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

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 38 (2011) 4387-4389

ISSN: 2229-712X

An enhanced data summarization for privacy preservation in incremental data mining

V.Rajalakshmi[1] and G.S.Ananthamala[2]

[1]Department of Information Technology, Sathyabama University, Chennai, India

[2]Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, India.

ARTICLE INFO

Article history:

Received: 22 August 2011;

Received in revised form:

26 August 2011;

Accepted: 31 August 2011;

Keywords

Privacy preservation,

Data summarization,

Incremental data,

Wavelet transformation,

B+ tree.

ABSTRACT

There has been a wide variety of research going on in the field of privacy preservation in data mining. Most of the methods are implemented for static data. But the world is filled with dynamic data which grows rapidly than what we expect. No technique is better than the other ones with respect to all criteria. This paper focus on a methodology that is well suited for incremental data that preserves its privacy while also performing an efficient mining .the method does not require the entire data to be processed again for the insertion of new data. The method uses data summarization technique which is used for both incremental data and providing privacy for such data. We develop the algorithm for making the environment flexible and cost effective.

Introduction

Data is information that has been translated into a form that is more convenient to move or process. The World Wide Web is filled with huge amount of data, which approximately doubles in every 50 days. Such data is a mixture of useful and non useful contents. Huge amount of data requires special consideration for its storage, retrieval and reasoning. In order to utilize the useful data efficiently data mining techniques are used. Data mining also called as knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. The aim of these algorithms is the extraction of relevant knowledge from large amount of data, while protecting at the same time sensitive information. Privacy is considered as the most challenging job today .The main objective of such privacy preserving algorithms are efficient mining of data and at the same time not compromising the privacy. No technique is better than the other ones with respect to all criteria.

Data mining has been combined with privacy preserving algorithms for more number of years which works efficiently only for static data. Such algorithms fail abruptly even if one record is added to it. If the data is updated in such a manner then again the complete algorithm should be run as though it is a new data. Is the algorithm going to recompile the whole data again or the new one alone is the question?

Data summarization is a method of representing raw information into summarized information so that it takes relatively less storage space and with acceptable degree of confidence. It is a method to represent the data coherently. This is the only method which can be used efficiently for hugely increasing data stream and since it generates new summarized data , and also it provides privacy without additional cost. The various summarization techniques are Aggregation, Pattern identification, Categorization, Feature extraction, Drift calculation & Generalization. In this paper an aggregation method called wavelet transformations is used.

Situation That Needs Privacy

We assume that the attacker has been keeping track of all the released tables; he thus possesses a set of released tables $\{T_0, \ldots, Tn\}$. We also assume that the attacker has the knowledge of who is and who is not contained in each table; that is, for each anonymized table Ti, the attacker also possesses a population table U_i which contains the explicit identifiers and the quasi-identifiers of the individuals in T_i. For instance, consider medical records released by a hospital. Although the attacker may not be aware of all the patients, he may know when target individuals in whom he is interested (e.g., local celebrities) are admitted to the hospital. Based on this knowledge, the attacker can easily deduce which tables may include such individuals and which tables may not. the goal of the attacker is to increase his/her confidence of attribute disclosure by comparing the released tables all together. This problem is called cross-version inferences. The various types of such attacks are difference attack, intersection attack and record tracking attack.

D Haar Wavelet Transform

Wavelets are a set of *non-linear* bases. When projecting (or approximating) a function in terms of wavelets, the wavelet basis functions are chosen according to the function being approximated.

Hence, unlike families of *linear* bases where the same, static set of basis functions are used for every input function, wavelets employ a dynamic set of basis functions that represents the input function in the most efficient way. Thus wavelets are able to provide a great deal of compression and are therefore very popular in the fields of image and signal processing.

A basic lifting scheme HAAR algorithm is used for transformation. Its time complexity is $N \log_2 N$.

$$D_{j+1,i} = S_{j,2i+1} - S_{j,2i} \quad \text{--- (1)}$$

Tele:

E-mail addresses: rajalakshmi.bala03@gmail.com

$$S_{j+1,i} = S_{j,2i} + \left\lfloor \frac{D_{j+1,i}}{2} \right\rfloor \quad (2)$$

The algorithm replaces the even elements with an average. The output of one step of the algorithm becomes the input for the next step. This results in a smoother input for the next step of the wavelet transform. The odd elements also represent an approximation of the original data set, which allows filters to be constructed. A simple lifting scheme forward transform is diagrammed in Figure 1.

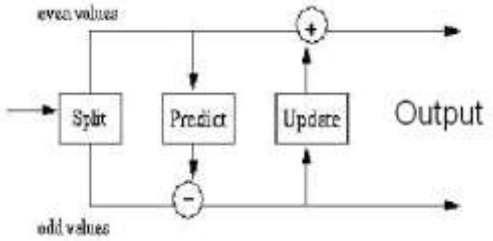

Fig. 1 Lifting scheme forward wavelet transform

The wavelet result for a relatively small data set (say 64 or 128 values) will tend to have few repeat values (e.g., a frequency close to 1/N for each value, in the case of N values).Hence the algorithm is advised for usage in a large database.

B+ TREES

B+ trees are used for representing the data , as they can be considered as basic means of representing hierarchical summary information efficiently. In earlier implementations B-trees were used which may use less tree nodes than a corresponding B+-Tree and sometimes possible to find search-key value before reaching leaf node. But there are some crucial disadvantages for B trees. In B-trees

- Only small fraction of all search-key values are found early.
- Non-leaf nodes are larger, so fan-out is reduced. Thus B-Trees typically have greater depth than corresponding B+-Tree.
- Insertion and deletion are more complicated than in B+-Trees .
- Implementation is harder than B+-Trees.
- In B+ Tree, since only pointers are stored in the internal nodes, their size becomes significantly smaller than the internal nodes of B tree.

For a b-order B+ tree with h levels of index

- The maximum number of records stored is
  $nmax = bh - 1$
- $nkmin = 2(b/2)h-1$
- The minimum number of keys is
- The space required to store the tree is O (n)
- Inserting a record requires O (logbn) operations
- Finding a record requires O (logbn) operations

Problem Definition

When such a threat for privacy in incremental data stream is there, there should be an efficient methodology which increases the interpretability of the data and not compromising the privacy which is also cost effective. The algorithm represents the data in the form of B+ tree, summarize the data into a lesser sized one using wavelet transformations and check for cross version interference.

Related Work

The data summarization based technique(3) uses clustering method for summarization of data. This will end up as a lossy transformation of the data. The method using random numbers(1) is also specified for stream of data, which introduces randomization of the data.

This leads to difficult retrieval of data. Both these methods does not specify any particular data structure for efficient storage and access of the data. Agrawal et al. [14] proposed a value distortion technique to protect the privacy by adding random noise from a Gaussian distribution to the actual data.

They showed that this technique appears to mask the data while allowing extraction of certain patterns like the original data distribution and decision tree models with good accuracy. Bradley [5] also proposed a summarization scheme to speed up the *k*-means clustering method.

Implementation

A method that is used to implement privacy of a data is checked for less information loss, less execution time and more data privacy and data utility. The methods that are used in my procedure are selected based on these metrics and has been proved so.

1. The input database is prepared for summarization by altering the important fields like name, account number etc.,
2. The remaining integer fields are extracted to implement summarization.
3. The 1D HAAR wavelet transformations using lifting scheme on these data, which reduces the size of the data.
4. These data are added to the existing B+ tree as a new node
5. The B+ tree is height balanced for efficient utilization which summarizes the database and restricts to specific amount of data.

The data set that is used is the data collected from a popular series of hospitals in US. The data is highly dynamic as the patients and their diagnosis is a continuous process. A sample of the database is as shown:

Fig. 2 : Flow diagram of the method

Experimental Result

The method is implemented on a dataset of different sized records and the following results are obtained. The graphs show that the information loss is reduced as the number of records is increased. The efficiency of the algorithm increases as the data size is increased. Therefore, the algorithm works effectively for large sized database.

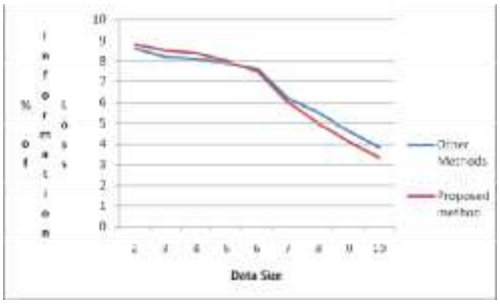

Fig 3: Graph demonstrating the information loss

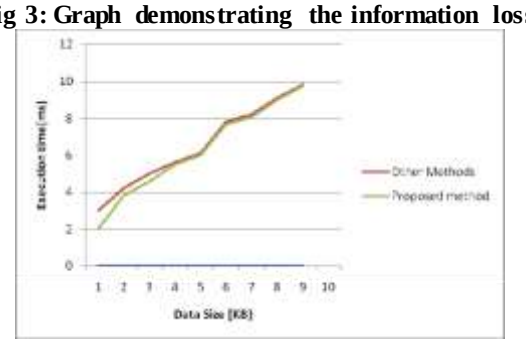

Fig 4: Graph demonstrating the Executing time

Conclusion and Future Enhancements

The method is used to compress the data and making it easier to store a frequently altered database. The method also stores the data in B+ tree enabling the efficient storage, access and updation of data. The method uses 1D HAAR transformation for summarizing the data which makes the execution faster. The method is implemented and proved as efficient compared to existing techniques. It can be further enhanced to solve specific type of attacks in privacy preservation of data.

References

[1] J.Gitanjali1, Dr.J.Indumathi2, Dr.N.Ch.Sriman Narayana Iyengar3," A Pristine Clean Cabalistic Foruity Strategize Based Approach for Incremental Data Stream Privacy Preserving Data Mining", 2010 IEEE 2nd International Advance Computing Conference,pp 410-415

[2] Faraz Rasheed, Young-Koo Lee, Sungyoung Lee," Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops",2006 IEEE

[3] Bidyut Kr. Patra, Sukumar Nandi, P. Viswanath," Data summarization based fast hierarchical clustering method for large datasets" , 2009 International Conference on Information Management and Engineering, pp278-282

[4] Chia-Han Yang and Don-Lin Yang," IMBT - A Binary Tree for Efficient Support Counting of Incremental Data Mining", International Conference on Computational Science and Engineering, 2009.

[5] P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In KDD, pages 9–15, 1998.

[6] Fatih Altiparmak, Hakan Ferhatosmanoglu," Incremental Maintenance of Online Summaries over Multiple Streams", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 2, FEBRUARY 2008

[7] Bi-Ru Dai, Li-Hsiang Chiang," Hiding Frequent Patterns in the Updated Database", 2010 IEEE

[8] Jia Yubo,Duan Yuntao, Wang Yongli," An Incremental Updating Algorithm forOnline Mining Association Rules", 2009
International Conference on Web Information Systems and Mining.

[9] Huidong Jin, K.-S. Leung", Scalable Model-Based Clustering for Large Databases Based on Data Summarization", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 11, NOVEMBER 2005.

[10] Shuguo Han, Wee Keong Ng, Li Wan," Privacy-Preserving Gradient Descent Methods", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,March 2010.

[11] Xindong Wu,Gong-Qing Wu, Fei Xie,Zhu Zhu, Xue-Gang Hu, Hao Lu, Huiqian Li," News Filtering and Summarization on the Web Intelligent Systems, IEEE, Sept.-Oct. 2010 Vol. 25, Issue:5 .

[12] Bulut, A.; Singh, A.K.," SWAT: hierarchical stream summarization in large networks " , Data Engineering, 2003. Proceedings. 19th International Conference, March 2003.

[13] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: applications to image and text data", In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 245–250, New York, USA, 2001.

[14] Jaideep Vaidya and Chris Clifton. Privacy-Preserving Association Rule Mining in Vertically Partitioned Data. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining, 2002.

[15] R. Agrawal and R. Srikant, "Privacy preserving data mining", In Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 439–450, Dallas, TX, May 2000

Table I A Sample Database

STATE PROFILES								
	Unit	U.S.	AL	AK	AZ	AR	CA	CO
POPULATION								
Total persons: (July 1)								
1993	1000's	257908	4187	599	3936	2424	31211	3566
2000	1000's	276242	4485	699	4437	2578	34888	4059
Percent increase:								
1990 to 1993	Percent	3.7	3.6	8.9	7.4	3.1	4.9	8.2
1990 to 2000 (1)	Percent	11.1	11	27.1	21.1	9.7	17.2	23.2
65 yrs and over, 1993	Percent	12.7	13	4.4	13.4	15	10.6	10
Residing in a metro area, 1992	Percent	79.7	67.4	41.8	84.7	44.7	96.7	81.8