ISSN: 0976-3104
SPECIAL ISSUE [APSM]
Susan and Christopher

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

ANONYMISING THE SPARSE DATASET: A NEW PRIVACY PRESERVATION APPROACH WHILE PREDICTING DISEASES

V. Shyamala Susan* and T. Christopher
PG and Research Department of Computer Science, Government Arts College, Udumalpet, Tamil Nadu, INDIA

ABSTRACT

Data mining techniques analyze the medical dataset with the intention of enhancing patient's health and privacy. Most of the existing techniques are properly suited for low dimensional medical dataset. The proposed methodology designs a model for the representation of sparse high dimensional medical dataset with the attitude of protecting the patient's privacy from an adversary and additionally to predict the disease's threat degree. In a sparse data set many non-zero values are randomly spread in the entire data space. Hence, the challenge is to cluster the correlated patient's record to predict the risk degree of the disease earlier than they occur in patients and to keep privacy. The first phase converts the sparse dataset right into a band matrix through the Genetic algorithm along with Cuckoo Search (GCS).This groups the correlated patient's record together and arranges them close to the diagonal. The next segment dissociates the patient's disease, which is a sensitive value (SA) with the parameters that determine the disease normally Quasi Identifier (QI).Finally, density based clustering technique is used over the underlying data to create anonymized groups to maintain privacy and to predict the risk level of disease. Empirical assessments on actual health care data corresponding to V.A.Medical Centre heart disease dataset reveal the efficiency of this model pertaining to information loss, utility and privacy.

Received on: 19th-May-2016
Revised on: 21st -June-2016
Accepted on: 29th – June-2016
Published on: 29th – June-2016

KEY WORDS

Privacy Preservation, Clustering, sparse high dimensional dataset, band matrix, health care data, anatomisation, Genetic Algorithm (GA) and Cuckoo Search Algorithm (CSA).

***Corresponding author:** Email: shyamalasusan@gmail.com, **Tel:**+91-9916913776

INTRODUCTION

In the recent days, data mining techniques play a primary role within the healthcare domain [1] and medical industry, with the aim of improving the health and preserving the patient's privacy. Heart disorder is among the predominant causes of mortalities in several countries, inclusive of India. So there is a need for medical practitioners to predict heart disease earlier than they occur in patients. At the same, time anonymizing the heart disease data set is a task with a large undertaking given that of unstructured or semi-structured datasets. For instance, within the case of relational data (e.g., gender, age, BMI), records consists of only one value for every attribute. On the contrary, set-valued data (e.g., diagnostic codes and lab tests) have one or more values (cells) for every attribute.

Many data mining techniques such as naïve bayes algorithm, bagging algorithm, the neural network algorithm, decision tree, kernel density, k-mean clustering, etc were used for diagnosing of heart disease [2-4]. Anonymisation is managed through various techniques such as k-anonymity [5], l-diversity [6] and t-closeness [7]. T-closeness is the upgraded variant of k- anonymity and l- diversity qualities. These methods anonymise the dataset by generalization and attribute suppression, but the end result is the data loss. In the recent works, permutation technique stick to a procedure of Anatomy [8] where it groups the distinctive sensitive data from a transaction, and later permute them to dissociate the relationship. However, this method failed to provide its significance for high dimensional data.

The objective of this research attempts to be equipped to predict the probability of getting affected with heart disease with the given sparse patient dataset and to preserve the patient's disease from an adversary.

BACKGROUND STUDY

Health information systems have largely helped to increase the possibility of constructing the medical documents available to researchers, public health organizations, and the others who hold interest in medical data. However, health care data generally includes a large amount of patient privacy. Sharing this kind of data directly will be a great threat in the case of patient privacy. Hence it becomes a necessity for practical techniques to be developed

THE IIOAB JOURNAL
ISSN: 0976-3104

for balancing the healthcare data sharing and privacy preservation. In recent times, the relevance concerned with Privacy-Preserving Data Mining (PPDM) methods is analyzed thoroughly and studied by Mat win [9]. Usage of certain techniques showed their capability of avoiding the discriminatory utilization of data mining. Few techniques provided a proposal that any stigmatized group must not be focused over a number of generalization of data compared over the common population.

Recently, K Anonymization technique additionally utilized for ensuring the assurance of patient information [7] [9][10]. Once the security of protection is pertinent, the information is readied for investigation and the learning which helps choice making is extricated. It is confused, misfortune in utility of data [7] [9]. Zhu and Peng [11] initially depicted obviously the present state of China Restorative informatization level and clarified the need of cross-authoritative data sharing. [10] The specific issue known as connecting assault is additionally considered.
The K-anonymity is then again integrated with data mining technique for protecting the identity disclosure of the. Once the protection of privacy is applicable, the data is prepared for analysis and the knowledge which assists decision making is extracted. It is complicated, and there is loss in utility of data. Zhu and Peng [11] first described clearly the current condition of China Medical informatization level and explained the need of cross-organizational information sharing. The particular problem known as linking attack is also studied. Then a respective K Anonymization model is formulated along with Suppression techniques which are utilized for preserving privacy.

Machanavajjhala et al [6] proposed l-diversity which, not like k-anonymity, had the knowledge of the distribution of values pertaining to the sensitive attributes and regarding the impacts of background knowledge. l -diversity, a framework which provides stronger privacy assurance is employed to Inpatient Micro data that is gathered from adult dataset samples. They help in protecting the privacy of the user, either by making changes to quasi identifier values or through the addition of noise. The identity of patients has to be protected while the patient data are shared.

Gal et al [12] introduced a privacy model which is an upgraded version of K anonymity and l-diversity along with multiple sensitive attributes. Here the patient data set is got from the Kentucky Cancer Registry. But this model finds difficulty in distinguishing QIs and SAs.

T-closeness which is the upgraded variant of k- anonymity that is introduced by Soria-Comas et al [13] is highly correlated with ε- differential privacy. This approach is useful in improving the quality of anonymity and minimization of the loss of information for Patient Discharge Data.

Soria-Comas et al [14] introduced new kind of refinements with respect to k-anonymity, in which t-closeness performs better as the one providing with the guarantees of strictest privacy. It depends on generalization and suppression and the benefits of micro aggregation are examined, and then multiple micro aggregation algorithms for the purpose of k-anonymous t-closeness are proposed and later evaluated empirically.

Loukides and Gkoulalas-divanis [15] presented a new mechanism for anonymizing the data by meeting the data publishers' utilization requirements with a low information loss experience. A measurement of accurate information loss and an efficient anonymization algorithm are brought into use for minimizing the information losses. Experimental investigations related to on click-stream and medical data exposed that the new technique permitted more robust query answers rather than the state sophisticated techniques that are equal to them in terms of efficiency. The need for privacy is imposed by implementing a partition in the patient record dataset PR into sets patient records which are disjoint that are called as anonymized groups. The probability of the association of any kind of transaction in G with that item amounts to half [16].

CBA[17] presented an efficient technique to preserve privacy with minimal information loss by modifying prognosis codes and measured the data loss due to generalization and suppression and anonymise the patient's identity thro Clustering-established Anonymizer (CBA).However this clustering approach is not suitable for high dimensional data and suffers from high information loss and fails to protect against l-diversity. Biomedical researchers anonymise the data with improved utility, but it doesn't help in high dimensionality problem. In DRC[18] approach, utility metrics were particular to the data recipient's requirements. Utility loss is observed to increase with the variation in the data recipient's requirement.

The problem that exists with the available methods is that a big portion of the initial terms are typically absent from the anonymized dataset and every other method is applicable to low dimensional dataset samples. In Anatomy [10], the quasi identifiers are isolated from sensitive values and are provided protection against attribute disclosure. As it generates and issues the quasi identifiers directly it also renders a compromise over data utility.

THE IIOAB JOURNAL
ISSN: 0976-3104

Protection of privacy employing disassociation is observed as a complex issue in [10].There are only less works [18-19] which aid in preserving the original data, without the addition of noise, on the basis of an anatomy[10] idea.

PROBLEM SPECIFICATION

The proposed approach designs a model for the case of representation of sparse high dimensional medical data with a perspective of shielding the patient's privacy and in addition to aid the sufferers in having skills over the sickness's threat degree .The heart disease dataset comprises less number of SA that determines the risk stage of heart disease as well as QI values which are the parameters that identify the disease. As the dataset are arbitrarily distributed over the complete area, the task lies in efficiently grouping patients' records with similar QI values collectively to predict the hazard stage and anonymising the patient record with the disease to preserve privacy that has no sensitive value. The distinct contribution of this work is given beneath:

The first segment minimises the bandwidth of the sparse patient report by Genetic algorithm with Cuckoo Search (GCS).This permutes the and thus yields the adjacent rows correlated and brings them near to the diagonal. As it gets hold of the correlation well, it maximizes the utility and reduces the search space. When the patients' records are regrouped as a band matrix, the subsequent step is the construction of anonymised group of the patient with a view to guard the privacy and to predict the sickness. Anatomisation method is used over the regrouped band with the intention to dissociate the QI values with SA. This results in two tables which are the QIT and ST. Then the density based clustering algorithm anonymises the Sensitive attributes in ST with QI values in QIT through the clustering the closest non-sensitive QI values with SA. This helps in protecting the privacy and each group predict the risk level of the patient. Empirical assessments on original health care data corresponding to V.A.Medical Centre heart disease dataset illustrate the efficiency of this model corresponding to information loss, utility and privacy.

Case study

The V.A.Medical Centre database is the well-known heart disease data set extensively used by ML researchers for dealing with heart diseases. This database comprises of 76 attributes; nonetheless all the experiments that are published refer about making use of a subset of 14 among them.

The **Table- 1** below is the sample test report of the patients used by the medical practitioner for determining the measure of heart disease earlier than they occur in patients. Every one of these attributes can be grouped into three kinds:
1. Identifier attributes: a minimal set of attribute which can make the explicit identification of individual records.
2. Sensitive Attributes (SA): The sensitive attribute are regarded to be a privacy breach in case related to a specified individual, and are provided in the right of the Table.
3. Quasi Identifiers (QI) attributes: The remaining of the attributes that are non-sensitive, are utilized for determining the level of heart disease and can also is utilized by an adversary for re-identifying individual patient record.

These Experiments has been carried out with the V.A.Medical Centre database on determining the probability related to the existence of values 1 and 2, for the case of the non-sensitive classes. In this work, only 14 attributes are considered by the ML researchers that are given below

Age,sex,type of chest pain(Cpt),blood pressure at rest(Rbp),serum Scestoral(Sc),blood sugar at fasting(Fbs),electrographic at rest(Recg),maximum heart rate(Thalach),ST depression made by exercise corresponding to rest(Old peak), exercise induced angina(Exang), slope of the peak exercise ST(Slope), number of major vessels(Ca), blood disorder (Thal), the predicted attribute (Class).

Eight patients' treatment history records are considered for evaluation as depicted in **Table- 2** and **Table- 3** indicates the ranges for diagnosing the diseases. In **Table- 2** only two patients' disease (P1 and P8) are diagnosed. Now the challenge is to predict heart risk level for the other patients and at the same time to preserve privacy of the patient (P1 and P8) by anonymising the table.

The initial phase removes the attributes Recg, Exang and slope as they have no values in **Table-1**. **Table- 4** is replaced with "1" on the basis of the range specified in **Table- 3** and 0 if not in the range or else it is left blank

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 1. Samples of Sparse heart disease dataset

Age	Sex	Cpt	Rbp	Sc	Fzs	Recg	Thalah	Exang	Oldpeak	Slope	Ca	Thal	Claas
66	1	1					120	0					
67			160	226	0				2.5			3.5	
67			175	290	0		129						
47		2	130		0		172		1.4	1	0	3	
56	1	2		236	0	0	178		0.8		0	3	
57	0				3						0		
63	1		130		0		147		1.4		1	7	2
44	1	2		289									
52	1		120						1.9			8	
57	1					2					2		
54	1		143		3				1.7			3	1

Table: 2. Eight patients' Samples of heart disease dataset

Patient id	Quasi Identifiers (QI)								Sensitive
	Cp	trestbps	chol	fbs	Thalach	oldpeak	ca	thal	Attributes
P1	1				120				2
P2		160	226	0		2.5		3.5	
P3		175	290	0	129				
P4				3			0		
P5	2		289						
P6		120				1.9		8	
P7				2			2		
P8		143		3		1.7		3	1

Table: 3. Attributes ranges of V.A.Medical Centre heart disease dataset

Attributes	Class one-severe	Second class-moderate
Cpt	2	1
Rbp	140-160	160-180
Sc	200-230	230-290
Fbs	0-2	2-4
Thalach	130-155	110-130
Old peak	2.5-3	1-2
Ca	1	2
thal	8	3

Then the patient's record is reorganized as a band matrix by employing GCS approach on the basis of their correlation which is depicted in **Table- 5**. The subsequent step improves the quality of anonymisation by employing disassociation is carried out that outputs two Tables namely QIT and ST which discloses privacy. QIT indicates the patient history of records and ST refers to the patient's level of heart disease. This is provided in Table 6.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 4. Converted sparse heart disease dataset

Patient id	QI								SA
	Cpt	Rbp	Sc	Fbs	Thalach	Old peak	Ca	thal	
P1	1	0	0	0	1	0	0	0	2
P2	0	1	1	0	0	1	0	1	
P3	0	1	1	0	1	0	0	0	
P4	0	0	0	1	0	0	1	0	
P5	1	0	1	0	0	0	0	0	
P6	0	1	0	0	0	1	0	1	
P7	0	0	0	1	0	0	1	0	
P8	0	1	0	1	0	1	0	1	1

Table: 5. Final Reorganized Table

Patient id	QI								SA
	ca	fbs	thal	oldpeak	Trestbps	chol	thalach	Cp	
P7	1	1	0	0	0	0	0	0	
P4	1	1	0	0	0	0	0	0	
P8	0	1	1	1	1	0	0	0	1
P6	0	0	1	1	1	0	0	0	
P2	0	0	1	1	1	1	0	0	
P3	0	0	0	0	1	1	1	0	
P5	0	0	0	0	0	1	0	1	
P1	0	0	0	0	0	0	1	1	2

Table: 6. Final Published groups

Patient id	QIT									ST
	ca	fbs	thal	oldpeak	trestbps	chol	thalach	Cp		
P7	1	1	0	0	0	0	0	0		
P4	1	1	0	0	0	0	0	0		
P8	0	1	1	1	1	0	0	0		1
P6	0	0	1	1	1	0	0	0		
P2	0	0	1	1	1	1	0	0		
P3	0	0	0	0	1	1	1	0		
P5	0	0	0	0	0	1	0	1		
P1	0	0	0	0	0	0	1	1		2

Then Modified Density Based Clustering (MDBSCAN) is performed over QIT and ST in Table 6 which again gives the anonymised Table as result by splitting the medical records into two clusters as illustrated in **Table- 7** and **Table- 8**. This protects the privacy in such way that P7, P4, P6, P3, P1 patients' are groups as one cluster and P2, P5, P8 in other. In a similar manner, the diseases are predicted in such a manner that p7, p4, p6, p2 are more susceptible to type 2 and p2 p5 may be vulnerable to type 1 and the patients are advised to meet the clinicians before the disease becomes worsened. The GCS approach with clustering approach proved its efficiency in terms of utility, information loss and execution time.

THE IIOAB JOURNAL
ISSN: 0976-3104

Table: 7. Anonymised patient record

Patient id	QI								SA
	Ca	**Fbs**	**thal**	**Old peak**	**Rbp**	**Sc**	**Thalach**	**Cpt**	
P7	1	1	0	0	0	0	0	0	2
P4	1	1	0	0	0	0	0	0	
P6	0	0	1	1	1	0	0	0	
P3	0	0	0	0	1	1	1	0	
P1	0	0	0	0	0	0	1	1	

Table: 8. Anonymised patient record

Patient id	QI								SA
	Ca	**Fbs**	**thal**	**Old peak**	**Rbp**	**Sc**	**Thalach**	**Cpt**	
P8	0	1	1	1	1	0	0	0	1
P2	0	0	1	1	1	1	0	0	
P5	0	0	0	0	0	1	0	1	

PRELIMINARIES

Notation and description

The goal of the paper is the anonymization of the V.A.Medical Centre heart disease data which comprises of a set of patient records$PRD = (P_1, \dots P_n)$, n = |T|, Each patient record $P \in PRD$ contains attributes that are received from an attribute set$A = (a_1, \dots a_d)$d = |A|. The data is specified as a binary matrix A with n number of rows and d number of columns. For example **Table- 1** has 16 patient records with 14 attributes which is discussed elaborately again in the section that follows.

$$A[i][j] = \begin{cases} 1 & a_j \in P_i \\ 0 & a_j \notin P_i \end{cases} \quad (1)$$

For instance, the matrix for heart disease data shown in table 4 is expressed as

$$A[8][8] = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \quad (2)$$

In the set of attributes $A[i][j]$for the patient records$PR = (P_1, \dots P_n)$, some are sensitive to privacy, such as the heart disease is severe or medium condition seen in running example **[Table -1]**.

Definition 1 (Sensitive Attributes (SA)): The set $SA \in A$ having the attributes denoting a privacy threat if they are associated with certain patient records, renders the sensitive attributes,$S = \{sa_1 \dots, sa_m\}$ m = |S|. The rest of the attributes in the table indicated as Quasi Identifiers (QIs) are insensitive, which means that their association with a specific individual is not dangerous. In contrast, these items that are harmless can be used by an adversary for re-identification of individual patient records, as shown in the introductory section. These items are specified by QID attributes.

Definition 2 (Quasi Identifier (QI) attributes): The collection of attributes in A that an attacker can get knowledge about, for re-identifying separate patient records comprise the set of QIs. Typically, any attribute that is non-sensitive is QIs, therefore= $A \backslash SA$, $QI = (qi_1, qi_2, .. qi_m)$ .The records constituting of attributes from SA

THE IIOAB JOURNAL
ISSN: 0976-3104

are represented as Sensitive Attributes (SA), again the one havingjust the attributes from QI are regarded as non-sensitive.

Definition 3 (Privacy): A transformation that is privacy-preserving of patient set PR is said to contain privacy degree p if the probability corresponding to the association of any patient records $pr \in PR$ with a particular sensitive item $sa \in SA$ does not exceed 1/p. This is same as thinking that the respective patients records of a patient data can have correlation to a specific sensitive item having probability at the most 1/p within p − 1 other patient records. It has to be noticed that, the having an association of an individual patient data with an item in QI could not be considered as a privacy breach. A

Permutation-based technique is used same as [10], for achieving privacy preservation.
The need for privacy is imposed by implementing a partition in the patient record dataset PR into sets patient records which are disjoint that are Called as anonymized groups. For each group G, the exact QI attributes of the patient records are exposed, in addition to a summary corresponding to the frequencies of sensitive items which are in G. The probability of the association of any kind of transaction in G with that item amounts to half [16]. In general, let $f_1^G \dots f_1^m$ represent the number of occurrences with respect to the sensitive items $sa_1 \dots sa_m$ seen in group G. Hence group G gives the privacy degree,

$$p^G = \min |G|/f_i \qquad 3)$$

The privacy degree with respect to a whole partitioning $\mathcal{P}$ of PR is expressed by

$$p^{\mathcal{P}} = \min p^G$$

PROPOSED METHODOLOGY

Data Reorganization By Genetic With Cuckoo Search Algorithm

The V.A.Medical Centre heart disease data set that is used in this work involves sparse patient record. As the heart diseases data is less dense, it is randomly spread in the overall set. In the first phase, the sparse patient set is represented as a matrix format for the arrangement of band matrix. The Band matrix organization has been accepted as an advantageous mode to denote the sparse data in different scientific applications [20].

For the matrix A, a graph G = (V,E) is constructed where V has one vertex for every patient record there exists an edge from vertex vi to vertex vj for the entire non-zero elements .If matrix A is symmetric, and then G can be inferred to be undirected. GCS algorithm is based on the observation that indicates that a permutation pertaining to the patient records of the matrix is linked with a vertices re-labeling for G. For the transformation of heart diseases data matrix is required to reduce the distance ‘b’ between non-zero elements by doing a permutation of the rows and columns from the diagonal.

Reverse Cut hill-McKee Algorithm best deals with the symmetric matrices for the purpose of bandwidth reduction. This method is usually inexpensive, even though the result got may not have good quality, specifically if the original matrix is far from symmetric. Computing [A×A]T is an additional computational overhead; although the quality corresponding to the solution is far better (i.e. the resultant bandwidth is very much lesser). For the purpose of providing a solution to this issue, GCS is presented for the reduction of band matrix for both the case of symmetric and unsymmetric matrices.

To minimize the band width of the matrix A, GCS traversal begins from an initially selected patient records as root node. Each patient record lying at the same distance from the root inside the traversal path constitutes a level set. Each node in the traversal is considered as chromosomes. Nonetheless, the genetic algorithm can get easily slipped off in local optima. With a purpose to remedy this, a Cuckoo Search Algorithm (CSA) [21] is employed for searching for the nearby identical distance patient records from the initial patient records matrix ‘A’ which is generated by GA. The cuckoo bird lays one egg at a point of time and after which leaves this egg in any randomly chosen nest. The number of available nearest distance value corresponding to the patient records is pre-determined, and the probability Pa(0,1) is computed. Here, in this case the current patient records are selected as same and the rest of the rows are permutated in case it gives zero elements. The best nest patient records along

THE IIOAB JOURNAL
ISSN: 0976-3104

with the diagonal element (eggs) will be carried over to the next n generation when it is a good one or if it is stopped.

For the global optimal finding of the non-zero elements to patient records distance value is computed and rearrangement is also carried out between records exploiting genetic operations such as Multipoint crossover, and K swap mutation.

The Multipoint Crossover operator is employed for the initial band matrix solution hoping to generate a new better band matrix solution population by the interchange of many cross sections that are either odd or even.

K swap mutation operator is then applied over a parent patient records chromosome by randomly choosing 2 labels and then swapping them. Then swap operation is carried out k times.

The entire process is continued iteratively till no improvement is seen. At each step, the vertices that are in the same level which share the same parent (attributes such asca, Sc etc.,) are ordered in an increasing sequence on the basis of the vertex degree. By reversing the order obtained, the permutation that needs to be applied over the patients of matrix A is used.

Genetic Cuckoo Search (GCS)

Encode population of heart disease data points as chromosomes
Set population size, chromosome, max-gen, gen=0
Generate
Initial a population of n host
nests by randomly picking
samples (Ai, i =1,2,3....n);
In each patient records population fitness value(FT) is evaluated;
While (gen < max-gen)
If (FTi < FTj)
Replace row j by new patient record
End
Perform genetic operations
Get number of cuckoos randomly by Levy flights
Choose a nest as patient records among n randomly
Host birds abandon
Pa in (0,1)nests, and search Pa with non-zero elements
Select fitness(FT) with high Pa
perform permutation
Bandwidth reduced matrix B
Gen=gen+1
End while

Anonymization method

When the data in the sparse patient records is reorganized the subsequent step is the creation of anonymized groups. Within the first part, the relationship dissociation between QI and SA is finished and produced as two new Tables that might be the QIT and SAT that is involved in disclosing the privacy. The next phase is the anonymisation mechanism [22] in which the non-sensitive attributes gets organized with different sensitive attributes for the preservation of privacy in such a way that the degree of privacy is way much below 1/p in every one of the grouping.

Modified Density Based Clustering Method (MDBSCAN) [23] is employed to the anotomised QIT and ST for performing the anonymisation. The DBSCAN clustering approach has the dilemma of border objects. The border features often don't take into account foremost QIT and ST data points. So here MDBSCAN is offered which normally considers the core QIT and ST data points while clustering. At first, a group is created for every distinct touchy attributes by way of on the grounds that a collection of facets (P) within the ST. This sensitive values' QI is then mapped to the identical team and eliminated from the QIT. A point *P* is considered as a core point if it comprises at the least minPts facets that consists three or 4 quasi identifiers in distance ε (degree of privateness) of it, and additionally, these facets can also be reached directly from p.. Now the rest of the QI values in the QIT

THE IIOAB JOURNAL
ISSN: 0976-3104

Table are mapped to the group on the basis of the distance ε along with the sensitive attributes QI value. This procedure goes on till all the QI values are mapped onto the group.

Each group helps in assisting the patient by identifying the risk level of the disease. With the objective of preservation of privacy, each QIT with none of the sensitive value has to be grouped with sensitive transactions in such a way that the privacy degree is below 1/p in each one of the group.

MDBSCAN needs two parameters: ε and the minPts required in order to create a dense region. It begins with random sensitive attributes (unclassified) that has not been come across to be visited. This sensitive attributes ε-neighborhood is then regained to each QI, and in case it has adequately several points, a cluster is created. Else, the QI point is then labeled as outliers to SAT. This process is carried on till all the QI values are mapped onto the group. During the last step of the MDBSCAN, each core object is now allocated to its best density-reachable chain when all density-reachable chains that reach to the core object are known. Just in case any one of the clusters contains only one single QIT, the clusters are merged employing a linkage or amalgamation rule which determines when the 2 clusters are similar enough to be connected together.

Anonymization (Band matrix B)

Perform disassociation using anatomy
Release new two namely QIT and SAT
Clustering
(set of patients records (P∈SAT) , Eps, MinPts ,i∈QIT)

```
ClusterId := nextclusterId(outlier);
For(i=1;i<n; i++) DO
 Point := P.get(i);
  If Point. ClId = UNCLASSIFIED then
   If Expand Cluster(P, ClusterId, Eps, MinPts) then
ClusterId := nextclusterId(ClusterId)
   End If
  End If
 End For
End; // MDBSCAN
For x_border in the border list
result.size(x_border)=Retrieve Neighbours(x_border,Eps)
Assign x_border  to ClId
End for
End
```

The entire patient record is represented as SetOfPoints.The global density parameters that are determined manually are represented as Eps and MinPts. The function P.get(i) returns the i-th element of QIT. Given below ExpandCluster is the important function utilized by MDBSCAN .

```
Expand Cluster(set of patients records(P) ClId, Eps, MinPts) : Boolean;
seeds:=P.regionQuery(P,Eps);
If seeds. size<MinPts then // no core point
Else // all patient records in seeds are density- reachable from Point
P.changeClIds(seeds,ClId);
seeds. Delete(P);
While seeds <> Empty Do
current P := seeds. first();
result := P.regionQuery(current P, Eps);
If result.size>= MinPts then
P.changeClId(P,NOISE);
Return False;
Add all UNCLASSIFIED then seeds. Append(result P);
Label  UNCLASSIFIED and NOISE as BORDER DATAPOINTS
End If; // If result.size>= MinPtsthen
End While; // seeds <> Empty RETURN True;
End If ;//If seeds. size<MinPts then
```

THE IIOAB JOURNAL
ISSN: 0976-3104

End; // Expand Cluster

EXPERIMENTALEVALUATION

The heart disease dataset that is available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/heart-disease/processed.va.datais used to evaluate the GCS technique with MDBSCAN. The data set contains 76 raw attributes. However, the most published experiments best has reference to 14 of them with classification.

The data set comprises of 200 rows. The efficiency of this methodology is checked with respect to utility and effectiveness and proved that this method performs better than the existing techniques like Clustering-Based Anonymizer (CBA) [17], Data Recipient Centered (DRC) [18] approach.

Utility: The anonymized data set Utility is measures by calculating the distance present between the original and estimated pdf over all cells. It is measured by KL-divergence [24] as a metric which provides meaning for the evaluation of the amount of data loss endured by data anonymization.
KL_Divergence(Act,Est) = ∑_(∀cell C)⁡〖Act_C^S log⁡〖(Act_C^S)/(Est_C^S)〗〗 (5)
The actual pdf value of Sensitive Attribute (SA) for a cell C is expressed by
Act_C^S=(Occurrences of S in C)/(Occurrences of S in PR) (6)
The estimated pdf Est_C^SC is Calculated in a similar way, excepting that the numerator consists of equ(5) that is summed over all groups intersecting cell C
a • b/|G| (7)
The number of occurrences of the item s in G is represented by a, and the number of transactions that match the QIT selection predicate (last line of (5)) are denoted by b. For each (p, r) setting, 100 group-by queries gets generated by means of the random choice of SA and qi1 . . . qir from this the average reconstruction error is computed. The reconstruction error is measured with the following query.
SELECT COUNT (∗) FROM T WHERE (Sensitive Item sa is present) AND (qi1 = val1) ∧ . . . ∧ (qir= valr)

The reconstruction error is measured by changing the p the degree of privacy, m is the number of sensitive item that is arbitrarily selected, and r is the number of QI values that varies. More than 100 group-by queries are created by arbitrarily choosing q1,q2,q3...qnand s1,s2,s3.....sm. The average reconstruction error is determined and the clustering accuracy is measured. The experimental result is illustrated in the **Figure- 1**.The clustering algorithm preserves correlations in a better manner between the patient records for better utility.

Figure-2 illustrates the execution time of CBA, DRC and MDBSCAN for p = 20, that is the most important factor possessing influencing on runtime performance. MDBSCAN is time-effective, with completion time ranging at most 16 sec for the heart disease dataset. A more considerable overhead is suffered by GCS execution, which needs 185 sec for the heart disease dataset. Nevertheless, this overhead is only observed for the input transformation only once, regardless of p values. However the execution time corresponding to the new MDBSCAN with GCS is less compared with the other clustering methods. Since the MDBSCAN that hashigh dimension problem is solved by using of GCS.DRC only has the capability of dealing with execution times that are in in the range of 300 sec for the heart disease dataset.

The efficiency of QI is evaluated in terms of clustering accuracy. For this p=10 is fixed upon and the number of QI is varied in {2,4,6,8}. **Figure-3** illustrated the results on learning a QI attribute. It can be observed that clustering accuracy reduces only with a slight increase of QI, as the most correlated attributes are yet in the same column. In every case, MDBSCAN with GCS is seen with better clustering accuracy in comparison with other clustering techniques.

Figure-4 indicates the utility loss vs. privacy loss with regard to various privacy requirements (p=10 and p=20). The results that are affected in terms of privacy and utility are measured. If one selects a different measure for privacy (or utility), then the figure may appear in a different way. From the figure, GCS outperforms compared to other privacy requirements. The results indicates that GCS yields considerable better data utility rather than RCM and RCM with greedy (RCMG), as in the GCS method unsymmetric band matrix reconstruction is also taken into consideration.

THE IIOAB JOURNAL
ISSN: 0976-3104

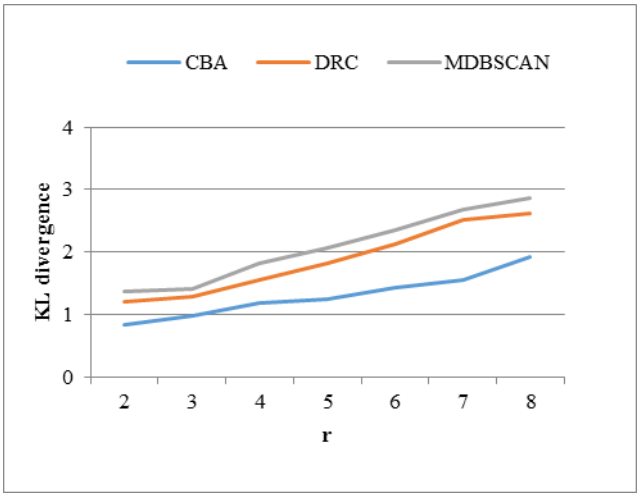

Fig: 1. Reconstruction Error vs r (p = 10)

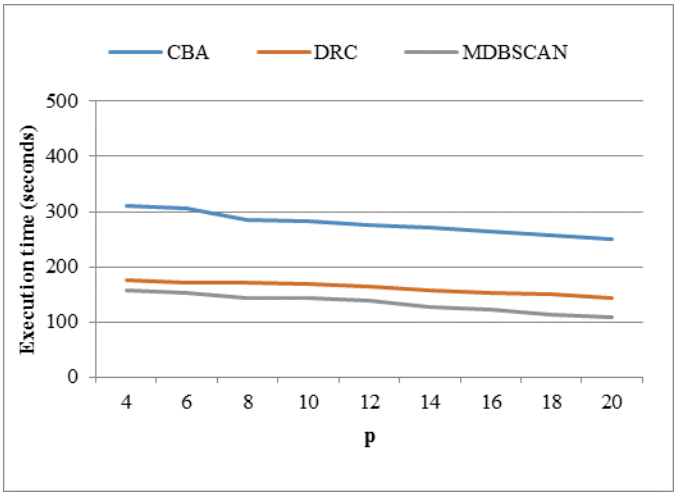

Fig: 2.Execution Time

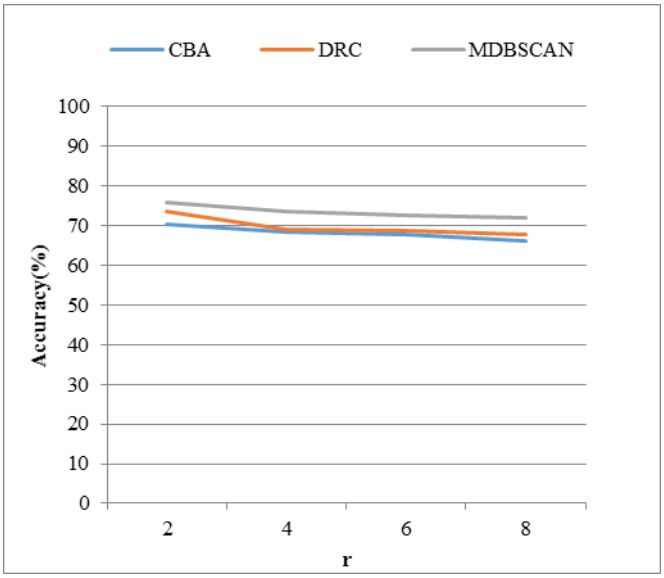

Fig: 3. Clustering accuracy (r=8 ,p=10)

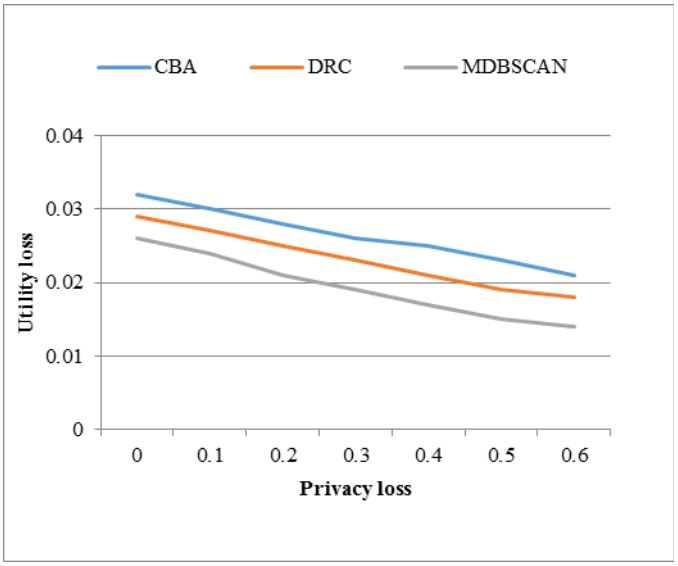

Fig: 4. Privacy-Utility Trade-off by Varied privacy requirements Clustering accuracy (r=8, p=10)

CONCLUSION

The proposed methodology designs a model for the representation of the sparse high dimensional medical dataset obtained from V.A.Medical Centre heart disease dataset with the attitude of protecting the patient's privacy from an adversary and additionally to predict the disease's threat degree. The CGS algorithm converts the sparse high-dimensional heart disease data set as a band matrix that limits the search space and maximizes the utility.

THE IIOAB JOURNAL
ISSN: 0976-3104

Anatomizations technique reduces the data loss and discloses privacy. The Modified Density Based Clustering (MDBSCAN) Method proved its resourcefulness through the minimization of the time, privacy and complexity in comparison to the available highly performing techniques like CBA,DRC and it also renders a reduction in the computational overhead.

COMPETING INTEREST

The authors hereby declare that they possess no competing interest.

AUTHORS' CONTRIBUTION

The author V.Shyamala Susan conceived the idea and developed the algorithm on the data set. The author Dr.T.Christopheranalyzed the result and provided suggestions to modify the algorithm. Both the authors read and approve the final submission.

ACKNOWLEDGEMENT

None

FINANCIAL DISCLOSURE

The work has been supported by the University Grants Commission (UGC)-New Delhi,India and the grant number is MRP-5711/15(SERO/UGC)January 2015.

REFERENCES

[1] Asghar MH, Mohammadzadeh N, Negi A.[2015] Principle application and vision in Internet of Things (IoT), in Computing, Communication & Automation (ICCCA), 2015 International Conference on,.427–431, 15–16 May.

[2] Luigi Atzori, Antonio Iera, Giacomo Morabito.[2010]The Internet of Things: A survey, Computer Networks, 54(15): 2787–2805, ISSN 1389–1286.

[3] Hicks D, Mannix K, Bowles HM, Gao BJ.[2015] SmartMart: IoT-based In-store mapping for mobile devices, in Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference on ,616-621:20-23

[4] Al Nuaimi, K Kamel, H. [2011]A survey of indoor positioning systems and algorithms, in Innovations in Information Technology (IIT), 2011 International Conference on185–190, 25–27 April

[5] Pereira PP, Jens E, Rumen K, Jerker D, Asma R, Mia J.[2013] Enabling cloud connectivity for mobile internet of things applications. In: Proceeding of the IEEE 7th international symposium on service oriented system engineering (SOSE), pp 518–526.

[6] Asghar MH, Mohammadzadeh N, Negi A.[2015] Principle application and vision in Internet of Things (IoT), in Computing, Communication & Automation (ICCCA), 2015 International Conference on, 427–431, 15–16 May.

[7] Hicks D, Mannix K, Bowles HM, Gao BJ.[2015] SmartMart: IoT-based In-store mapping for mobile devices, in 9thInternational Conference onCollaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 616-621, 20-23 Oct. 2013.

[8] Al Nuaimi, K Kamel H. [2011]A survey of indoor positioning systems and algorithms, in Innovations in Information Technology (IIT), 2011 International Conference on, 185–190, 25–27 April

[9] Luigi Atzori, Antonio Iera, Giacomo Morabito.[2010]The Internet of Things: A survey, Computer Networks, 54(15) : 2787–2805, ISSN 1389–1286.

ABOUT AUTHORS

***V Shyamala Susan**, received her post graduate degree in MCA from Coimbatore Institute of Technology, Coimbatore and M.phil degree is earned from M.S University, Tirunelveli. At present, she is working as the Head, Department of Computer Science in A.P.C. Mahalaxmi College for women, Thoothukudi, India. She has published 5 papers in international/national journal s. She has got eleven years of teaching experience and her area of interest is data mining.*

Dr. T. Christopher is presently working as Asst Professor, PG and Researchdepartment of Computer Science,Government Arts College,Coimbatore. Hehas published 20 papers in international/national Journals; His area ofinterest include Data mining, Network security. He has to credit 23 Yearsof teaching and research experience