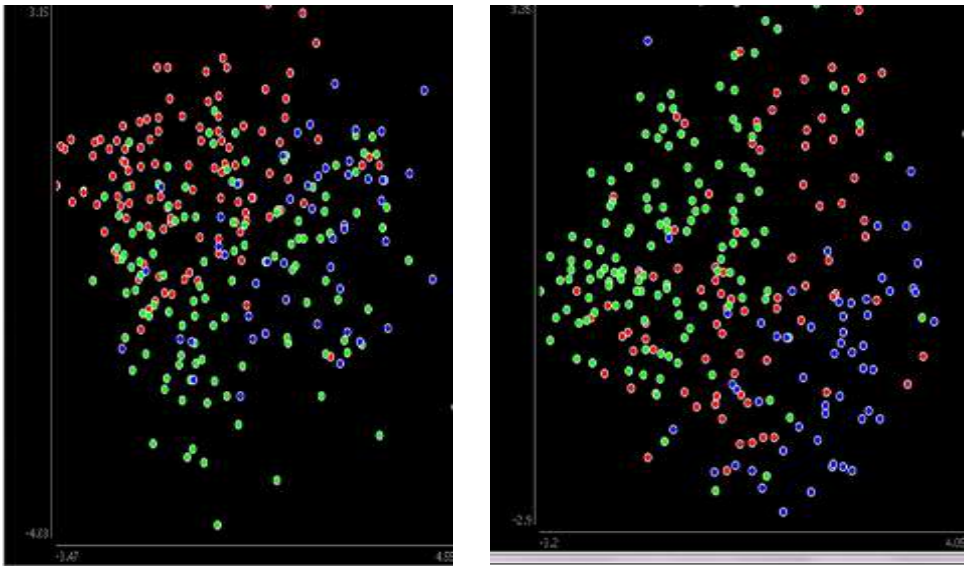
(a) **(b)**

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



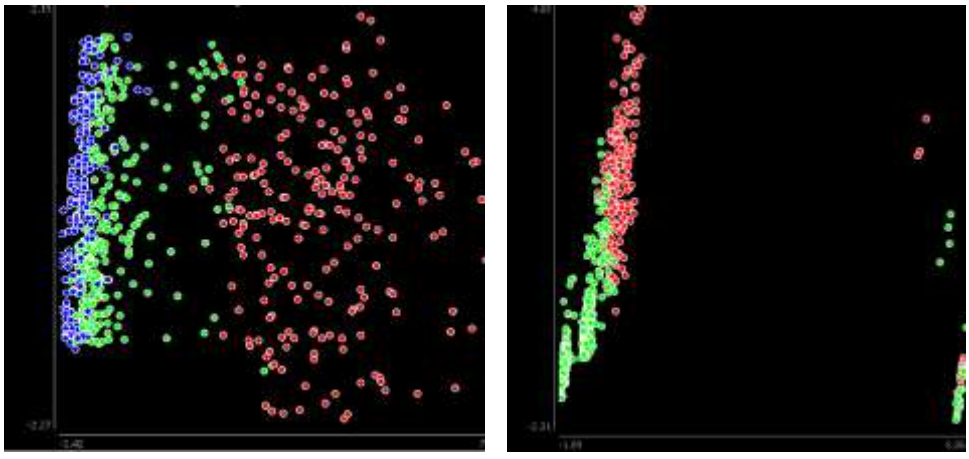
(a) **(b)**

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



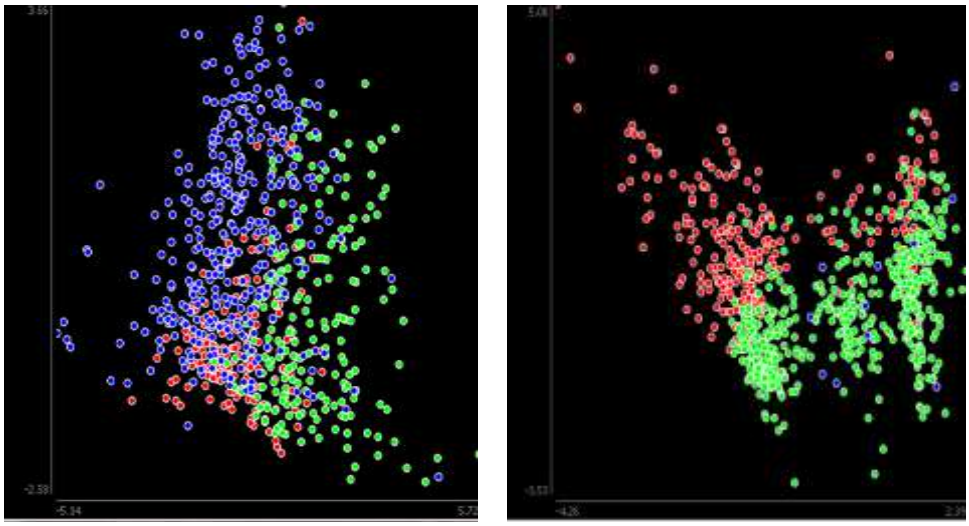
(a) **(b)**

Fig 8 The Privacy Ratio Results for Heart Stat Log Dataset Using Copy Technique. (a) Before privacy , (b) After privacy.



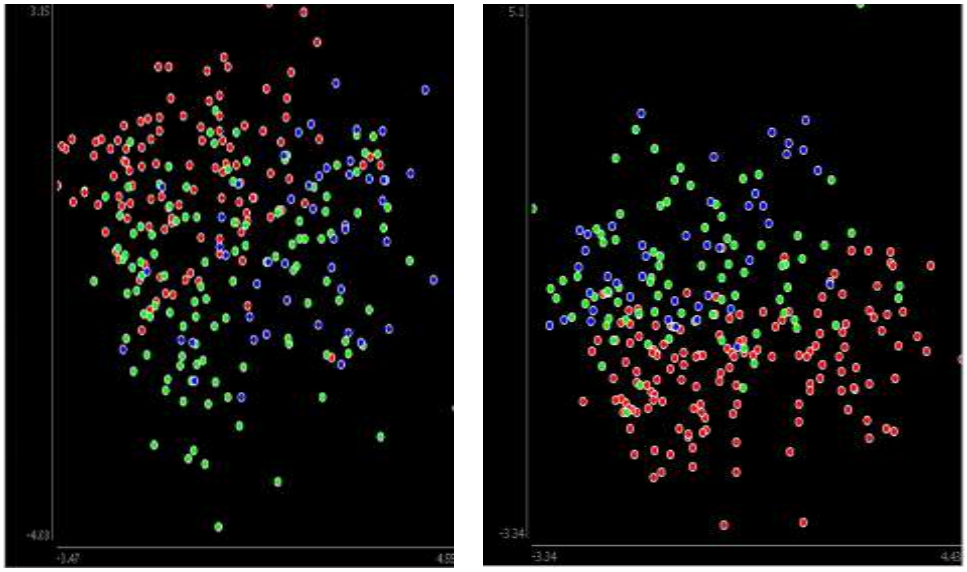
(a) **(b)**

Fig 9 The Privacy Ratio Results for Breast Cancer Wisconsin Dataset Using Additive Noise Technique. (a) Before privacy , (b) After privacy.



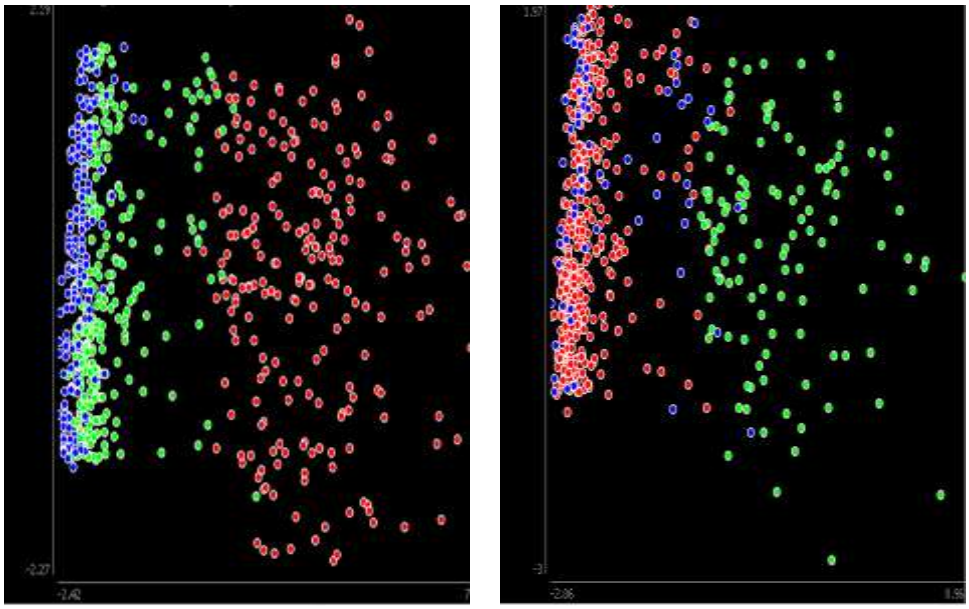
(a) **(b)**

Fig 10 The Privacy Ratio Results for Diabetes Dataset Using Additive Noise Technique. (a) Before privacy , (b) After privacy.



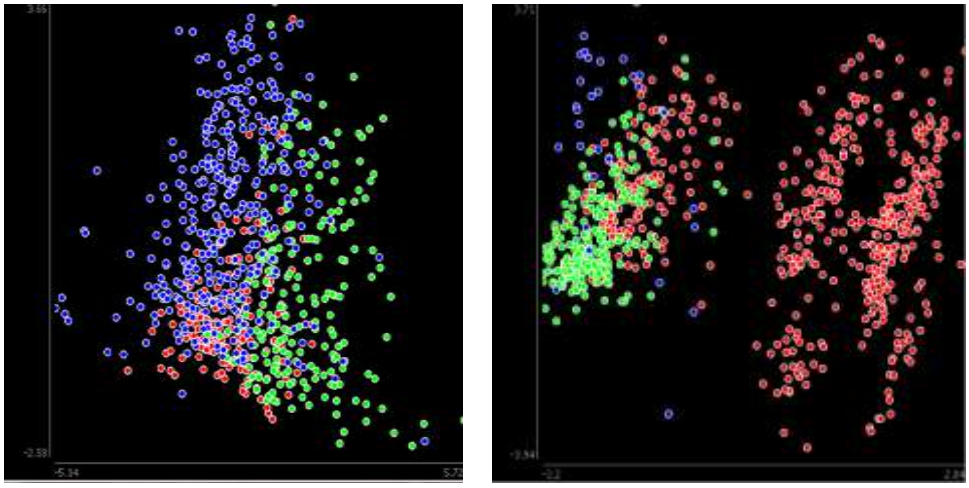
(a) **(b)**

Fig 11 The Privacy Ratio Results for Heart Stat Log Dataset Using Additive Noise Technique. (a) Before privacy , (b) After privacy.



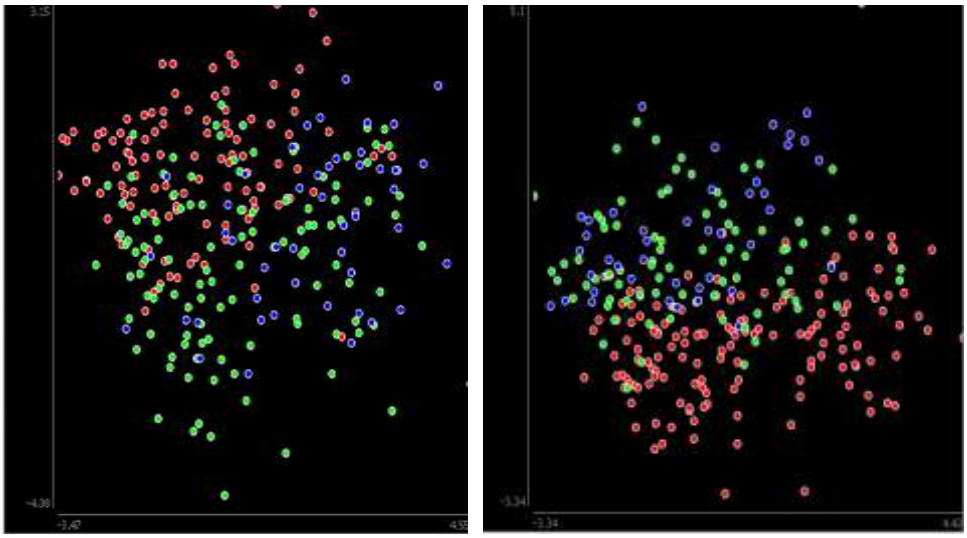
(a) **(b)**

Fig 12 The Privacy Ratio Results for Breast Cancer Wisconsin Dataset Using Swapping Technique. (a) Before privacy , (b) After privacy.



(a) **(b)**

Fig 13 The Privacy Ratio Results for Diabetes Dataset Using Swapping Technique. (a) Before privacy , (b) After privacy.



(a) **(b)**

Fig 14 The Privacy Ratio Results for Heart Stat Log Dataset Using Swapping 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 (2011) 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,2006:1–224.

[9] S.Vijayarani and M.Sathiya Prabha." Association Rule Hiding using Artificial Bee Colony Algorithm". International Journal of Computer Applications (0975 – 8887) Volume 33– No.2, November 2011.