TELKOMNIKA, Vol.17, No.6, December 2019, pp.3155~3160
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i6.13059

Classifying confidential data using SVM for efficient cloud query processing

Huda Kadhim Tayyeh*[1], Ahmed Sabah Ahmed Al-Jumaili[2]
[1]Department of Informatics Systems Management (ISM), College of Business Informatics, University of Information Technology and Communications, Baghdad, Iraq
[2]Department of Business Information Technology (BIT), College of Business Informatics, University of Information Technology and Communications, Baghdad, Iraq
*Corresponding author, e-mail: haljobori@uoitc.edu.iq[1], asabahj@uoitc.edu.iq[2]

Abstract

Nowadays, organizations are widely using a cloud database engine from the cloud service providers. Privacy still is the main concern for these organizations where every organization is strictly looking forward more secure environment for their own data. Several studies have proposed different types of encryption methods to protect the data over the cloud. However, the daily transactions represented by queries for such databases makes encryption is inefficient solution. Therefore, recent studies presented a mechanism for classifying the data prior to migrate into the cloud. This would reduce the need of encryption which enhances the efficiency. Yet, most of the classification methods used in the literature were based on string-based matching approach. Such approach suffers of the exact match of terms where the partial matching would not be considered. This paper aims to take the advantage of N-gram representation along with Support Vector Machine classification. A real-time data will used in the experiment. After conducting the classification, the Advanced Encryption Standard algorithm will be used to encrypt the confidential data. Results showed that the proposed method outperformed the baseline encryption method. This emphasizes the usefulness of using the machine learning techniques for the process of classifying the data based on confidentiality.

Keywords*: advanced standard encryption, cloud database, cloud query processing, support vector machine*

1. Introduction

The emergence of cloud computing services such as Hardware-as-a-Service (HAAS), Infrastructure-as-a-Service (IAAS) and Software-as-a-Service (SAAS) have contributed toward facilitating daily-basis business's transactions [1, 2]. In the past, an organization is required to provide a full-support hardware, platform and software for accomplishing its goals. This comes with high expensive cost of installation, licensing and maintenance [3]. Therefore, the cloud computing services have offered a great opportunity which represented by enabling organizations and corporations to use specific service per usage [4].

The Database-as-a-Service (DAAS) was one of the services that have been offered by several cloud service providers [5]. DAAS facilitates the process of initiating a database for a specific organization where such organization is able to use such database including querying and storing tasks without bothering the expenses of maintenance and backup operations [6]. However, this has posed a serious challenging issue which is the privacy [7]. In some fields, it is highly risky to let their own data vulnerable to be violated by any third party such as the medical domain [8]. Therefore, organizations tend to use an encryption task in order to protect their own data.

Apparently, this would challenge the use of cloud computing services especially for the DAAS where the organizations are accommodating regular tasks daily such as the storing, backing up and querying. Considering the encryption and decryption required to perform these tasks, a significant time consuming would indeed occur [9]. In order to improve the efficiency of such issue, several researchers have proposed various types of encryption methods that would have efficient performance in terms of the time and memory consumption [10-12]. However, using light encryption would lead to several potential attempts to violate the data. Therefore,

Received May 8, 2019; Revised July 2, 2019; Accepted July 18, 2019

some authors have recently examined a trade-off mechanism in which the data is being classified firstly in terms of confidentiality and then based on its status a partial encryption will take a place.However, most of the classification methods used to categorize the confidential data were relying on string-based matching approach.

Various research studies have examined the efficiency of cloud data storage for example, Wang et al. [13] presented a ranking approach for improving the efficiency of search within data stored in the cloud. In fact, the search within an encrypted data is too complicated and may yield inaccurate results. Therefore, the authors have took the advantage of statistical techniques such as term frequency and mutual information and in order to rank the documents within the cloud data. In this regard, the search query typed by the user will be examined in terms of the term frequency and mutual information in order to retrieve the most relevant documents.

In addition, Ren et al. [10] proposed an efficient query processing over the cloud based on a k-nearest neighbour classification method. The proposed method aims to index the documents within the cloud storage in order to efficiently and effectively retrieve the most relevant data. Meanwhile, the authors have used the random perturbation approach in order to insure optimal confidentiality.

Apart from the encryption, some researchers have attempted to classify the data based on the confidentiality prior to migrate the data into the cloud. Graepel et al. [14] presented a classification method for categorizing the data prior to the migration. Similarly, Zardari et al. [15] proposed a classification technique for distinguishing the confidentiality of the data. Their technique was intended to provide several classes for the confidentiality.

Recently, Albadri & Sulaiman [16] have examined the classification of data before migrating it into the cloud with a rule-based classification technique. The authors have used a real-time data of students and manually annotate each data instance into their confidentiality status. Consequentially, the authors have developed a set of rules in order to distinguish the data. The classification was built based on string-based matching among the data attributes. Based on such classification, a partial encryption has been performed for the confidential data in order to reduce the load of query processing.

Zardari et al. [17] have examined the role of machine learning in terms of classifying confidential data using KNN classification method. The proposed classification method aimed to identify which data needs to be encrypted based on its confedintiality. The classification was relying on string-based similarity of data attributes. Renu et al. [18] have proposed a binary tree classification method for protecting confidential data. The proposed method is based on pre-defined dictionary along with a string-based matching. Such dictionary contains confidential data and the string-based matching will compare the new data or unseen data with the predefined ones. From the literature, one could notice that most of the classification methods used for categorizing confidential data were relying on string-based matching. Taking the advantage of other data representation such as N-gram and utilizing the frequency of terms would facilitate toward improving the classification. Therefore, this study aims to propose a bag-of-word (BoW) representation or so-called N-gram along with Support Vector Machine (SVM) classifier for the process of confidentiality classification.

2. Research Method

The proposed method of this study consists of five steps as shown in Figure 1. First step is related to the dataset used in the experiment which will be used for the classification. Next step is related to the adjustment required for making the data suitable for the classification task. Such adjustment is known as N-gram representation. After that, the classification step will take place by categorizing the data into confidential and non-confidential. Based on the results of such classification, an encryption process will be performed upon the confidential data using Advanced Encrypted Standard (AES). Finally, an evaluation for the encrypted data will be done using query processing. Next sub-sections will tackle each step independently.

2.1. Data

The data used in this study is the one that has been introduced by Albadri and Sulaiman [16] which consists of University student's information. Such data contains regular information about students such as their basic information and information related to their

courses, grading and payments. The data has been manually annotated based on four class labels including (i) sensitive data, (ii) Confidential data, (iii) Internal data, and (iv) public data. Sensitive data is related to the basic information such as date of birth and mother name. Confidential data is the most restricted information which is related to the payments, student's grades and other information that is not tolerated to be violated. The internal data is the data that is being permitted to use by the staff of the university such as the progress reports of the students only. Finally, the public data which is the normal information that is not private such as the name of the student. In fact, each class label requires specific encryption mechanism in which the most confidential needs sophisticated encryption and vice versa. Table 1 shows the details of the dataset.

2.2. N-gram Representation

In order to enable the SVM classification, it is necessary to turn the data into vectors. For this purpose, the N-gram representation has been used. Such representation aims to process all the terms that have been occurred within the dataset [19]. Then, the distinct occurrence of terms will be maintained. In other words, the redundant terms will be discarded. This is to insure that all the unique terms are being considered. After that, the unnecessary terms will be removed such as the stopwords. This is due to their insignificant impact in terms of determining the class label. Hence, all these terms will be used as columns or attributes where the data instance will be examine in terms of these attributes based on the occurrence. Table 2 depicts an example of this representation.

As shown in Table 2, each data instance will be examined in terms of the terms that located in the columns. Such examination refers to whether the instance contain this term or not. Containing the term will be represented as '1', while the absence will be represented as '0'. Now each data instance will be represented as a vector like the following vector:

00001000000

in this regard, each vector will contains a value of '1' which refers to the occurrence of a corresponding term [20, 21].

Table 1. Dataset Details

Description	Quantity
No. of tables	35
No. of fields	362
No. of class labels	4

Table 2. N-gram Representation

Data	Term 1	Term 2	Term n	Class
Instance 1	1	0	0	Sensitive
Instance 2	0	1	0	Internal
Instance 3	0	0	0	Public
Instance 4	0	0	1	Public
Instance 5	0	0	0	Public

2.3. Classification using SVM

This algorithm is intended to create a training model based on examples of the data. This means that the data instances within the dataset that is being given a class label will be used for the training [22]. Such model will make the algorithm is able to predict the class label of each data instance. The prediction of this algorithm is based on a margin which known as Hyperplane. In the vector space where the data is seen as vector, SVM will aim to assign an accurate Hyperplane that is dividing the data into two classes [23, 24]. Figure 2 depicts such division by the Hyperplane. The way of computing such Hyperplane can be illustrated based on the following equation.

$$f(\vec{x}) = sgn\big((\vec{x} \times \vec{w}) + b\big) = \begin{cases} +1: & (\vec{x} \times \vec{w}) + b > 0 \\ -1: & Otherwise \end{cases} \quad (1)$$

2.4. AES Encryption

After classifying the confidential data using the proposed SVM, an encryption task will take a place. For this purpose, the Advance Encryption Standard will be used to encrypt the confidential data. AES has been widely used for encryption purposes regarding to its various

key lengths such as 128 bits, 192 bits and 256 bits [25]. Basically, the aforementioned lengths will be used for the three classes Internal, Sensitive and Confidential respectively. This is since the fourth class label will not be encrypted.

2.5. Evaluation

After encrypting the confidential data, a query processing will take a place where multiple types of queries are being used. The evaluation of query processing will be based on the time consumed to retrieve or to execute queries.

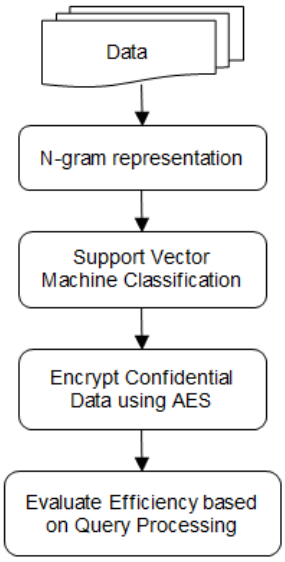

Figure 1. The proposed method steps

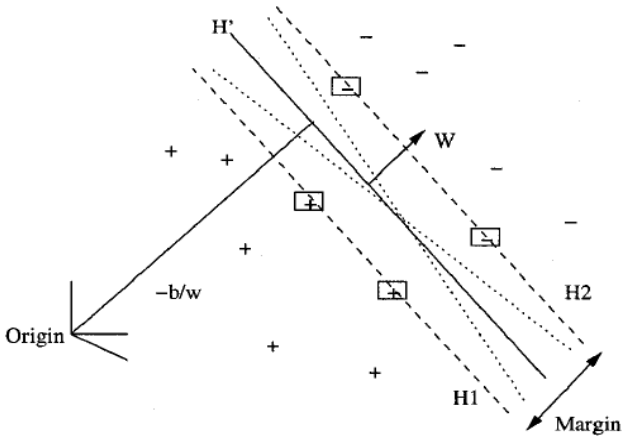

Figure 2. Hyperplane separation

3. Results and Analysis

In this section, the results of time consumption will be depicted. As mentioned earlier, based on the results of classification, the data will be encrypted in accordance to each class label. For this purpose, three types of query will be used in the experiments including Add, Select and Delete queries. For each type, 50 queries will be used for the evaluation. In addition, the baseline encryption of [16] will be used to compare its performance against the proposed method. Figures 3, 4, and 5 will show the results of baseline along with the proposed method for each type of queries.

For the Add query results as shown in Figure 3, approximately both the baseline and the proposed method has similar performance. However, the proposed method has slightly better performance by consuming less time. In general, the add query is the most query that consume time compared to other queries. For the select query results shown in Figure 4, the proposed method has outperformed the baseline by consuming lesser time. Similarly, for the results of delete query shown in Figure 5, the proposed method showed superior performance compared to the baseline.

The reason behind the superiority of the proposed method lies on the effectiveness of SVM classification method which facilitated toward classifying the data in accordance to the confidentiality. This has increase the number of data that may need lesser encryption length or even might need an encryption at all. In fact, the way of representing the terms based on N-gram has demonstrated better performance of classification compared to the string-based matching representation.

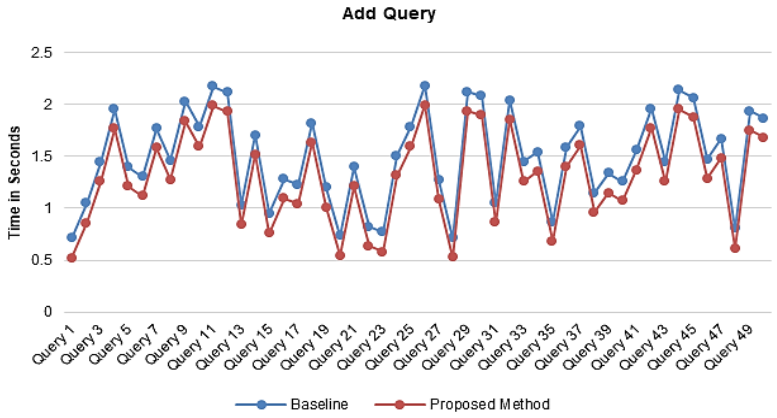

Figure 3. Results of add query for the proposed method and the baseline

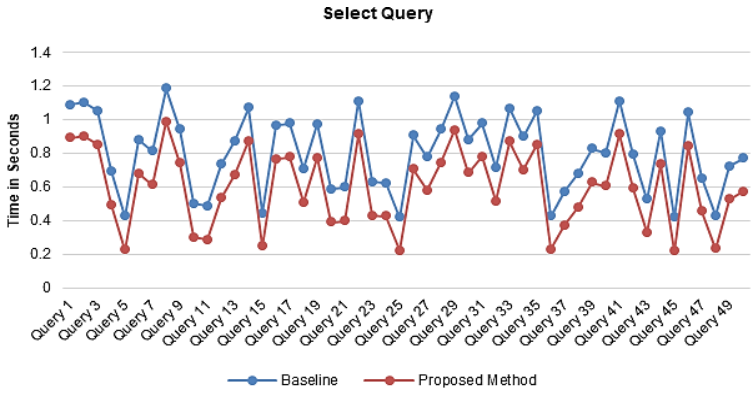

Figure 4. Results of select query for the proposed method and the baseline

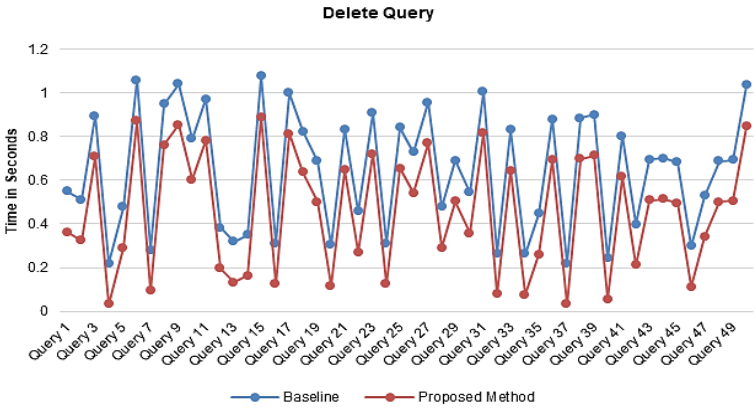

Figure 5. Results of delete query for the proposed method and the baseline

4. Conclusion

This paper has proposed an improved classification method for categorizing confidential data using SVM. With the use of N-gram representation, SVM has showed better classification accuracy. Based on such classification results, an encryption has been conducted using the AES algorithm. Results of encryption showed that the proposed method has outperformed the baseline encryption based on the efficiency of query processing. This emphasizes the usefulness of N-gram representation compared to the string-based matching when categorizing confidential data. For future studies, addressing different encryption methods would yield better performance.

References

[1] Krutz RL, Vines RD. Cloud security: A comprehensive guide to secure cloud computing: Wiley Publishing. 2010.
[2] Behrend TS, Wiebe EN, London JE., Johnson EC. Cloud computing adoption and usage in community colleges. *Behaviour & Information Technology*. 2011; 30: 231-240.
[3] Dillon T, Wu C, Chang E. *Cloud computing: issues and challenges*. Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference. 2010: 27-33.
[4] Zhang Q, Cheng L, Boutaba R. Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications*. 2010; 1: 7-18.
[5] Curino C, Jones EP, Popa RA, Malviya N, Wu E, Madden S, Balakrishnan H, Zeldovich N. *Relational cloud: A database-as-a-service for the cloud*. 5th Biennial Conference on Innovative Data Systems Research. 2011.
[6] Al Shehri W. Cloud Database Database as a Service. *International Journal of Database Management Systems*. 2013; 5: 1.
[7] Agrawal D, El Abbadi A, Emekci F, Metwally A. *Database management as a service: Challenges and opportunities*. Data Engineering, 2009. ICDE'09. IEEE 25th International Conference. 2009: 1709-1716.
[8] Sammour M, Hussin B, Othman MFI, Doheir M, AlShaikhdeeb B, Talib MS. DNS Tunneling: a Review on Features. *Int. J. Eng. Technol*. 2018; 7: 1-5.
[9] Bethencourt J, Sahai A, Waters B. *Ciphertext-policy attribute-based encryption*. Security and Privacy, 2007. SP'07. IEEE Symposium. 2007: 321-334.
[10] Ren Y, Xu J, Wang J, Kim JU. Designated-verifier provable data possession in public cloud storage. *International Journal of Security and Its Applications*. 2013; 7: 11-20.
[11] Li R, Liu AX. *Adaptively secure conjunctive query processing over encrypted data for cloud computing*. 2017 IEEE 33rd International Conference on Data Engineering (ICDE). 2017: 697-708.
[12] Sahin C, Allard T, Akbarinia R, El Abbadi A, Pacitti E. *A Differentially Private Index for Range Query Processing in Clouds*. 2018 IEEE 34th International Conference on Data Engineering (ICDE). 2018: 857-868. doi:10.1109/ICDE.2018.00082.
[13] Wang C, Cao N, Ren K, Lou W. Enabling secure and efficient ranked keyword search over outsourced cloud data. *IEEE Transactions on parallel and distributed systems*. 2012; 23(8): 1467-1479.
[14] Graepel T, Lauter K, Naehrig M. *ML confidential: Machine learning on encrypted data*. International Conference on Information Security and Cryptology. 2012: 1-21.
[15] Zardari MA, Jung LT, Zakaria MN. *Hybrid Multi-cloud Data Security (HMCDS) Model and Data Classification*. Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference. 2013: 166-171.
[16] Albadri H, Sulaiman R. A Classification Method For Identifying Confidential Data To Enhance Efficiency Of Query Processing Over Cloud. *Journal of Theoretical & Applied Information Technology*. 2016; 93(2): 412-420.
[17] Zardari MA, Jung LT. Classification of File Data Based on Confidentiality in Cloud Computing using K-NN Classifier. *International Journal of Business Analytics (IJBAN)*. 2016; 3: 61-78.
[18] Renu S, Veni SK. An Enhanced CIA tree Using String Matching Algorithm. *International Journal of Applied Engineering Research*. 2017; 12: 6123-6126.
[19] Konchady M. Text mining application programming: Charles River Media, Inc. 2006.
[20] Sun Q, Liu H, Ma L, Zhang T. A novel hierarchical Bag-of-Words model for compact action representation. *Neurocomputing*. 2016; 174: 722-732.
[21] Elshourbagy M, Hemayed E, Fayek M. Enhanced bag of words using multilevel k-means for human activity recognition. *Egyptian Informatics Journal*. 2016; 17: 227-237.
[22] Yin C, Xiang J, Zhang H, Wang J, Yin Z, Kim JU. *A new svm method for short text classification based on semi-supervised learning.* Advanced Information Technology and Sensor Application (AITS), 2015 4th International Conference. 2015: 100-103.
[23] Pan X, Yang Z, Xu Y, Wang L. Safe screening rules for accelerating twin support vector machine classification. *IEEE transactions on neural networks and learning systems*. 2018; 29: 1876-1887.
[24] Ishida H, Oishi Y, Morita K, Moriwaki K, Nakajima TY. Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions. *Remote sensing of environment*. 2018; 205: 390-407.
[25] Hoang T. *An efficient FPGA implementation of the Advanced Encryption Standard algorithm*. Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference. 2012: 1-4.