Sensitive Outlier Protection in Privacy Preserving Data Mining

S.Vijayarani
Assistant Professor,
School of Computer Science and Engineering
Bharathiar University
Coimbatore

S.Nithya
M.Phil Research Scholar,
School of Computer Science and Engineering
Bharathiar University,
Coimbatore

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 their 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. Sensitive outlier protection is novel research in the data mining research field. Clustering is a division of data into groups of similar objects. One of the main tasks in data mining research is Outlier Detection. In data mining, clustering algorithms are used for detecting the outliers efficiently. In this paper we have used four clustering algorithms to detect outliers and also proposed a new privacy technique GAUSSIAN PERTURBATION RANDOM METHOD to protect the sensitive outliers in health data sets.

General Terms
Data mining, Clustering, Outlier detection, Privacy

Keywords
Data Mining, Privacy, Clustering, PAM, CLARA, CLARANS, ECLARANS, Outlier Detection, Gaussian Perturbation Random Method

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.

The efficient database management systems have been very important resources for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection at any time needed. The propagation of database management systems has also contributed to recent massive gathering of all sorts of information. 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 primary requirement of privacy-preserving data mining is to guard the input data, yet still allow data miners to extract useful knowledge models [1]. 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"[1]. 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 [2]

- Cannot be traced to an individual
- Does not constitute an intrusion.

Any data does not give us completely accurate knowledge about a specific individual meets these criteria. At the other extreme, any improvement of knowledge about an individual could be considered an intrusion. The latter is particularly likely to cause a problem for data mining, as the goal is to improve the knowledge. Even though the target is often groups of individuals, knowing more about a group does increase the knowledge about individuals in the group. It means need to measure both the knowledge gained and ability to relate it to a particular individual, and determine if these exceed thresholds.

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. A number of techniques such as randomization and *k*-anonymity [2, 3, and 4] have been suggested in recent years in order to perform privacy-preserving data mining.

The rest of the paper is organized as follows. In Section 2 describes clustering outlier detection techniques. The related works are discussed Section 3. In Section 4, problem formulation and the proposed privacy technique for sensitive outliers is given. The experimental results of the proposed privacy technique are discussed in Section 5. Conclusions and the are given in Section 6.

2. OUTLIER DETECTION BASED ON CLUSTERING APPROACHES

A few clustering algorithms such as PAM,CLARA,CLARANS,DBSCAN, BIRCH, ROCK, STING, Wave Cluster can also handle outliers, but their main concern is to find clusters and the outliers in the context of clustering are often regarded as noise. In general, outliers are typically just ignored or tolerated in the clustering process for these algorithms are optimized for producing meaningful clusters, which prevents giving good results on outlier detection.

Most methods on outlier mining in the early work are based on statistics. These methods can be mainly classified into two categories: *distribution-based* and *depth-based* ones. The *distribution-based* methods use standard distribution to fit the dataset. Outliers are defined on the basis of probability distribution. The main problem with *distribution-based* method is that it assumes that the underlying data distribution is known a prior. However, for many applications, the prior knowledge is not always obtainable, and the cost for fitting data with standard distribution is significantly considerable. *Distance-based* scheme declares a point as an outlier if its neighbourhood contains less than *pct*% of a whole dataset. This notion generalizes many concepts from *distribution-based* method and enjoys better computational complexity. *Distance-based* scheme is further extended to improve the effectiveness of detection.

Deviation-based method identifies outliers by inspecting the main characteristics of objects in a dataset and the objects that "deviate" from these features are considered as outliers. It uses the "local outlier factor" (*LOF*) to measure how strong an object can be an outlier. Since the *LOF* value of an object is obtained by comparing its density with those in its neighbourhood, it has stronger modelling capability than *distance-based* scheme, which is based only on the density of the object itself.

Cluster-based outlier detection techniques were recently developed. Jiang, M.F., [19] proposed an outlier finding process, named *OFP*, which based on k-means algorithm and regarded small clusters as outliers. He, Z. et al., [26] proposed the concept of cluster-based local outlier and outlier detection method *FindCBLOF*, which used "*cluster-based* local outlier factor" for identifying the outlierness of each object. The method *OFP* and *FindOut* can only process numerical-attribute data; on the contrary, *FindCBLOF* can only process categorical attribute data. Jiang, S., [12] presented outlier detection *TOD*, which improves the efficiency of *FindCBLOF* method and can process mixed-attribute data.

3. RELATED WORKS

Loureiro, A., Torgo, L. And Soares, C. [18] describes a methodology for the application of hierarchical clustering methods to the task of outlier detection. The methodology is tested on the problem of cleaning official statistics data. The goal of this paper is the detection of erroneous foreign trade transactions in data collection. The methodology discussed here is able to save a large amount of time by selecting a small subset of suspicious transactions for manual inspection which includes most of the erroneous transactions. The authors compared several alternative hierarchical clustering methodologies for this task. The results they have obtained here confirmed the validity of the use of hierarchical clustering techniques for this task. Their comparison results show that their methodology improves previous results by keeping similar number of erroneous transactions identified with significantly.

Jiang, S., and An, Q., [12] generalizes local outlier factor of object and proposed a clustering-based outlier detection scheme (CBOD). The method consists of two phases, the first phase cluster dataset by one-pass clustering algorithm and second phase determine outlier cluster by outlier factor. The time complexity of CBOD is nearly linear with the size of dataset and the number of attributes, which results in good scalability and suitable to large dataset. The theoretic analysis and the experimental results show that the detection process is effective and feasible.

John Peter.S., et al., [14] discussed about the Minimum Spanning Tree based clustering algorithm for detecting outliers. They mentioned Minimum Spanning Tree based clustering algorithm is capable of detecting clusters with irregular boundaries. The algorithm partition the dataset into optimal number of clusters. Small clusters are then determined and considered as outliers. The rest of the outliers (if any) are then detected in the remaining clusters based on temporary removing an edge (Euclidean distance between objects) from the data set and recalculate the weight function. They introduce a new cluster validation criterion based on the geometric property of data partition of the dataset in order to find the proper number of clusters. The algorithm works in two stages. The first stage of the algorithm creates optimal number of clusters, where as the second stage of the algorithm detect outliers. The key feature of their algorithm is it finds noise-free/error-free clusters for a given dataset without using any input parameters.

Al-Zoubi, M., [3] proposes a method based on clustering approaches for outlier detection. They first perform the PAM clustering algorithm in that, small clusters are detected in the remaining clusters based on calculating the absolute distances between the results show that their method works well. The experimental results show that the proposed approaches give effective results when applied to different data sets.

Murugavel. P. et al., [20] compared three partition based algorithms with k-medoid distance based method for outlier detection. Here they improve the time efficiency and accuracy of detection. The main advantages of all these approaches is that they are using only Unsupervised methods, which means new data can be added to the database can be tested for outliers in future in an efficient manner. Experiments showed that CLARANS is the best algorithm while considering outlier detection, followed by CLARA and PAM.

Aggarwal C.C, Yu P.S., [1] they have addressed the issue of privacy preserving data mining. Specifically, they consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Their work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multi-party computation and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and, furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. They focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

Sheng-yi Jiang., Qing-bo- An., [23] they generalize the concept of outlier factor of object to the case of cluster and put forward a clustering-based outlier detecting method. They regard the cluster which comes by clustering process as a unit and identify it as “normal” or “outlier” (the id of a cluster is also the id of its objects). The method is made up of two stages: the first stage is grouping dataset with clustering algorithm; the second stage is to identify the gained clusters as “normal cluster” or “outlier cluster” according to their outlier factor. The goal of clustering is that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. Many efficient clustering algorithms have been proposed by the database research community. Clustering algorithm can be selected according to data, objective of clustering and application. In this paper, they use one-pass clustering algorithm divide dataset into hyper spheres with almost the same radius. On the basis of the outlier factor of cluster, they present a clustering-based outlier detection method (*CBOD*), which consists of two stages. The theoretical analysis explains that the time complexity of *CBOD* is nearly linear with the size of dataset, the number of attributes and the final number of clusters, *CBOD* suit to detect outlier in large dataset. Finally, they give some experimental results to demonstrate the effectiveness.

E. M. Knurs., and R. T. Ng., [16] assumes data is distributed. The data are allowed to use, but disclosing it to others is a privacy violation. The problem is to find distance-based outliers without any party gaining knowledge beyond learning which items are outliers. Ensuring that data is not disclosed maintains privacy, i.e., no privacy is lost beyond that inherently revealed in knowing the outliers. Even knowing which items are outliers need not be revealed to all parties, further preventing privacy breaches. The approach duplicates the results of the outlier detection algorithm. The idea is that an object o is an outlier if more than a percentage p of the objects in the data set is farther than distance d from o. The basic idea is that parties compute the portion of the answer they know, and then engage in a secure sum to compute the total distance. The key is that this total is (randomly) split between sites, so nobody knows the actual distance. A secure protocol is used to determine if the actual distance between any two points exceeds the threshold; again the comparison results are randomly split such that summing the splits (over a closed field) results in a 1 if the distance exceeds the threshold or a 0 otherwise. For a given object o, each site can now sum all of its shares of comparison results (again over the closed field). When added to the sum of shares from other sites, the result is the correct count; all that remains is to compare it with the percentage threshold p. This addition/comparison is also done with a secure protocol, revealing only the result, if o is an outlier.

4. PROBLEM FORMULATION AND METHODOLOGY

The main objective of this research work is, detecting the outliers by applying clustering algorithms. The outliers detected are considered as sensitive outliers. Protecting the sensitive outliers 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 outlier detection.Now, verify whether outliers are detected or not. The performance of the clustering algorithms for outlier detection and the privacy technique are analyzed.

4.1. Methodology

1. Input the Data set
2. Preprocessing
3. Outlier detection
   3.1 Existing Algorithms
   - PAM
   - CLARA
   - CLARANS
   - ECLARANS
4. Outlier Protection
   4.1 GAUSSIAN PERTURBATION RANDOM Method

4.2 Dataset as Input

Breast Cancer Wisconsin and Pima Indians Diabetes Data Set are used for outlier detection and outlier protection. These datasets are collected from http*://archive.ics.uci.edu/ml/datasets.html*.

4.2.1 Breast Cancer Wisconsin Dataset

This dataset consists of 699 instances and 10 attribute. The dataset characteristics are Multivariate. The attribute characteristics are Integer.

4.2.2 Pima Indians Diabetes Data Set

This dataset consists of 699 instances and 10 attribute. The dataset characteristics are Multivariate. The attribute characteristics are Integer.

4.3 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. Data cleansing is used mainly in databases and it refers to identifying incomplete, incorrect, inaccurate, irrelevant parts of the data and missing data can be replaced, modified, deleted.

Data cleansing may involve removing typographical errors and validating and correcting values against a known list of entities. The validation may be strict such as rejecting any address that does not have a valid postal code or fuzzy such as correcting records that partially matches the existing records. After cleansing, a data in a database will be consistent to different sets of data that have been merged from separate databases. Sophisticated software applications are available to clean a database's data using algorithms, rules, and look-up-tables, a task that was once prepared manually and therefore still subject to human error. In this research, k-nearest neighbor technique is used for preprocessing.

4.4 An Approach for Outlier Detection

Outliers detection is an outstanding data mining task, referred to as outlier mining. Outliers are objects that do not comply with the general behavior of the data. By definition, outliers are rare occurrences and hence represent a small portion of the data. The following clustering algorithms are used for detecting the outliers.

4.4.1 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 in contains two phases Build phase and Swap phase [1].

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

Swap phase: Calculates the total cost for each pair of selected and non-selected object.

PAM Procedure:

1. Input the dataset D
2. Randomly select k objects from the dataset D
3. Calculate the Total cost T for each pair of selected Si and non selected object Sh
4. For each pair if T si < 0, then it is replaced Sh
5. Then find similar medoid for each non-selected object
6. Repeat steps 2, 3 and 4, until find the medoids.

4.4.2 Clara (Clustering Large Applications)

CLARA is introduced to overcome the problem of PAM. This works in larger data set than PAM. This method takes only a sample of data from the data set instead of taking full data set. It randomly selects the data and chooses the medoid using PAM algorithm [1].

CLARA Procedure:

1. Input the dataset D
2. Repeat n times
3. Draw sample S randomly from D
4. Call PAM from S to get medoids M.
5. Classify the entire dataset D to Cost1.....costk
6. Calculate the average dissimilarity from the obtained clusters

4.4.3 Clarans (Clustering Large Applications Based On Randomized Search):

This method is similar to PAM and CLARA. It starts with the selection of medoids randomly. Draws the neighbour dynamically. It checks "maxneighbour" for swapping. If the pair is negative then it chooses another medoid set. Otherwise it chooses current selection of medoids as local optimum and restarts with the new selection of medoids randomly. It stops the process until returns the best.

CLARANS Procedure:

1. Input parameters numlocal and maxneighbour.
2. Select k objects from the database object D randomly.
3. Mark these K objects as selected Si and all other as non-selected Sh.
4. Calculate the cost T for selected Si
5. If T is negative update medoid set. Otherwise selected medoid chosen as local optimum.
6. Restart the selection of another set of medoid and find another local optimum.
7. CLARANS stops until returns the best.

4.4.4 Enhanced Clarans (Eclarans)

This method is different from PAM, CLARA AND CLARANS.Thus method is produced to improve the accuracy of outliers. ECLARANS is a new partitioning algorithm which is an improvement to CLARANS form clusters with selecting proper arbitrary nodes instead of selecting as random searching operations. The algorithm is similar to CLARANS but these selected arbitrary nodes reduce the number of iterations of CLARANS

ECLARANS Procedure

1. Input parameters numlocal and maxneighbour. Initialize i to 1, and mincost to a large number.
2. Calculating distance between each data points
3. Choose n maximum distance data points
4. Set current to an arbitrary node in n: k
5. Set j to 1.
6. Consider a random neighbor S of current, and based on 6, calculate the cost differential of the two nodes.
7. If S has a lower cost, set current to S, and go to Step 5.
8. Otherwise, increment j by 1. If j maxneighbour, go to Step 6.
9. Otherwise, when j > maxneighbour, compare the cost of current with mincost. If the former is less than mincost, set mincost to the cost of current and set best node to current.
10. Increment i by 1. If i > numlocal, output best node and halt. Otherwise, go to Step 4.

4.5 Privacy Technique Based On Gaussian Perturbation Random Method

Data perturbation is a form of privacy-preserving data mining for electronic health records (EHR). There are two main types of data perturbation appropriate for EHR data protection. The first type is known as the probability distribution approach and the second type is called the value distortion approach. Data perturbation is considered a relatively easy and effective technique in for protecting sensitive electronic data from unauthorized use. Data perturbation has been hailed as a more effective application of data protection in health care than de-identification/re-identification due to the higher probability that attacks could take place which link public data sets to original identifiers or subjects. Data perturbation is hailed as a more solid application when it comes to EHR security.

The probability distribution approach takes the data and replaces it from the same distribution sample or from the distribution itself. The value distortion approach perturbs data by multiplicative or additive noise, or other randomized processes. It is considered to be more effective than the former type of perturbation. This approach builds decision tree classifiers where each element is assigned random noise from the Gaussian distribution, for instance. By data mining, the original data distribution is rebuilt from its perturbed version. However, critics point to the fact that random additive noise can be filtered which can result in EHR privacy compromises.

After the outlier detection process, we consider the outlier detected as the sensitive information that is to be protected. Thus the next step is to propose an approach for preserving privacy to the sensitive information. Here we are proposing a new approach of privacy technique based on the Gaussian Perturbation Random Method. The main objective of this proposed approach is to perform a rounding technique to the outlier information and thus provides a protection to the data. This method is to initialize their ensemble for enhanced CLARANS data representation.

4.5.1. Data perturbation by Gaussian Perturbation Random Method

Input: Dataset Objects
Output: Perturbed dataset

Method

1. Consider the sensitive outliers from the data set
2. Select the clusters one by one
   - 2.1 Consider the data items
   - 2.2 Combine the nearest outliers and cluster data items
   - 2.3 Find out the mean value of outliers and mean value of cluster data items
   - 2.4 Find out the difference between outliers mean and the cluster data items mean
   - 2.5 Select the outliers randomly and perturb the outlier value by add or subtract with the difference
3. Repeat the same process for all the clusters.

Thus by applying the Gaussian Perturbation Random method could be able to protect the sensitive outlier information. Later the dataset is modified based on the above result of preserved outlier information. Now, the ECLARANS algorithm is applied to the modified dataset in order to detect the outliers. All the sensitive outliers are protected and no outlier is detected by the ECLARANS algorithm.

5. EXPERIMENTAL RESULTS AND DISCUSSION

This research work has implemented in MATLAB 7.10(R2010a) and executed in the processor Intel(R) Core(TM) i3 CPU M370 @ 2.40 GH_Z 2.40 GH_Z. Breast Cancer Wisconsin and Pima Indians Diabetes Data Set are used for outlier detection and outlier protection

5.1 Outlier Accuracy

Outlier detection accuracy is calculated, in order to find out more number of outliers detected by the clustering algorithms PAM, CLARA, CLARANS and ECLARANS.

Table 5.1.1 Number Of Outliers Detected.

DATA SET	PAM	CLARA	CLARANS	ECLARANS
PIMA INDIAN DIABETES	6	38	219	286
WISCONSIN BREAST CANCER	109	130	314	319

It could be seen that the ECLARANS algorithm has produced more outliers compared with PAM, CLARA and CLARANS. In Pima, PAM detected 6 outliers, CLARA detected 38 outliers, CLARANS detected 219 outliers, and ECLARANS detected 286 outliers, the same algorithms are used for detecting outliers in cancer dataset is also shown in Table1.Thus it could be shows that the ECLARANS algorithm improves the accuracy of detecting the outliers.

Algorithms Performance of Accuracy

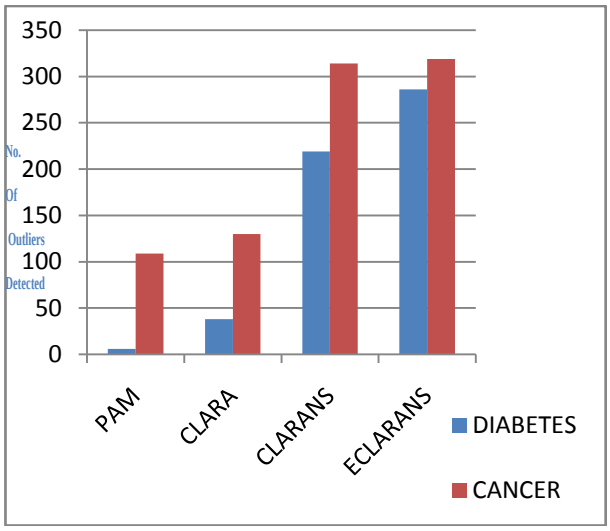

Figure 1: Outlier Accuracy

The chart shows that the number of outliers detected by the existing clustering algorithms PAM, CLARA, CLARANS and the proposed clustering algorithm ECLARANS. The new proposed clustering technique ECLARANS has detected more number of outliers compared to the existing techniques.

5.2 Time Complexity Of Clustering Algorithms

The Time complexity performance factor is measured in terms of the time required for detecting the outliers by the clustering algorithms PAM, CLARA, CLARANS and ECLARANS.

Table 5.2.1 Time Complexity

DATASETS	PAM (in secs)	CLARA (in secs)	CLARANS (in secs)	ECLARANS (in secs)
PIMA INDIAN DIABETES	238.92	269.41	3.73	30.97
WISCONSIN BREAST CANCER	374.52	206.32	3.47	16.06

Performance of Time Complexity

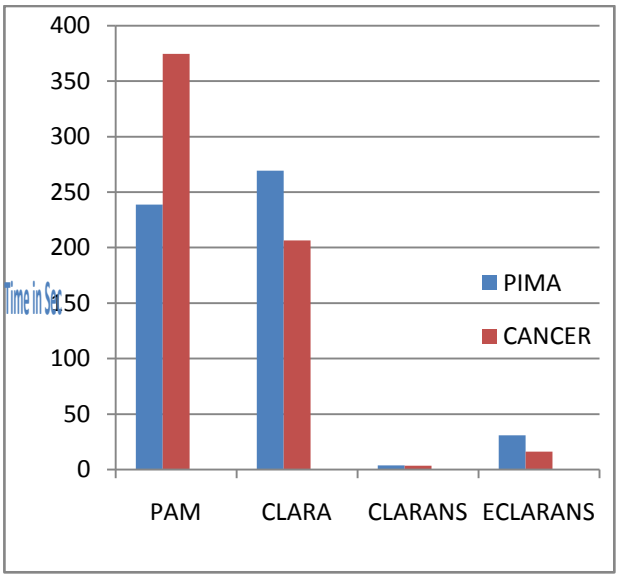

Figure 2: Time Complexity

Comparing the performance of time complexity of the clustering algorithms, the CLARANS algorithm has taken less time.

5.3 Results of Outlier Protection

This proposed approach Gaussian Perturbation Random Method is to perform a rounding technique to the outlier information and thus provides a protection to the data. This method is to initialize their ensemble for enhanced CLARANS data representation.

Outlier Detection Using Eclarans clustering In Diabetes Data Set

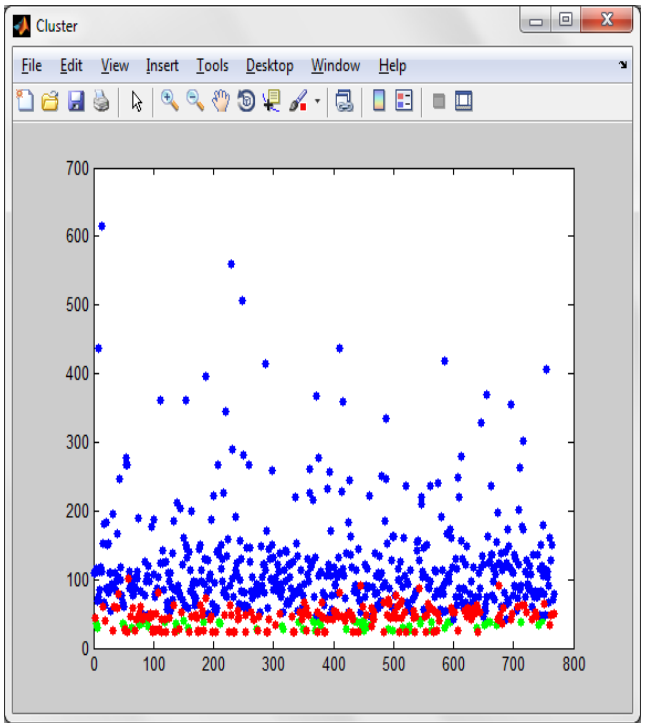

Figure 3: Outlier Detection

Protected Outliers Using Privacy Technique in Diabetes Data set

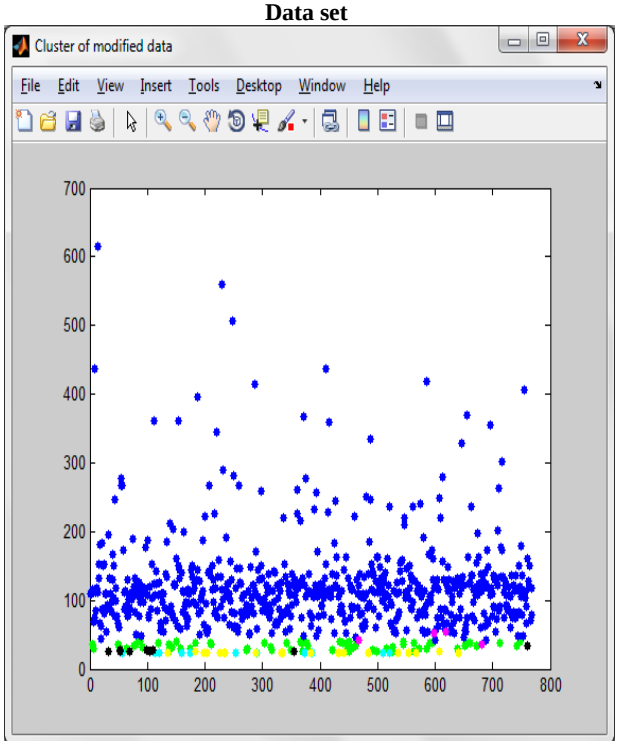

Figure 4: Protected Outliers

Outlier Detection Using Eclarans clustering In Cancer Data Set

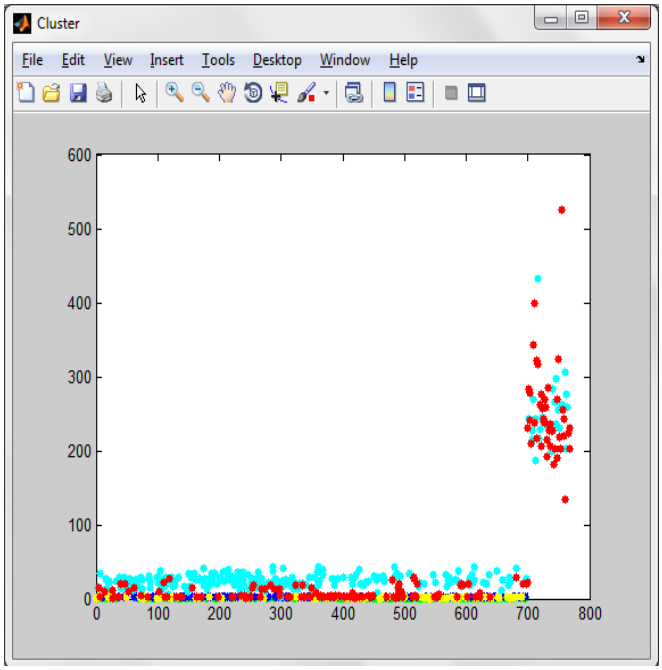

Figure5: Outlier Detection

Protected Outliers Using Privacy Technique

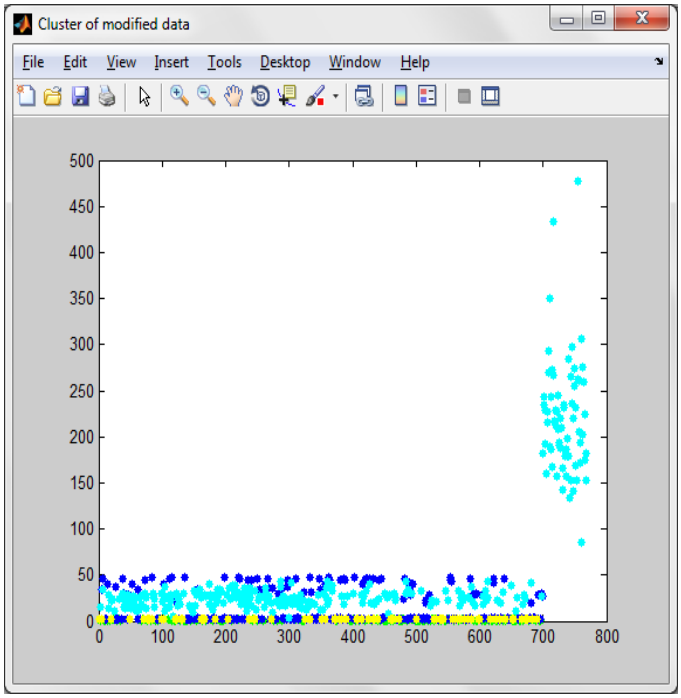

Figure 6: Protected Outliers

5.3 Time Complexity of Privacy Technique

Time has taken by the Gaussian Perturbation Random Method for the protection of outliers.

Table 5.3.1 Time Complexity

DATASETS	TIME COMPLEXITY(in seconds)
DIABETES DATA SET	0.02
CANCER DATA SET	0.01

Time Complexity Performance of Privacy Technique

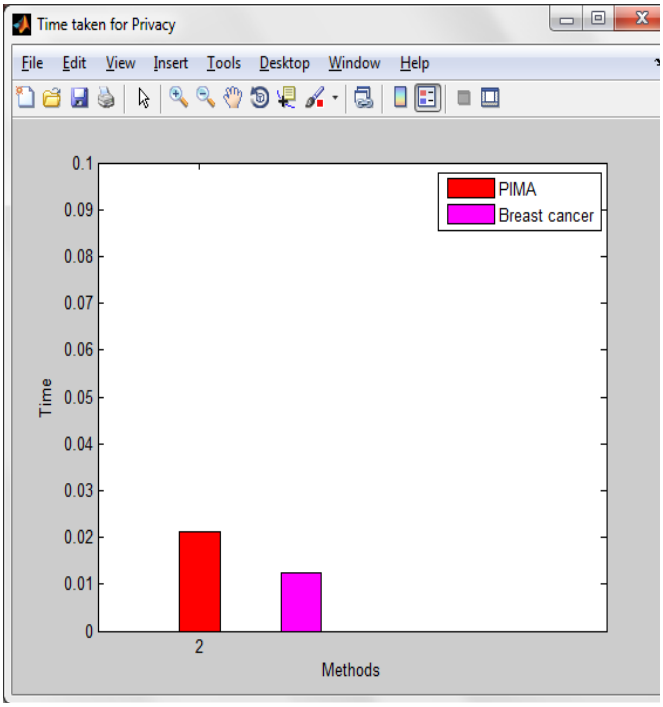

Figure7: Time Complexity

The figures show that the results of modified data (i.e.) protected the sensitive outliers (produced by ECLARANS algorithm) information of two data sets using a privacy technique Gaussian Perturbation Random Method.

6. CONCLUSION

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 four clustering algorithms PAM, CLARA, CLARANS and ECLARANS for detecting the outliers in the health datasets. The outliers given by the ECLARANS algorithm is considered as sensitive outliers. These sensitive outliers are protected by a proposed privacy technique Gaussian Perturbation random method. Different performance factors are used for measuring the efficiency of the clustering algorithms and the privacy technique Gaussian Perturbation random method. 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.

7. REFERENCES

[1] Aggarwal C.C, Yu P.S., "Models and Algorithms: Privacy-Preserving Data Mining", Springer, 2008.

[2] Ajay Challagalla,S.S.Shivaji Dhiraj ,D.V.L.N Somayajulu,Toms Shaji Mathew,Saurav Tiwari,Syed Sharique Ahmad " Privacy Preserving Outlier Detection Using Hierarchical Clustering Methods" 2010 34th Annual IEEE Computer Software and Applications Conference Workshops.

[3] 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

[4] Al-Zoubi, M. (2009) "An Effective Clustering-Based Approach for Outlier Detection", European Journal of Scientific Research.

[5] "An Effective Clustering-Based Approach for Outlier Detection", Moh'd Belal Al- Zoubi, European Journal of Scientific Research, ISSN 1450-216X Vol.28 No.2 (2009).

[6] Antonio Loureiro, Luis Torgo, and Carlos Soares, "Outlier Detection Using Clustering Methods: a data cleaning application", in Proceedings of the Data Mining for Business Workshop, 2009.

[7] Elisa Bertino , Dan Lin and Wei Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms", in Privacy-Preserving Data Mining (Models and Algorithms), Charu C. Aggarwal and Philip S. Yu (Eds.), Springer-Verlag, 2008.

[8] E. M. Knorr and R. T. Ng. "Algorithms for mining distance based outliers in large datasets". In Proceedings of 24th International Conference on Very Large Data Bases (VLDB 1998), New York City, NY, USA, Aug.24-27 1998.

[9] Friedman A., Wolff R., Schuster A. "Providing *k* - anonymity in data mining", The VLDB Journal , Vol.17 ,2008.

[10] Jaideep Vaidya, Chris Clifton, W.Lafayette "Privacy Preserving Data Mining" Springer 2006.

[11] Jeffrey W. S., "Data mining: An overview", CRS report RL 31798.

[12] Jiang, S. And An, Q. (2008), "Clustering Based Outlier Detection Method", Fifth International Conference on Fuzzy Systems and Knowledge Discovery.

[13] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006.

[14] John Peter.S., Department of computer science and research center St.Xavier's College, Palayamkottai, "An Efficient Algorithm for Local Outlier Detection Using Minimum

Spanning Tree", International Journal of Research and Reviews in Computer Science (IJRRCS), March 2011.

[15] Jyothsna R.Nayak and Diane J.Cook, "Approximate Association Rule Mining", the Florida AI Research Society Conference FLAIRS, 2001.

[16] Knurs, E.M. and Ng, R.T. (1998) Algorithms for mining Distance-based outliers in Large Datasets, VLDB

[17] Liu, H., Shah, S. and Jiang, W. (2004) "On-line outlier detection and data cleaning", Computers and Chemical Engineering .

[18] Loureiro,A., Torgo, L. and Soares, C. (2004) "Outlier Detection using Clustering Methods: a Data Cleaning Application", in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[19] Mahfouz, M.A. and Ismail, M.A. (2009)" Fuzzy relatives of the CLARANS algorithm with application to text clustering", World Academy of Science, Engineering and Technology.

[20] Murugavel. P. et al, "Improved Hybrid Clustering And Distance-Based Technique for Outlier Removal", International Journal on Computer Science and Engineering (IJCSE), 1 JAN 2011

[21]]Poovammal E., Ponnavaikko M., "An Improved Method fo r Privacy Preserving Data Mining", International Advance Computing Conference, 2009.

[22] Samarati P, Sweeney L. Protecting "Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression". IEEE Symp. on Security and Privacy, 1998.

[23] Sheng-yi Jiang, Qing-bo- An, "Clustering-Based Outlier Detection Method", Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'08, 2008.

[24] Sweeney L. "AI Technologies to Defeat Identity Theft Vulnerabilities". AAAI Spring Symposium, AI Technologies for Homeland Security, 2005.

[25] Sweeney L. "Privacy Technologies for Homeland Security". Testimony before the Privacy and Integrity Advisory Committee of the Department of Homeland Security, Boston, MA, June 15, 2005.

[26] Zenyou He *,Xiaofei Xu, Shenchun Deng ., "Discovering Cluster Based Local Outliers", Department of computer science and Engineering, Harbin Institute of Technology, Harbin 150001,P.R.China.