ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Privacy Feedback System Using Data Mining and Outlier Detection Algorithm

Savita Lohiya, Priyanka Ramayi, Ashwin Pillai, SriramVeturi

Assistant Professor, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

ABSTRACT: In this project, we intend to provide an elimination of a third party requirement in order to maintain the confidentiality in the organization Data is shared across various organizations at different levels in business, marketing, hospital and entertainment domains. When any data is shared between organizations, there is a possibility that there could be some exceptional data that is different from the rest of the dataset in terms of behaviour, property and value. Considering the data related to the hospitals, there could be such data which deviates from the rest of the data, which would result in incorrect medical assumptions and computations. We propose a method called Privacy Feedback System in Data Mining. First we detect the outliers from the original data. Then, we notify the dataset owner about the presence of the outliers and the characteristics of the outliers so that the dataset owner has an idea which parts of the computation can get affected due the presence of theoutliers and can accordingly manipulate the data to get the desired results from the computational processes. The sensitive information which reveals the dataset owner's identity would be hidden from the world by making the sensitive attributes anonymous. This would preserve the privacy of the dataset owner as well as work as a feedback system which notifies the user about the anomalies present and their significance in the dataset.

KEYWORDS: Privacy preserving; confidentiality; k-Anonymity algorithm; Anomaly Detection; feedback system; Median based outlier detection algorithm.

I. INTRODUCTION

In statistics, outliers are considered as the anomaly which behaves differently from the rest of the dataset. The privacy feedback system ensures that outliers are detected from the dataset provided by the user and provides the user with the feedback regarding the behavior of the outliers detected. The privacy feedback system is a web based application that anonymizes the confidential data and provides only the necessary information to the people who require to access the dataset. The system would ask the hospital organizations to store the dataset with the information of the patients. The information may include the personal data about the patient like name, age, gender, address and phone number. The system would ask the hospital organization about the parameters and attributes which need to be anonymous. The attributes and the parameters would then be made anonymous by displaying the character '*' in place of the data element. This dataset with hidden confidential information would be provided to the research organizations for further computations.

II. RELATED WORK

For better accuracy of the original data, the data has undergo randomization and generalization, also it can reconstruct the original data and provide data with no information loss, ensures usability of the original data. Thus, the data when transmitted, the sensitive data is hidden and there is scope of the third part interceptions in the sensitive data transmission[1].

[2],This paper proposes a novel "knowledge-free" anomaly detection method for spacecraft based on Kernel Feature Space and directional distribution, which constructs a system behaviour model from the past normal telemetry data

DOI: 10.15680/IJIRCCE.2017. 0503106

from a set of telemetry data in normal operation and monitors the current system status by checking incoming data with the model.
[3]One of the commonest ways of finding outliers in one-dimensional data is to mark as a potential outlier any point that is more than two standard deviations, say, from the mean .[4] But the presence of outliers is likely to have a strong effect on the mean and the standard deviation, making this technique unreliable.

III. PROPOSED SYSTEM

A. *Design Considerations:*

- The Privacy Feedback system allows the hospital organization to upload the data sets in the web application portal.
- The system would ask the organization to input the attributes that need to be anonymous while displaying the dataset to the research organizations.
- The system would process the data with the specified attributes and replace the data elements with "*" and displays the anonymized dataset using k-anonymity algorithm.
- The system would process the dataset more for the detection of outliers with the Median based Extreme Outlier detection algorithm**.**
- The dataset with the outliers are displayed to the user in the form of a feedback.

B. *Features of the System:*

- Increased reliability.
- Low cost and maintenance.
- Low error rate.
- High feasibility.

C. *Block Diagram:*

The two main phases incorporated in the system are data set uploading and anonymization and anomaly detection.

1) Dataset Uploading

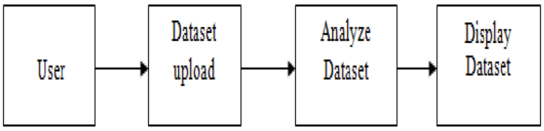


Fig.1. Dataset Uploading

The diagram given above depicts the first stage of the user that is Data Uploading wherein user has to load its desktop or online dataset into the application.The user's dataset will be loaded and displayed in the application in a tabular form which helps in easy recognition of dataset.

2) Anonymization and Anomaly detection

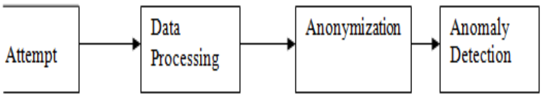


Fig.2. Anonymization and Anomaly detection

The diagram above depicts as to the anonymization and anomaly detection phase.
In this phase the user species the attributes that has to be camouflaged to ensure confidentiality. The selected attribute is anonymized and the anonymized dataset is displayed. Also the user is made vigil about the vulnerable or deviated

dataset by detecting anomalies. Finally the dataset with anomalies is displayed in the form of a feedback with anomalies highlighted thus providing easy recognition to the user.

D. *DFD Diagrams*

1. DFD level 0:

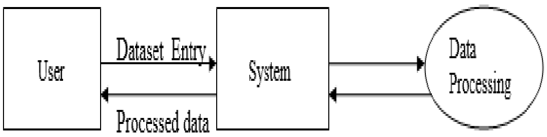


Fig.3. DFD level 0

This figure depicts the first data flow diagram. The initial phase shows that the user inputs his dataset into the system. The system then performs various data processing technique to ensure confidentiality. The result of this processing is provided to the user in the form of feedback.

2. DFD level 1:

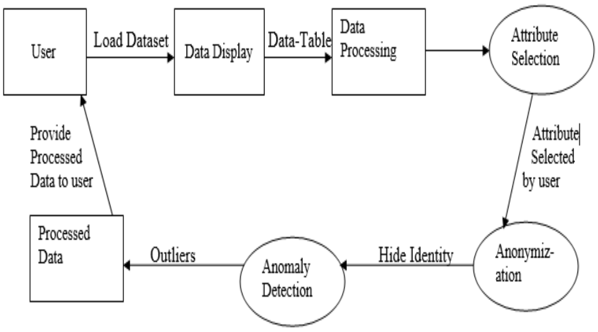


Fig.4. DFD level 1

This figure depicts the second data flow diagramdataset into the system. After loading the dataset the user provides the attributes from the dataset that incorporates sensitive details and is required to be camouflaged. The system software hides these details using anonymization algorithm wherein the details are replaced by special characters such as ‘*’. The anonymized dataset is then further processed for detection of any vulnerable or errorneous data or presence of any anomalies using outlier detection algorithm. Finally the processed dataset is then returned to the user in the form of a feedback.

E. *System Algorithm:*
The system includes two algorithms for its efficient functioning.

k-anonymity algorithm:
Data anonymization is the use of one or more techniques designed to make it impossible or at least more difficult to identify a particular individual. The k-anonymity algorithm provides a formal protected model to ensure confidentiality of dataset.

Unless otherwise stated, the term data refers to person-specific information that is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a tuple. A tuple contains a relationship among the set of values associated with a person. Tuples within a table are not necessarily unique. Each column is called an attribute and denotes a field or semantic category of information that is a set of possible values; therefore, an attribute is also a domain. Attributes within a table are unique. So by observing a table, each row is an ordered n-tuple of values <d1, d2, ...,dn> such that each value dj is in the domain of the j-th column, for j=1, 2, ..., n where n is the number of columns. In mathematical set theory, a relation corresponds with this tabular presentation, the only difference is the absence of column names.

Definition 1. Attributes
Let B (A1,...,An) be a table with a finite number of tuples. The finite set of attributes of B are {A1,...,An}. Given a table B(A1,...,An), {Ai,...,Aj} ⊆ {A1,...,An}, and a tuple t∈B, I use t[Ai,...,Aj] to denote the sequence of the values, vi,...,vj, of Ai,...,Aj in t. I use B[Ai,...,Aj] to denote the projection, maintaining duplicate tuples, of attributes Ai,... Aj in B. Throughout the remainder of this work each tuple is assumed to be specific to one person and no two tuples pertain to the same person. This assumption simplifies discussion without loss of applicability.

Definition 2. Quasi-identifier: Given a population of entities U, an entity-specific table T(A1,...,An), fc: U →T and fg: T → U', where U ⊆ U'. A quasi-identifier of T, written QT, is a set ofattributes {Ai,...,Aj} ⊆ {A1,...,An} where: ∃pi∈U such that fg(fc(pi)[QT]) = pi.

Example 1. Quasi-identifier: Let V be the voter-specific table described earlier in Figure 1 as the voter list. A quasi-identifier for V, written QV, is {name, address, ZIP, birth date, gender}.

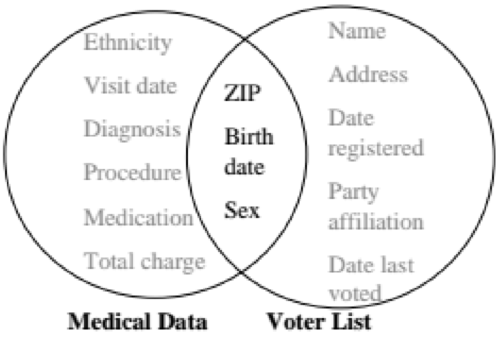


Fig.5. Linking to re-identify data

Linking the voter list to the medical data as shown in Figure 1, clearly demonstrates that {birth date, ZIP, gender} ⊆ QV. However, {name, address} ⊆QV because these attributes can also appear in external information and be used for linking. In the case of anonymity, it is usually publicly available data on which linking is to be prohibited and so attributes which appear in private data and also appear in public data are candidates for linking; therefore, these attributes constitute the quasi-identifier and the disclosure of these attributes must be controlled. It is believed that these attributes can be easily identified by the data holder.

Assumption (quasi-identifier).
The data holder can identify attributes in his private data that may also appear in external information and therefore, can accurately identify quasi-identifiers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

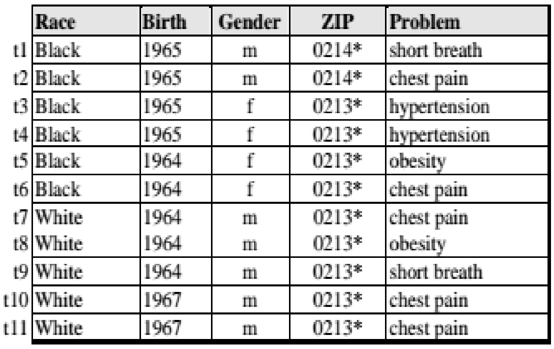
Vol. 5, Issue 3, March 2017

To determine how many individuals each released tuple actually matches requires combining the released data with externally available data and analysing other possible attacks. Making such a determination directly can be an extremely difficult task for the data holder who releases information. Although I can assume the data holder knows which data in PT also appear externally, and therefore what constitutes a quasi-identifier, the specific values contained in external data cannot be assumed. I therefore seek to protect the information in this work by satisfying a slightly different constraint on released data, termed the k-anonymity requirement. This is a special case of k-map protection where k is enforced on the released data.

Definition 3. k-anonymity

Let RT(A1,...,An) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT[QIRT] appears with at least k occurrences in RT[QIRT].

Table1: Example of k-anonymity where k=2 and QI={Race, Birth, Gender, ZIP}.



	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Example 3. Table adhering to k-anonymity

Table 1 provides an example of a table T that adheres to k-anonymity. The quasi-identifier for the table is QIT= {Race, Birth, Gender, ZIP} and k=2. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T. That is, for each sequence of values in T[QIT] there are at least 2 occurrences of those values in T[QIT]. In particular, t1[QIT] = t2[QIT], t3[QIT] = t4[QIT], t5[QIT] = t6[QIT], t7[QIT] = t8[QIT] = t9[QIT], and t10[QIT] = t11[QIT].

Lemma.

Let RT(A1,...,An) be a table, QIRT =(Ai,..., Aj) be the quasi-identifier associated with RT, Ai,...,Aj⊆ A1,...,An, and RT satisfy k-anonymity. Then, each
sequence of values in RT[Ax] appears with at least k occurrences in RT[QIRT] for x=i,...,j.

Example 4. k occurrences of each value under k-anonymity

Table T in Table 1 adheres to k-anonymity, where QIT= {Race, Birth, Gender, ZIP} and k=2. Therefore, each value that appears in a value associated with an attribute of QI in T appears at least k times. |T[Race ="black"]| = 6. |T[Race ="white"]| = 5. |T[Birth ="1964"]| = 5. |T[Birth ="1965"]| = 4. |T[Birth ="1967"]| = 2. |T[Gender ="m"]| = 6. |T[Gender ="f"]|= 5. |T[ZIP ="0213*"]| = 9. And, |T[ZIP ="0214*"]| = 2. It can be trivially proven that if the released data RT satisfies k-anonymity with respect to the quasi-identifier QIPT, then the combination of the released data RT and the external sources on which QIPT was based, cannot link on QIPT or a subset of its attributes to match fewer than k individuals. This property holds provided that all attributes in the released table RT which are externally available in combination (i.e., appearing together in an external table or in a possible join of external tables) are defined in the quasi-identifier QIPT associated with the private table PT. This property does not guarantee individuals cannot be identified in RT; there may exist other inference attacks that could reveal the identities of the individuals contained in the data. However, the property does protect RT against inference from linking (by direct matching) to known external sources; and in this context, the solution can provide an effective guard against re-identifying individuals.

Median based Extreme Outlier detection Algorithm:

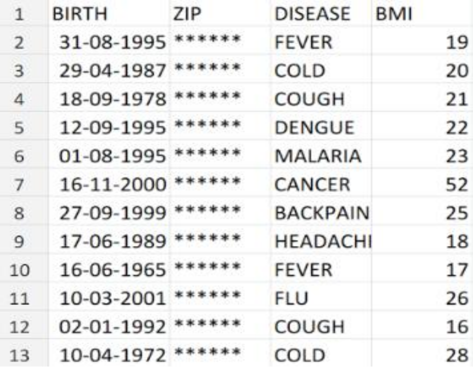
The medical records presented by the hospital organizations contain many numerical data attributes. In such one-dimensional data, outliers can be identified using the mean and the standard deviations of the univariate data set. However, the mean and the standard deviations are hugely affected by the outliers. Thus, the approach of identifying the anomalies is eliminated. Irrespective of the values of the outliers, the median of any data set is unaffected. The median approach is said to be the best method to identify the anomalies

The algorithm has its base on one assumption. This method of outlining the anomalies assumes that the medical records with numerical data contain the deviating data at the extreme ends of the data set, when arranged in a sequence. Following are the steps of the algorithm:

1. The column or the attribute in which the anomalies are to be detected is identified.
2. It is mandatory to have the numerical data arranged in a sequence (smallest to the largest or largest to the smallest).
3. Every element in the data set sequence is indexed with a number as N1, N2, ..,Nn.
4. The median (M), smallest(S) and the largest (L) element in the column of the numerical set of data are evaluated.
5. The number of elements between the smallest element and the median is calculated and denoted by D1.
6. Similarly, the number of elements between the median and the largest element is calculated and denoted by D2.
7. D1 and D2 are to be compared and checked whether which is greater.
8. After comparing, the distance which is smaller is placed over the larger one from the median index and the remaining extra data elements are considered as outliers.

Example of the median based extreme anomaly detection algorithm.

Consider a data set with an attribute BMI (Body Mass Index) with numerical input values. The Body Mass Index of any person lies between the range of 15 to 40. Any BMI value outside this range would be considered as an outlier. Consider a data set with 13 patients with BMI values as follows:

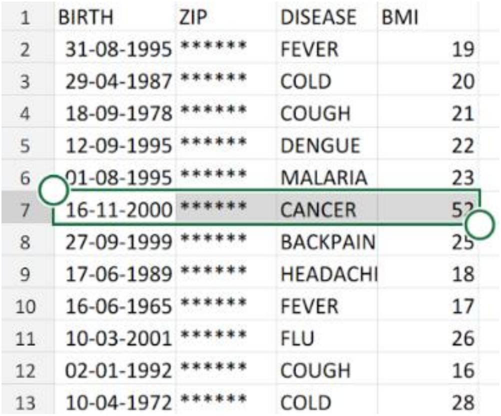


1	BIRTH	ZIP	DISEASE	BMI
2	31-08-1995	******	FEVER	19
3	29-04-1987	******	COLD	20
4	18-09-1978	******	COUGH	21
5	12-09-1995	******	DENGUE	22
6	01-08-1995	******	MALARIA	23
7	16-11-2000	******	CANCER	52
8	27-09-1999	******	BACKPAIN	25
9	17-06-1989	******	HEADACHI	18
10	16-06-1965	******	FEVER	17
11	10-03-2001	******	FLU	26
12	02-01-1992	******	COUGH	16
13	10-04-1972	******	COLD	28

Fig.6: Anonymized Dataset

The median, smallest and the largest elements are calculated as 22, 16 and 52 respectively. The distances are calculated as:
D1= Median - Smallest=22-16=8.
D2=Largest - Median=52-22=30.

Clearly, D2 is larger than D1. Thus, the distance of D1, 8 is placed after the median and the remaining elements are considered as outliers. From the data set, any value above Median + D1 is an outlier, which is 52 as shown in the figure:



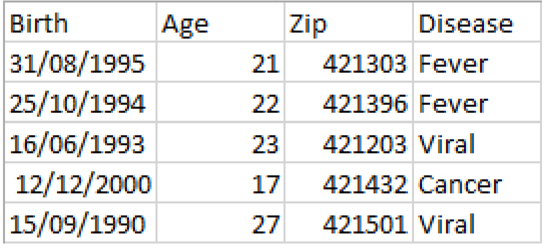
1	BIRTH	ZIP	DISEASE	BMI
2	31-08-1995	******	FEVER	19
3	29-04-1987	******	COLD	20
4	18-09-1978	******	COUGH	21
5	12-09-1995	******	DENGUE	22
6	01-08-1995	******	MALARIA	23
7	16-11-2000	******	CANCER	52
8	27-09-1999	******	BACKPAIN	25
9	17-06-1989	******	HEADACHI	18
10	16-06-1965	******	FEVER	17
11	10-03-2001	******	FLU	26
12	02-01-1992	******	COUGH	16
13	10-04-1972	******	COLD	28

Fig.7: Anonymized dataset with Anomalies detected

IV. SIMULATION RESULTS

The following figure depicts the original medical dataset that should be uploaded by the user

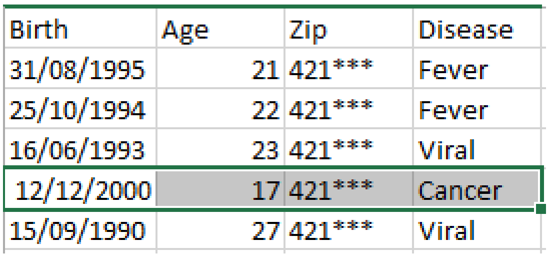
Table.2: Original Dataset



Birth	Age	Zip	Disease
31/08/1995	21	421303	Fever
25/10/1994	22	421396	Fever
16/06/1993	23	421203	Viral
12/12/2000	17	421432	Cancer
15/09/1990	27	421501	Viral

The following figure depicts the feedback that will provided to the user

Table.3: Feedback Dataset



Birth	Age	Zip	Disease
31/08/1995	21	421***	Fever
25/10/1994	22	421***	Fever
16/06/1993	23	421***	Viral
12/12/2000	17	421***	Cancer
15/09/1990	27	421***	Viral

Here in this example the values of QI{Disease} is repeated with a set of different zip codes. The digits of the zip code that are not similar are being anonymized using special characters and the deviated data in QI{Disease} is being heighted as an anomaly in the dataset.

If the parameters mentioned above are processed appropriately, the user gets the feedback of the dataset.

V. Conclusion and Future Scope

The project plays its role replacing the need ofexistence for a third party.Data set owners thenwould not be reluctant in using new technologies.Using sophisticated methods derived,a degree ofconfidence and reliability can be established for transferring datasets to those in need.

References

1) Savita Lohiya and LataRagha. ”Privacy Preserving in Data Mining Using Hybrid Approach”, published in Computational Intelligence and Communication Networks (CICN) IEEE Cnference, 2012 Fourth International Conference on 3-5 November, 2012.
2) Fujimaki, R., Yairi, T., and Machida, K. 2005., ”An approach to spacecraft outlier detection problem using kernel feature space”. In Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM Press, New York, NY, USA.
3)]Leys, C., et al., Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, Journal of Experimental Social Psychology, Volume 49, Issue 4, July 2013, pp. 764-766.
4)]Rousseeuw, P.J .and Croux C. (1993) Alternatives to the Median Absolute Deviation, Journal of the American Statistical Association, December 1993, pp. 1273-1283
5) L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty,Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
6) Study on K-anonymity Models of Sharing Medical Information, Yan ZHU Lin PENG, Research Centre for Contemporary Management, Tsinghua University,Beijing 100084, China.
7) Jian Wang,Yongcheng Luo, Yan Zhao,Jiajin Le. “A Survey on Privacy Preserving Data Mining”,2009 First International Workshop on Database Technology and Applications.

DOI: 10.15680/IJIRCCE.2017. 0503106