Privacy Preserving in Data Mining Using PAM Clustering Algorithm

Heba A. Raheem
Department of computer science, Karbala University, Iraq
Heba_ad74@yahoo.com
Safaa O. Al-Mamory
college of Information Technology, Babylon University, Iraq
safaa_vb@yahoo.com

Abstract

"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 the data warehouses. Privacy preserving data mining is a latest research area in the field of data mining which generally deals with the side effects of the data mining techniques. Privacy is defined as "protecting individual's information". Protection of privacy has become an important issue in data mining research"(S.Vijayarani et al.,2011) .

Clustering is a division of data into groups of similar objects. In this paper we have used PAM clustering algorithms in health datasets. The cluster selected to be hided are considered as sensitive cluster. This sensitive cluster is protected by using Additive Noise Perturbation random method.

Keywords: Data Mining, Clustering, PAM, Privacy

الخلاصة:

تعدين البيانات هي أنتزاع المعلومات التنبؤيه المخفيه من قواعد البيانات الكبيره وأيضا تقنيه جديده قويه ذات أمكانيه عظيمه لتحليل معلومات مهمه في مخازن البيانات. أبقاء السريه في تعدين البيانات هي آخر منطقة بحث في حقل تنقيب البيانات والتي هي بصورهعامه تتعامل مع الآثار الجانبيه لتقنيات تنقيب البيانات.السريه معرفه على أنها"حماية معلومات الفرد".حماية السريه أصبحت قضيه مهمه في بحث تعدين البيانات. في هذا البحث استخدمنا خوارزمية التجمعأو العنقده PAM في بيانات طبيه. والعنقده هي تقسيم البيانات الى مجموعة كيانات أوقيود متماثله.العنقود الذي يتم أختياره ليكون مخفيا يعتبر على أنه عنقود حساس(مهم).وتتم حمايته بأستخدام خوارزمية اضافة ضوضاء عشوائية Additive Noise Perturbation Randome Method .

اللكلمات المفتاحية : تنقب البيانات , تجمع pam, سري.

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."(S.Vijayarani *et al.*,2011) .

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. Confronted with huge collection of data, have to created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data so using data mining for analyzing and extracting the data from large databases.

Privacy is defined as “protecting individual’s information”. Protection of privacy has become an important issue in data mining research. A number of privacy-preserving data mining methods have recently been proposed which take either a cryptographic or a statistical approach. The cryptographic approach ensures strong privacy and accuracy via a secure multi-party computation, but typically suffers from its poor performance. The statistical approach has been used to mine decision trees, association rules, and clustering, and is popular mainly because of its high performance.

A standard dictionary definition of privacy as it pertains to data is "freedom from unauthorized intrusion"(Aggarwal C.C,2008). If users have given authorization to use the data for the particular data mining task, then there is no privacy issue. However if the user is not authorized that constitutes "intrusion". Privacy applies to “individually identifiable data”.

“A number of techniques such as randomization, k-anonymity, distributed privacy preservation, query auditing, and data publishing and cryptographic methods have been suggested in recent years in order to perform privacy-preserving data mining. A privacy-preserving data mining technique must ensure that any information disclosed

- Cannot be traced to an individual .
- Does not constitute an intrusion. “(Ajay Challagalla et al.,2010)

Despite the fact that this field is new, and that privacy is not yet fully defined, there are many applications where privacy-preserving data mining can be shown to provide useful knowledge while meeting accepted standards for protecting privacy. As an example, consider mining of supermarket transaction data. Most supermarkets now offer discount cards to consumers who are willing to have their purchases tracked. Generating association rules from such data is a commonly used data mining example, leading to insight into buyer behavior that can be used to redesign store layouts, develop retailing promotions, etc.

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 consideration in privacy preserving data mining is twofold . First, sensitive raw data should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded. 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(Chuang-Cheng Chiu et al,2007). The rest of the paper is organized as follows:

In Section 2 describes The related works. In Section3 problem formulation and the privacy technique is given. The experimental results of the privacy technique are discussed in Section4.Conclusions are given in Section5.

2.RELATED WORKS

Chuang-Cheng *et al.* (2007) proposed a novel clustering method for conducting the *k*-anonymity model effectively. In the proposed clustering method, feature weights are automatically adjusted so that the information distortion can be reduced. A set of experiments show that the proposed method keeps the benefit of scalability and computational efficiency when comparing to other popular clustering algorithms.

The similarity between this method and our proposal method in the reducing the information distortion. The difference in clutstering algorithm that is used and privacy technique.

JianWang *et. al.* (2009) 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. Each group has a size at least k, which is referred to as the level of that privacy preserving approach. The greater the level, the greater the amount of privacy. At the same time, 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.

R.Vidyabanu *et. al.* (2010) proposed technique for privacy preserving clustering using Principal component Analysis(PCA) based transformation approach. This method is suitable for clustering horizontally partitioned or centralized data sets .The framework was implemented on synthetic datasets and clustering was done using Self organizing Map(SOM).The accuracy of clustering before and after privacy preserving transformation was estimate.

S.Vijayarani *et al*. (2011) 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. Experimental results shows that the ECLARANS algorithm is the best algorithm for detecting the outliers and the sensitive outliers are protected efficiently by the Gaussian Perturbation random method. 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 .

S.Vijayarani *et al.* (2011) they have analyzed the performance of two perturbation masking techniques namely data transformation technique and bit transformation technique which are used for protecting sensitive numeric data in the micro data table. The experimental result shows that the data transformation technique has produced better results than bit transformation approach by using k_means clustering algorithm.

S.Vijayarani *et al*. (2011) presented a new perturbation masking technique called as modified data transitive technique(MDTT) is used for protecting the sensitive numerical attributes(s) .The performance of the proposed technique(MDTT) is compared with the existing masking techniques additive noise, rounding and micro aggregation. The experimental result shows the MDTT technique has produced better results than existing techniques by using k_means clustering algorithm. 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 the results shows the efficiency of this method in protecting the sensitive information.

Md Zahidul Islam *et al.* (2011) presented a framework for adding noise to all attributes (both numerical and categorical) in two steps; in the first step they add noise to sensitive class attribute values, which are also known as labels. Additionally, in the next step they add noise to all non-class attributes to prevent re-identification of a record with high certainty and disclosure of a sensitive class value. Noise addition to non-class attributes also protect the attributes from being disclosed. The main goal of them noise addition technique is to provide high level of security while preserving a good data quality by using a novel clustering technique. As mentioned in above 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 technique 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:

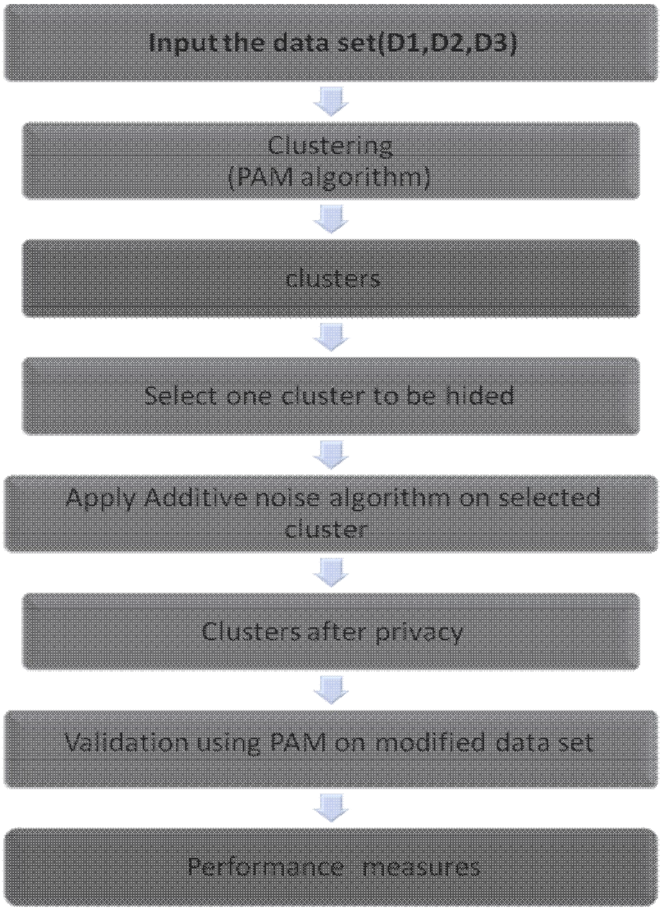

Input the data set(D1,D2,D3)
Clustering
(PAM algorithm)
clusters
Select one cluster to be hided
Apply Additive noise algorithm on selected cluster
Clusters after privacy
Validation using PAM on modified data set
Performance measures

Fig.1 The System Architecture

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. .

We modified the dataset by dealing with the null values .To do so we replace it with more repeating value of that attribute over the whole dataset.

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 An Approach for clustering

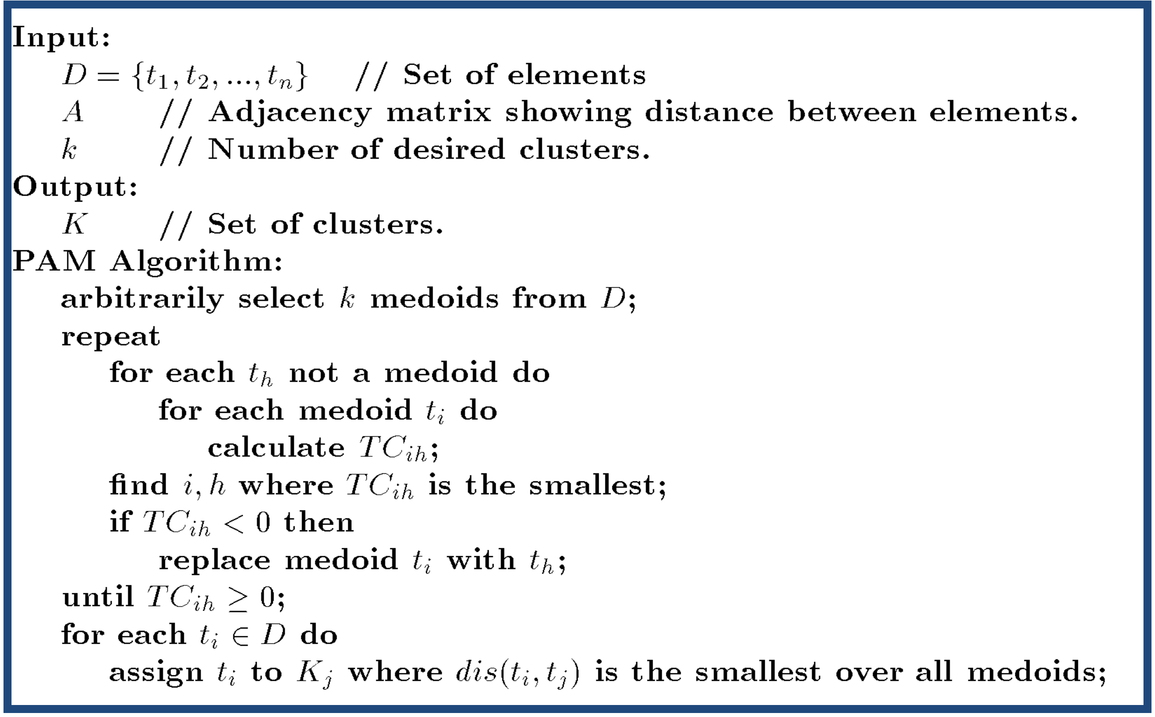
The following clustering algorithm are used in our research:

3.2.1 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. Mainly it contains two phases :

Build phase: This step is sequentially select k elements which is centrally located .This k elements to be used as k-medoids.

Swap phase: Calculates the total cost for each pair of selected and non-selected element.The pseudo-code of PAM clustering algorithm is summarized in Fig.2:



```
Input:
    $D = \{t_1, t_2, ..., t_n\}$    // Set of elements
    $A$      // Adjacency matrix showing distance between elements.
    $k$      // Number of desired clusters.
Output:
    $K$      // Set of clusters.
PAM Algorithm:
    arbitrarily select $k$ medoids from $D$;
    repeat
        for each $t_h$ not a medoid do
            for each medoid $t_i$ do
                calculate $TC_{ih}$;
        find $i, h$ where $TC_{ih}$ is the smallest;
        if $TC_{ih} < 0$ then
            replace medoid $t_i$ with $t_h$;
    until $TC_{ih} \geq 0$;
    for each $t_i \in D$ do
        assign $t_i$ to $K_j$ where $dis(t_i, t_j)$ is the smallest over all medoids;
```

Fig.2 The pseudo-code of PAM clustering algorithm (Margaret H.Dunham,2002).

Where TC is total cost for each pair of selected and non-selected element that is computed by Equation.1:

$$\mathrm{cost}(x, c) = \sum_{i=1}^{d} |x - c| \quad .....(1)$$

where *x* is any data object, *c* is the medoid, and *d* is the dimension of the element(object).

3.3 select the cluster to be hided

In this step the cluster selected to be hided manually 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 .

3.4 apply a privacy preserving data mining techniques on the selected cluster.

3.4.1 Additive Noise

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 .The pseudo-code of Additive Noise algorithm [S.Vijayarani *et al*,2011] is summarized in Fig.3.

```
Input: the cluster number to be hided.
Output: set of clusters after privacy

Additive Noise algorithm :
1. Consider a data base D with  n  tuples t = {t1, t2….tn}. for the selected cluster
Each tuples contains Set of attribute A = {A1, A2…..Am} A € ti.
2. Find the most three sensitive attributes SAR for all SAR € Ai € A (i=1,2…m) from each record in the selected cluster that have min distance value (min SSE error) between the record and the medoid(representative point of the selected cluster).
3. Calculate Average for all SAR (i).
4. Do
  For each  SAR (i)
Initialize the value i = (1, 2….n).
Check if (Average >=SAR (i.) to count the all values C1.
Calculate M1 = (2*Average)/C1
Replace SAR (i). With M1.
Check if (Average <= SAR (i.) to count the all values C2.
Calculate M2 = (2*Average)/C2
Replace SAR (i). With M2.
Increment the value of i.
5. While (i >=n)
6. End
```

Fig.3 The pseudo-code of Additive Noise algorithm

Where M1 and M2 is the noise value that it must add to the each *SAR (i)*,*C1* and *C2* are counters, min SSE error is minimum of sum square error value between the record and the medoid that is computed by Equation.2

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

Thus by applying Additive Noise Random method for each one of our health data set in our Research, 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 are hided or not. All the sensitive attributes in the requirement cluster are protected by using this technique.

4.EXPERIMENTAL RESULTS

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.

1-privacy ratio.

2-information loss ratio.

3- covering of data ratio
4- Run time.

4.1 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[**safaa 2008**]. The privacy ratio is calculated by Equation3.

$$PR = \left(1 - \frac{R(C')}{R(C)}\right) * 100 \qquad \text{.........}(3)$$

Where R(C□) is the number of records that remained near to the original cluster after privacy, R(C) is the number of records in the original cluster before privacy.

Table.1 privacy ratio results using three datasets

Data set	Privacy ratio
wisconsin breast cancer	97.727
diabetest	97.887
heart statlog	52.173

Fig.4:Shows privacy ratio results for our different datasets

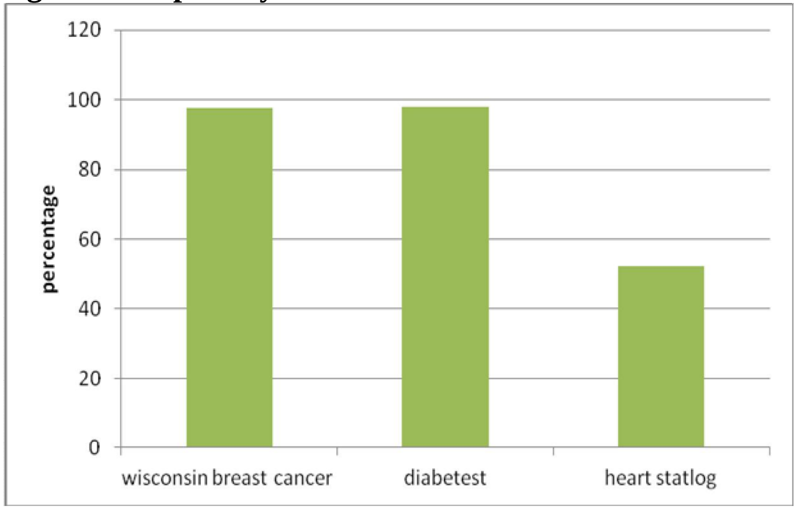


Fig4:privacy ratio

4.2 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 .it is measured by the percentage of the summation the difference between original value and modified value for the sensitive attribute in each record of the selected cluster and summation of original values for all data set[**S.Vijayarani 2011**] . The information loss ratio is calculated by Equation 4.

$$ILR = (\sum|original\ value - new\ value|\ )/(\sum|original\ values\ |) *100 \qquad \text{…….}(4)$$

Table.2 information loss ratio results using three datasets

Data set	information loss ratio
wisconsin breast cancer	0.3087
diabetest	0.4004
heart statlog	0.4306

Fig5: Shows the information loss results for our different datasets

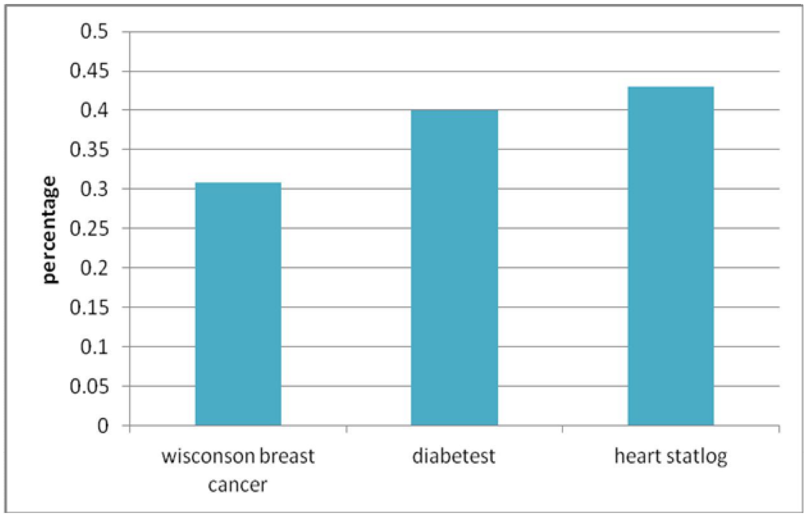


Fig5:information loss ratio

4.3 covering of data ratio

This performance factor is used to measure the percentage of average number of clusters covering in hidden cluster[**safaa 2008**]. it is calculated by Equation 5.

$$COD = \frac{C}{K} * 100 \quad ...(5)$$

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.

Table.3 covering of data ratio results using three datasets

Data set	covering of data ratio
wisconsin breast cancer	90
Diabetes	60
heart statlog	60

Fig6: Shows the covering of data ratio results for our different datasets

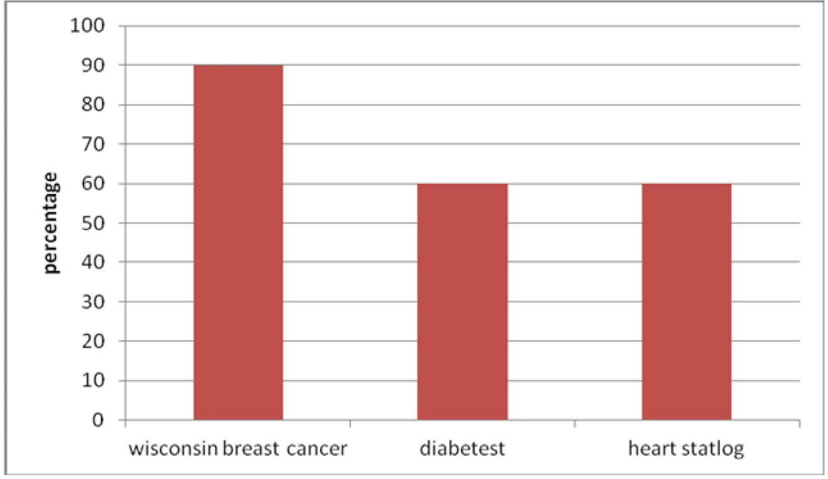


Fig6:covering of data ratio

4.4 Run time

Time requirements can be evaluated in terms of CPU time, or computational cost, or even the average of the number of operations required by the PPDM technique. In this work, the efficiency(time requirements) is calculated by using the CPU time.

Fig7: Shows the time efficiency for our different datasets

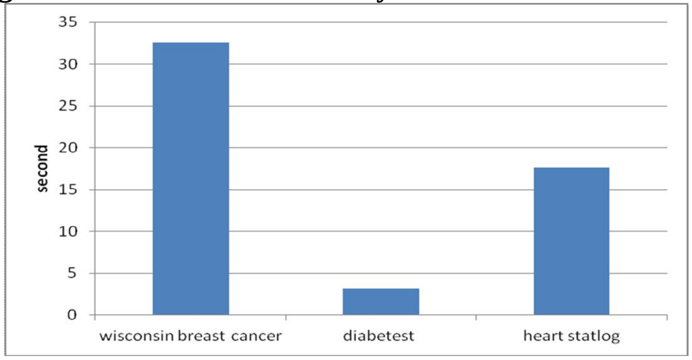


Fig7:time efficiency

5.CONCLUSIONS

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. In this paper we have used PAM clustering algorithm in health datasets. the cluster selected to be hided are considered as sensitive cluster. This sensitive cluster is protected by using Additive Noise Perturbation random method. Different performance factors are used for measuring the efficiency of the clustering algorithm and the privacy technique Additive Noise Perturbation random method. Experimental results shows that the PAM algorithm is the efficient algorithm for clustering in all data sets and the selected cluster are protected efficiently by the Additive Noise random method in Wisconsin breast cancer and diabetes data set ,except heart stat log data set, we show from the privacy ratio results ,our privacy technique has not work efficiently with this data set because these kinds of data sets have the special property that they are extremely *sparse*. The sparsely property implies that only a few of the attributes are non-zero, and most of the attributes take on zero values.

6.REFERENCES

Aggarwal C.C, Yu P.S. 2008 “Models and Algorithms: Privacy-Preserving Data Mining”, Springer, ISBN: 0-387-70991-8.

Ajay Challagalla,S.S.Shivaji Dhiraj ,D.V.L.N Somayajulu,Toms ShajiMathew,Saurav Tiwari,SyedSharique Ahmad. 2010 “ Privacy Preserving Outlier Detection Using Hierarchical Clustering Methods” ,4th Annual IEEE Computer Software and Applications Conference Workshops.

Al-Zoubi, M., Al-Dahoud, A. and Yahya, A.A. 2010 “New Outlier Detection Method Based on Fuzzy Clustering, WSEAS Transactions on Information Science and Applications”, Vol. 7, Issue 5

Al-Zoubi, M. 2009 “An Effective Clustering-Based Approach for Outlier Detection”, European Journal of Scientific Research.

Chuang-Cheng Chiu and Chieh-YuanTsai. 2007 A *k*-Anonymity Clustering Method for Effective Data

Privacy Preservation”, Springer, (Eds.): ADMA 2007, LNAI 4632, pp. 89–99,

JianWang , YongchengLuo, Yan Zhao JiajinLe. 2009” A Survey on Privacy Preserving Data Mining”, First International Workshop on Database Technology and Applications.

R.Vidyabanu and Dr N.Nagaveni. 2010” A Model Based Framework for Privacy Preserving

Clustering Using SOM”, International Journal of Computer Applications (0975 – 8887),Volume 1 – No. 13.

S.Vijayarani and S.Nithya. 2011” Sensitive Outlier Protection in Privacy Preserving Data Mining**”,** International Journal of Computer Applications (0975 – 8887), Volume 33– No.3, November.

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

S.Vijayarani, Dr. A.Tamilarasi, N.Murugesh.(2011)”A NEW TECHNIQUE FOR PROTECTING SENSITIVE

DATA AND EVALUATING CLUSTERING PERFORMANCE”, International Journal of Information Technology Convergence and Services (IJITCS) Vol.1, No.2, April 2011.

Md Zahidul Islam, Ljiljana Brankovic.(2011)” Privacy preserving data mining: A noise addition framework using a novel clustering technique ”, Elsevier .

Margaret H. Dunham.(2002).”Data Mining, Introductory and Advanced Topics”,Prentice Hall.