Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 74 (2014) 27243-27245

ISSN: 2229-712X

Security Maintaince System in Data Mining Using Distance Measure Technique

K Anitha and M Chandra Naik

Department of CSE, St.Mary's Group of Inistitutions Guntur St.Mary's Group of Inistitutions, Guntur, India.

ARTICLE INFO

Article history:

Received: 17 April 2013;

Received in revised form:

15 September 2014;

Accepted: 26 September 2014;

Keywords

Privacy,

Distance measure,

Closeness,

Anonymity.

ABSTRACT

*N*ow a day's Security is the main thing. In this paper we focus on distance measures applied to ensure the security of the separate sensitive information. Protecting data security is an key issue in data distribution. Security maintains system in data mining using distance measure techniques typically aim to protect separate security, with minimal impact on the quality of the released data. Now a days, a few of models are proposed to ensure the security protecting and/or to reduce the information loss as much as possible. i.e., they further improve the flexibility of the anonymous strategy to make it more closeness to reality, and then to meet the diverse needs of the people. Different proposals and algorithms have been designed for them at the same time. In this scenario we provide a survey of distance measure techniques for security preserving. We discuss the distance measure methods, the major achievement ways and the strategies of distance measure algorithms, and summarize their advantage and disadvantage. Then we give a demonstration of the work finished. Finally conclude further research directions of distance measure techniques by analyzing the existing work.

Introduction

The quality data publication has received important attention from the research community in recent technologies, due to the need of securing "linking attacks" [1] in many data dissimulate applications. Regard, for example, that an organization wants to contribute its payment records in Table, called the microdata. Attribute Salary is sensitive, i.e.,, 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 used in a linking attack to backup employees' identities. The preferment of ITs has enabled different organizations (e.g., census agencies, hospitals) to gather huge volumes of sensitive personal data (e.g., census data, medical records). Particularly, the anonymization should be conducted in a careful method; such that the published data not only secures a conflict from deduce sensitive information, but also remains useful for data analysis. The released table results give useful information; it presents detection risk to the separates whose data are in the table. Therefore, our objective is to limits the detection risk to an acceptable level while improving the uses. i.e., Achieved by the quality data before release. In the first step of anonymization is to remove clear identifiers. Anyway, this is not enough, as a conflict may existing know the quasi-identifier values of separates in the table.

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

Outlines of the security

Security with Multidimensional Opposed Knowledge: Security technique deals with data 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 conflict may also have access to some external data based knowledge. In a general case, we can model this external data based knowledge using a logical expression, possibly containing data changes. We say that an expression is ground if it contains no data changes. A ground expression can be evaluated on a possible original dataset, and it returns Boolean values. We say that reconstruction *R* satisfies expression *E* if and only if *E* is true on *R*(D*)[3].We consider as follows:

i. Knowledge about the target separate: An interesting class of instance-level knowledge [4] involves information that the conflict may know about the target separate. For example, Tom does not have cancer.

ii. *About others knowledge*: Likewise, the conflict may have information about separates other than the target. For example, Gary has flu.

iii. *Same-value families about knowledge:* We think the most intuitive type of knowledge relating various separates is the knowledge that a group (or family) of separates have the same sensitive value. For example, {Ann, Cary, Tom} could be a same family members, meaning if any one of them has a sensitive value (e.g., Flu), all the other tends also to have the same sensitive value detection 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 separate is not known clearly.

More security maintaining

Data mining methods apply a transformation which reduces the effectiveness of the underlying data when it is applied to

Tele:

E-mail addresses: Anitha.712@gmail.com

data mining methods or algorithms. In fact, there is a natural tradeoff between security and accuracy; though this tradeoff is affected by the particular algorithm which is used for security maintains. A key issue is to maintain maximum utility of the data without compromising the underlying security constraints. A broad survey of the different usability based methods for security maintaining data mining is given in such a way. The issue of designing usability based algorithms to work effectively with certain types of data mining problems is addressed.

Security constraints using mining association Rules: Since association rule mining is one of the important problems in data mining, we have attached a number of chapters to this problem. There are two ways to the security maintaining association rule mining problem: When the input to the data is regular state, it is a challenging problem to accurately determine the association rules on the regular state data. A various issue is that of output association rule security. In this scenario, 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 security maintains by the statistical community. The problem of output association rule security is briefly. A detailed description of someone of association rule hiding from the perspective of the database community is discussed

Information Sharing and security using Cryptographic Methods:

In some cases, multiple users may wish to share private information, without leaking any sensitive data at their end [6]. For example, various superstores with sensitive sales data may wish to communicate among themselves in knowing aggregate trends without leaking the trends of their separate stores. This needs secure and cryptographic protocols for sharing the information across the multiple users. The data may be distributed in two ways across various sites: In the area of security maintaining data mining is that of data streams, in which data grows quickly at an unlimited rate. In this case, the problem of security maintains is 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 observers that both the topics of data streams and security maintaining data mining are comparatively new, and there has not been much work on combining the two topics. Some work has We observers that both the topics of data streams and security maintaining data mining are comparatively 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 concentration based anonymization [8] of data streams. Both of these methods are explained in which surveys on security and randomization respectively are.

Closeness (nt):

This is very compatible version for the security maintains. In these methods we have use the distance estimate between the two attributes named as count. Here we have used the difference distance estimates for identifying the distance between the attributes based on the information gain and how closely those attributes are related. This can be very compatible compare to other anonymization techniques but this alone is not sufficient to estimate but we need multidimensional technique that can be used along with this approach can be very more helpful to keep the separate information confidentially. Now we will see this approach of N, T closeness [9] which can be distributed entire population of data.

Zipcode	Condition	Age
15850	Heart Disease	45
15853	Flu	29
15830	Cancer	67
15837	Flu	34
15879	Heart Disease	39
15845	Flu	32
15859	Cancer	28
16820	Heart Disease	57
16819	Heart Disease	79

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

Challenges of closeness (N, T):

i. There is no such calculations procedure to solve (N, T) closeness.

ii. There is efficient way till now of combining with generalizations and suppressions or parts.

iii. Lost co-relation between various attributes: This is because each attribute is generalized

Individually and so we lose their support on each other.

iv. Usability of data is crash if we use very small t.

Zipcode	Condition	Age
158**	Heart Disease	4*
158**	Flu	2*
158**	Cancer	6*
158**	Flu	3*
158**	Heart Disease	3*
158**	Flu	3*
158**	Cancer	2*
168**	Heart Disease	5*
168**	Heart Disease	7*

The Manhattan distance measure function computes the distance that would bet situation to get from a data point to the other if a grid-like path is followed. The Manhattan distance measure [10] between two it ems is the sum of the difference soft he is corresponding elements. The formula for this distance among a point X= (A_1, A_2, etc.) and a point Y= (B_1,B_2,etc.) is

$$\mathbf{d} = \sum_{i=1}^{n} (\mathbf{Ai} - \mathbf{Bi})$$

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 = (p_1, p_2,..., p_n) and q = (q_1, q_2,..., q_n) are two points in Euclidean n- space, then the distance from p to q, or from q to p is given by:

$$\mathbf{d(p,q)= d(q,p)} = \sqrt{(q1-p1)^2 + (q2-p2)^2 + (q3-p3)^2 + (q4-p4)^2}$$

$$= \sqrt{\sum_{i}^{n} (q1-p1)^2}$$

The position of a point in a Euclidean n-space is a Euclidean vector. So, p and q are Euclidean vectors, starting from the origin of the space, and their tips indicate two points. The Euclidean norm, or Euclidean length, or magnitude of a vector measures the length 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 security.

$$||P||=\sqrt{(p1)^2+(p2)^2+(p3)^2+(p4)^2\ldots.}=\sqrt{p.p}$$

Where the last equation involves the dot product. A vector can be described as a directed line segment from the origin of the Euclidean space [11] to a point in that space. If we consider that its length is actually the distance from it sail to it step, it become scalar that the Euclidean norm of a vector is just a special case of Euclidean distance: the Euclidean distance between it still and it stip. The distance between points p and q may have a direction (e.g. from p to q), so it may be represented by another vector, give

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

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. It may 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:

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

Which is equivalent to equation1, and also to?

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

One dimension: In one dimension, the distance between two points on the real line is the absolute value of the is numerical difference. Thus if *x* and *was* two points on the real line, then the distance between the misgiving 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,$p2$) 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 ($r2$,θ_2), then the distance between the points is,

$$\sqrt{(r1)^2+(r2)^2}+\overline{2\ r1r2\ \cos\theta}$$

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

Conclusion

This paper describes about the various distance measures used in the closeness technique to preserve the privacy of an individual while publishing the micro data like Hospital data senses 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.