Volume 3, No. 8, August 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICAL
Available Online at www.jgrcs.info

A DETAILED REVIEW ON PRIVACY PRESERVATION USING DISTANCE MEASURE TECHNIQUE IN DATA MINING

Kuntumalla Latha*[1] and Akkarabani Bharani Pradeep Kumar[2]

[1]Student, M.Tech (CSE), Anadapuram , Al-Ameer College of Engineering and Technology, VIZAG, A.P, India
[1]kuntumallalatha@gmail.com
[2]Asst.Professor, (CSE), Anadapuram , Al-Ameer College of Engineering and Technology, VIZAG, A.P, India

Abstract- Privacy is the main concerning in now a days. In this paper we concentrated on distance measures applied to ensure the privacy of the individual sensitive information.Protecting data privacy is an important problem in data distribution. Distance measure techniques typically aim to protect individual privacy, with minimal impact on the quality of the released data. Recently, a few of models are introduced to ensure the privacy protecting and/or to reduce the information loss as much as possible. That is, they further improve the flexibility of the anonymous strategy to make it more close to reality, and then to meet the diverse needs of the people. Various proposals and algorithms have been designed for them at the same time. In this paper we provide an overview of distance measure techniques for privacy preserving. We discuss the distance measure models, the major implementation ways and the strategies of distance measure algorithms, and analyze their advantage and disadvantage. Then we give a simple review of the work accomplished. Finally, we conclude further research directions of distance measure techniques by analyzing the existing work.

Keywords: Privacy, Distance measure, Closeness, Anonymity.

INTRODUCTION

Anonymized data publication has received considerable attention from the research community in recent years, due to the need of preventing "linking attacks" [1] in numerous data-dissemination applications. Consider, for example, that company wants to contribute its payment records in Table, called the microdata, to sociology scientists. Attribute Salary is sensitive, that is, the publication must ensure that no adversary can accurately infer the salary of any employee. Age and Zipcode are quasi-identifier (QI) [2] attributes, because they can be utilized in a linking attack to recover employees' identities. The advancement of information technologies has enabled various organizations (e.g., census agencies, hospitals) to collect large volumes of sensitive personal data (e.g., census data, medical records). Due to the great research value of such data, it is often released for public benefit purposes, which, however, poses a risk to individual privacy. A typical solution to this problem is to anonymize the data before releasing it to the public. In particular, the anonymization should be conducted in a careful manner, such that the published data not only prevents an adversary from infer-ring sensitive information, but also remains useful for data analysis. The released table gives useful information to researchers; it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table.

This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less specific but semantically consistent.

PRIVACY SKYLINE:

Privacy with Multidimensional Adversarial Knowledge:

Privacy technique deals with knowledge. Here data can be grouped into bins or buckets known a as D. we have to apply the technique then data becomes D*. The adversary may also have access to some external knowledge. In a very general sense, we can model this external knowledge using a logical expression, possibly containing variables. We say that an expression is **ground** if it contains no variables. A ground expression can be evaluated on a possible original dataset, and it returns true or false. We say that reconstruction *R* satisfies expression *E* iff *E* is true on *R*(**D***)[3].

We consider

a. ***Knowledge about the target individual:*** An interesting class of instance-level knowledge[4] involves information that the adversary may know about the target individual. For example, Tom does not have cancer.
b. ***Knowledge about others:*** Similarly, the adversary may have information about individuals other than the target. For example, Gary has flu.
c. ***Knowledge about same-value families:*** We think the most intuitive kind of knowledge relating different individuals is the knowledge that a group (or family) of individuals have the same sensitive value. For example,

{Ann, Cary, Tom} could be a same-value family, meaning if any one of them has a sensitive value (e.g., Flu), all the others tend also to have the same sensitive value disclosure and sanitizing data that improve computational efficiency several orders of magnitude over the best known techniques. This technique is efficient one when the knowledge is known. Another thing is that graph containing the relationship between the individual is not known clearly.

MOST PRIVACY PRSERVING

Data mining methods apply a transformation which reduces the effectiveness of the underlying data when it is applied to data mining methods or algorithms. In fact, there is a natural tradeoff between privacy and accuracy; though this tradeoff is affected by the particular algorithmwhich is used for privacy-preservation. A key issue is to maintain maximum utility of the data without compromisingthe underlying privacy constraints. A broad overview of the different utility based methods for privacy-preserving datamining is presented. The issue of designing utility based algorithms to work effectively with certain kinds of data mining problems is addressed.

Mining Association Rules under Privacy Constraints:

Since association rule mining is one of the important problems in data mining, we have devoted a number of chapters to this problem. There are two aspects to the privacypreserving association rule mining problem: When the input to the data is perturbed, it is a challenging problem to accurately determine the association rules on the perturbed data.A different issue is that of output association rule privacy. In this case, we try to ensure that none of the association rules in the output result in leakage of sensitive data. This problem is referred to as association rule hiding [5] by the database community, and that of contingency table privacy-preservation by the statistical community. The problem of output association rule privacy is briefly. A detailed survey of association rule hiding from the perspective of the database community is discussed.

Cryptographic Methods for information Sharing and Privacy:

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end [6]. For example, differentsuperstores with sensitive sales data may wish to coordinate among themselves inknowing aggregate trends without leaking the trends of their individual stores. This requires secure andcryptographic protocols for sharing the information across the different parties. The data may be distributed in two ways across different sites: In the area of privacypreserving data mining is that of data streams, in which data grows rapidly at an unlimitedrate. In such cases, the problem of privacy-preservationis quite challenging since the data is being released incrementally. In addition, the fast nature of data streams obviates the possibility of using the past history of the data.

We note that both the topics of data streams and privacy-preserving data mining are relatively new, and there has not been much work on combining the two topics. Some work has been done on performing randomization of data streams [7], and other work deals with the issue of condensation based anonymization [8] of data streams. Both of these methods are discussed in which are surveys on privacy and randomization respectively.

(N,T) CLOSENESS:

This is more flexible version for the privacy preservation. In these models we have use the distance measure between the two attributes named as count. Here we have used the various distance measures for identifying the distance between the attributes based on the information gain and how closely those attributes are related. This can be more flexible compare to other anonymization techniques but this alone is not sufficient to measure but we need multidimensional technique that can be used alongwith this approach can be very much helpful to keep the individual information confidentially. Now we will see this approach of N, T closeness [9] which can be distributed entire population of data.

Table 1. Original Table

ZipCode	Condition	Age
14850	Measles	33
14853	Allerg	26
14853	y Gout	22
14853	Cancer	32
14850	Flu	48
14850	Heart	47
14850	Flu	46
14853	Cancer	53
14853	Heart	51
13063	Flu	24
13063	Cancer	38
13068	Cancer	38
13068	Heart	30

After applying the above techniques the table is in the form of Anonymized version. Here we can see the anonymization by using the closeness by using thevarious distance measures the distance between the attributes is less age can say that they are close to each other,so replaced the last digits by '*' can hide the original details of the patient.

Limitations of (N,T)-closeness:

a. There is no computational procedure to enforce (N, T)-closeness.
b. There is effective way till now of combining with generalizations and suppressions or slicing.
c. Lost co-relation between different attributes: This is because each attribute is generalized separately and so we lose their dependence on each other.

d. Utility of data is damaged if we use very small t.(And small t will result in increasein computational time.

Table 2. Anonym zed Table

ZipCode	Condition	Age
148**	Measles	3*
148**	Allergy	2*
148**	Gout	2*
148**	Cancer	3*
148**	Flu	4*
148**	Heart	4*
148**	Flu	4*
148**	Cancer	5*
148**	Heart	5*
130**	Flu	2*
130**	Cancer	3*
130**	Cancer	3*
130**	Heart	3*

DISTANCE MEASURE

The Manhattan distance function computes the distance that would bet raveled to get from one data point to the other if a grid-like pathis followed. The Manhattan distance [10] between two it ems is the sum of the difference soft he ircorres ponding components. The formula for this distance betweena point ***X***= (***X1***,***X2***, etc.) and a point ***Y***= (***Y1***,***Y2***,etc.) is

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

The Euclidean distance between two points p and q is the length of the line segment connecting them p and q In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n- space, then the distance from p to q, or from q to p is given by:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + (q_4 - p_4)^2 \dots\dots} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The positionofapointinaEuclidean n-space is a Euclidean vector.So, p and q areEuclidean vectors, starting from the origin of the space,and theirtips indicate two points. The Euclidean norm, or Euclidean length, or magnitude of a vector measures thelength of the vector: values without packet loss, the increase in estimation error grows only linearly with the hop count and the growing speed is much slower than that of the hop count value. It shows that varying only the random forward hop count is not effective for providing better source location privacy.

$$\|P\| = \sqrt{(p_1)^2 + (p_2)^2 + (p_3)^2 + (p_4)^2 \dots\dots} = \sqrt{p.p}$$

Where the last equation involves the dot product. A vectorcanbedescribedasa directed line segment from the origin of the Euclidean space [11] (vector tail), to a point in that space (vector tip). If we consider that its length is actually the distance from it stail to it stip, it become sclear that the Euclideannormofavectoris just a special case of Euclidean distance: the Euclidean distance between it stailandit stip. Thedistance between points**p**and**q**may have a direction (e.g. from**p**to **q**), so it may be represented by another vector, given by,

$$(q - p) = (q_1 - p_1\ q_2 - p_2, q_4 - p_4, q_5 - p_5\ q_n - p_n$$

In a three-dimensional space (n=3), this is an arrow from p to q, which can be also regarded as the position of q relative to p. Itmay be also called a displacement vector if p and q represent two positions of the same point at two successive instants of time.

The Euclidean distance between p and q is just the Euclidean length of this distance (or displacement) vector:

$$\|q\text{-}p\| = \sqrt{q - p}$$

Whichisequivalenttoequation1, andalsoto

$$\|q\text{-}p\| = \sqrt{||p||^2 + ||q||^2 + 2.p.q}$$

One dimension:

In one dimension, the distance between two points on the real line is the absolute value of the irnumerical difference. Thus if *x* and *was* two points on the real line, then the distance between the mis given by:

$$\sqrt{(x - y)^2} = |x - y|$$

In one dimension, there is a single homogeneous, translation invariant metric (in other words, a distance that is induced by a norm), up to a scale factor of length, which is the Euclidean distance. In higher dimensions there are other possible norms.

Two dimensions:

In the Euclidean plane, if **p**= (p_1,p_2) and**q**= (q_1,q_2) then the distance is given by

$$d(p,q) = (q_1 - p_1)^2 + (q_2 - p_2)^2$$

This is equivalent to the Pythagorean theorem [12]. Alternatively, it follows from that if the polar coordinate soft the point **p**are (r_1,θ_1) and those of **q**are (r_2,θ_2), then the distance between the points is,

$$\sqrt{(r_1)^2 + (r_2)^2 + 2\, r_1 r_2 \cos\theta}$$

The above distance measures are use dinthecloseness function (n,t) for achieving better privacy while publishing the sensitive information of the individual.

CONCLUSION

This paper describes about the various distancemeasures used in the closeness technique to preserve the privacy of an individual while publishing the micro data like Hospital datasenses data etc.

REFERENCES

[1]. Adam N., Wortmann J. C.: Security-Control Methods for Statistical Databases: A Comparison Study. ACM Computing Surveys, 21(4), 1989.

[2]. Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

[3]. Agrawal R., Srikant R., Thomas D. Privacy-Preserving OLAP. Proceedings of the ACM SIGMOD Conference, 2005.

[4]. Agrawal R., Bayardo R., Faloutsos C., Kiernan J., Rantzau R., Srikant R.: Auditing Compliance via a Hippocratic database. VLDBConference, 2004.

[5]. Agrawal D. Aggarwal C.C. On the Design and Quantification of Privacy- Preserving Data Mining Algorithms.ACM PODS Conference, 2002.

[6]. Aggarwal C., Pei J., Zhang B. A Framework for Privacy Preservation against Adversarial Data Mining.ACM KDD Conference, 2006.

[7]. Aggarwal C.C. On k-anonymity and the curse of dimensionality.VLDB Conference, 2005.

[8]. Aggarwal C. C., Yu P. S.: A Condensation approach to privacy preserving data mining. EDBT Conference, 2004.

[9]. Aggarwal C. C., Yu P. S.: On Variable Constraints in Privacy- Preserving Data Mining. SIAM Conference, 2005.

[10]. Aggarwal C. C.: On Randomization, Public Information and theCurse of Dimensionality. ICDE Conference, 2007.

[11]. Bawa M., Bayardo R. J., Agrawal R.: Privacy-Preserving Indexing of Documents on the Network. VLDB Conference, 2003.

[12]. Aggarwal. G., Feder. T., Kenthapadi. K., Motwani. R., Kiran. S Approximation Algorithms for k-anonymity. Journal of Privacy Technology, paper 2005.