Research on Privacy-Preserving Technology for Cloud Computing

Xiaolong Wang
College of Physical Science and Technology, Central China Normal University, Wuhan, China
flysky_221@163.com

Abstract - Consumers will be able to access applications and data anywhere in the world on demand by cloud computing which promises reliable services delivered through next-generation data centers. Cloud computing has a very broad application prospects such as virtualization, large-scale, dynamic configuration and many other characters. At the same time, there are many security risks such as privacy information leakage in the network by the rapid growing of network security threats. Security issues is the key issues constraining the development of cloud computing. The Privacy Protection Support Vector Machine (PPSVM) is widely concerned in secure multi-party computation (SMC). We propose a new optimized Privacy Protection Support Vector Machine classifier without Secure Multi-Party Computation for vertically partitioned data set which is not disclosing the private data. The novel approach is proved as being greater than traditional classification SVM on privacy-preserving by some experiments.

Index Terms - Cloud computing, Security risks, SVM, Optimized PPSVM, Secure Multi-Party Computation.

I. Introduction

As a new and convenient information media, network has been gradually recognized for its large scale distributed computing resources and great application scenarios since 1990's. Cloud computing is a hot topic in recent years as research and application. Most of the IT companies and the industry believe cloud computing is the core architecture of the next generation computer network application technology. The US government projects that between 2010 and 2015, it's spending on cloud computing will be at approximately a 40-percent compound annual growth rate and will pass $7 billion by 2015 [1].

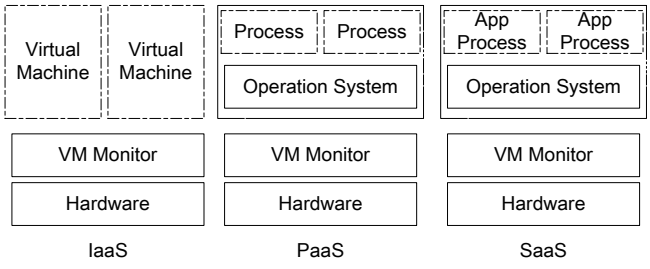

Fig. 1 Typical cloud computing platform

Cloud computing which promises reliable services delivered through next-generation data centers that are built on compute and storage virtualization technologies. Consumers will be able to access applications and data from a "Cloud" anywhere in the world on demand. In other words, the Cloud appears to be a single point of access for all the computing needs of consumers. The consumers are assured that the Cloud infrastructure is very robust and will always be available at any time. A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers [2]. The typical cloud computing platform is shown in Fig.1 which include three primary cloud-computing serving models: Iaas (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service), SaaS (Software-as-a-Service).

At the same time, there are many security risks such as privacy information leakage in the network by the rapid growing of network security threats. With the strong dependence of vast network data centers in the application process, the maintenance and data management of all the services are entrusted to the cloud computing service providers to complete. Security issues is the most important issue faced by people in the use of cloud computing data storage services [1].

More and more organizations and institutions devoted to the research and development of cloud computing security standards. The Cloud Security Alliance (CSA) is a not-for-profit organization with a mission to promote the use of best practices for providing security assurance within Cloud Computing, and to provide education on the uses of Cloud Computing to help secure all other forms of computing. The Cloud Security Alliance is led by a broad coalition of industry practitioners, corporations, associations and other key stakeholders [3]. Open cloud manifesto document [4] is intended to initiate a conversation that will bring together the emerging cloud computing community (both cloud users and cloud providers) around a core set of principles.

Cloud computing has a very broad application prospects such as virtualization, large-scale, dynamic configuration and many other characters. According to the Cloud Security Alliance, there are many security threats need to be considered in the development of cloud computing such as malicious use, unsafe application interface and data security issues [5]. Security issues is the key issues constraining the development of cloud computing. In order to achieve better development, the privacy protecting problems in cloud computing should have good solutions.

The rest of this paper is organized as follows. Cloud computing and data security principle is briefly reviewed in Section 2. In Section 3, we simply review privacy protection

techniques. In Section 4, we propose a novel optimized privacy protection support vector machine approach. Experimental results on two standard datasets are presented in Section 5, respectively, comparing novel method with multi-class classification SVM, which is followed by the conclusions and future works of this paper.

II. Cloud Computing and Data Security

The concept of cloud computing has been evolving for more than 40 years. The term "cloud" originates from the telecommunications world of the 1990s, when providers began using virtual private network (VPN) services for data communication. VPNs maintained the same bandwidth as fixed networks with considerably less cost: these networks supported dynamic routing, which allowed for a balanced utilization across the network and an increase in bandwidth efficiency, and led to the coining of the term "telecom cloud." Cloud computing's premise is very similar in that it provides a virtual computing environment that's dynamically allocated to meet user needs [1].

TABLE I Advantage and Disadvantage of Three Types Cloud

Types	Advantage	Disadvantage
Public Cloud	Simplest to implement and use; Minimal upfront costs; …	Most expensive long-term; Susceptible to prolonged services outages; …
Private Cloud	Allows for complete control; Minimal long-term costs; …	Large upfront costs; Susceptible to prolonged services outages; …
Hybrid Cloud	Most cost-efficient; Less susceptible to prolonged service outages; …	Difficult to implement; Requires moderate amount of space; …

Cloud computing is a kind of new resource-sharing policies, while providing a variety of applications and required services based on resource-sharing mechanism. Cloud computing is characterized as follows:

1) *Virtualization and large-scale*
2) *Resource sharing*
3) *Flexibility and reliability*
4) *Service-on-demand*

The decision each business faces is not quite as clear-cut as whether cloud computing could prove beneficial to it. Instead, each business must make a decision as to what type of cloud it will utilize: public, private, or hybrid [6]. Each cloud type has its own set of positives and negatives. These benefits and drawbacks are listed in Table.1.

The three major cloud computing security strategies are shown in Table.2. The basic idea for encryption-on-demand is that they try to use the encryption-on-demand server which can provide some kind of encryption service; the trusted virtual data center can separate each customer workloads to different associated virtual machines; the trusted computing is based on a smart design which uses a so-called trusted platform module chip which has a kind of endorsement private key.

TABLE II Advantage and Disadvantage of Three Security Methods

Types	Advantage	Disadvantage
Encryption on-Demand	Using random encrypt-system good for security	Need local machine to encrypt the data
Trusted virtual data center	Reduce the payload for one physical machine and avoid the misconfiguration problem, good for users data security	The user should check the data frequently to make sure it won't be changed
Trusted Cloud Computing	Good security by using private key and public key	It can also have some problems if the attackers use the playback attack

Shamir [7] introduced a novel type of cryptographic scheme, which enables any pair of users to communicate securely and to verify each other's signatures without exchanging private or public keys, without keeping key directories and without using the services of a third party.

Boneh [8] proposed a fully functional identity-based encryption scheme (IBE). The scheme has chosen ciphertext security in the random oracle model assuming an elliptic curve variant of the computational Diffie-Hellman problem. Our system is based on the Weil pairing. We give precise definitions for secure identity based encryption schemes and give several applications for such systems.

Baek [9] constructed the first identity-based threshold decryption scheme secure against chosen-ciphertext attack. A formal proof of security of the scheme is provided in the random oracle model, assuming the Bilinear Diffie-Hellman problem is computationally hard.

Joseph [10] provided a privacy-preserving data integrity protection by enabling public auditability for cloud storage and implements a scalable framework that addresses the construction of an interactive audit protocol to prevent the fraudulence of prover and the leakage of verified data in cloud storage by reducing the overhead in computation, communication and storage.

III. Privacy Protection Techniques

A. Traditional Privacy Protection Techniques

The classic privacy protection techniques can be roughly broken into four broad categories [11]. These categories include encryption and security mechanisms, anonymizing mechanisms, infra-structures and labeling protocols.

1) Encryption and Security Mechanisms

Confidentiality, reliability and integrity of information are concerned in encryption and other information security technology. The primary data privacy protection of existing systems adopts access control mechanism combined with the original data encryption protocol. These techniques can be applied to cloud computing data store and share up can make use of the information security mechanisms to achieve data privacy protection

2) Anonymizing Mechanisms

The literature [12] introduced a formal protection model named k-anonymity and a set of accompanying policies for deployment. A release provides k-anonymity protection if the

information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. It also examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, μ-Argus and k-Similar provide guarantees of privacy protection.

3) Infra-structures

Infrastructure-based privacy protection mechanism build middleware platform to develop privacy protection applications. Its essence is considering the privacy mechanisms throughout the process of program design. For example, marking the degree of user privacy preferences of all user data in the development process and providing different users with personalized service. The reminding and the right to make a different choice are given to the user when collecting and analyzing data.

4) Labeling protocols

Labeling protocol is the protocol language between users and information gathering person in order to achieve the privacy agreement between each other. This requires information collection person gives the statement and description of the preferences of user data, the searching range of data, and the purpose of data collection. P3P (Platform for Privacy Preferences Project) [13] promoted by W3C (World Wide Web Consortium) is the largest influence and the widest range of applications.

B. Privacy Protection Information Retrieval Techniques

Private information retrieval (PIR) provides a cryptographic means for retrieving data from a database without the database or database administrator learning any information about which particular item was retrieved [14]. PIR is firstly proposed in 1995 by Chor [15] to protect the user's query privacy when the users want to be able to query information in a public database. The researchers began to explore solutions from different directions in subsequent studies which are divided into two directions: improving the communication complexity [16]; solving other problems, such as the calculation of the cost-sharing server-side overhead and improve query robustness [17].

IV. Optimized Privacy Protection Support Vector Machine Approach

Most of the privacy protection technologies of distributed data are the cryptography-based privacy protection technologies, especially in distributed environment. Secure multi-party computation (SMC) [18] is the most commonly used protocol. The basic cryptographic algorithms applied in SMC include a variety of public key cryptosystem, in particularly, the semantic security homomorphic public key encryption system. With the development of privacy protection technology, the Privacy Protection Support Vector Machine (PPSVM) is also widely concerned in many literatures [21].

Vapnik [22] first proposed Support Vector Machine in 1995 which found a maximum interval classified hyper-plane to separate two types of data. The standard SVM model is to solve the optimization problem as:

$$\begin{cases} \min\left(\frac{1}{2}w'w + C\sum_{i=1}^{m}\xi_i^2\right). \\ d_i((w\cdot x_i)+b)+\xi_i \geq 1 \end{cases} \quad (1)$$

Where $\xi_i \geq 0, i = 1,\ldots,m$ is the slack variable in the constraint, m is the number of all samples, C is interval parameter to adjust interval and deviation. The weight vector w and bias b are computed and decision function $f(x) = w\cdot x + b$ is to determine the class of new data sample x. The kernel function is defined as $K(x_i, x_j) = x_i \cdot x_j$ for linear SVM and other $K(x_i, x_j)$ for Non-linear SVM such as Radial Basis Function (RBF) [23] kernel function is

$$K(x_i, x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{g}). \quad (2)$$

We propose a new optimized Privacy Protection Support Vector Machine classifier without Secure Multi-Party Computation for vertically partitioned data set which is not disclosing the private data. The assumption is: matrix A is vertically partitioned as $A = (A_1, A_2, \ldots, A_q)$ which every column is owned by one entity that does not want to public or share their data with other entities. One entity discloses its local Gram matrix $A_i A_i^{'}$ which determine kernel matrix and A_i is not disclosed by this way. So the kernel matrix is:

$$K(A, A^T) = AA^T. \quad (3)$$

The public non-linear classification optimization problem is defined as:

$$\begin{cases} \min_{\tau}\left(\frac{1}{2}(\omega\tau)' K(x^T, A^T)(\omega\tau) - e'\tau\right). \\ \sum_{i=1}^{m}\omega_i\tau_i = 0 \end{cases} \quad (4)$$

Where $K(A, A^T)$ is computed by Eq.2. ω is the class label and τ is the parameter of data. According to Eq.1, the weight vector w and decision function $f(x)$ are computed by Eq.5 and Eq.6. And the class of sample x_i is determined by the positive or negative symbol of decision function:

$$w = \sum_{i=1}^{m}\omega_i\tau_i x_i. \quad (5)$$

$$f(x) = \prod_{i=1}^{m}\omega_i\tau_i K(x_i^T, A_i^T) + b. \quad (6)$$

In fact, the unique matrix A_i could not be calculated if the kernel matrix $K(A, A^T)$ is the only known. Therefore, we conclude that the sample x_i could not be calculated if the kernel matrix $K(x_i^T, A_i^T)$ is the only known. So there is no any entity disclose its information from A_i or x_i .

V. Experiments on Datasets

The performance of proposed approach and standard SVM is compared in two datasets: part of PAMAP2 Physical Activity Monitoring dataset and full SMS Spam Collection Data Set. The PAMAP2 Physical Activity Monitoring dataset [24] contains data of 18 different physical activities (such as walking, cycling, playing soccer, etc.), performed by 9 subjects wearing 3 inertial measurement units and a heart rate monitor. The dataset can be used for activity recognition and intensity estimation, while developing and applying algorithms of data processing, segmentation, feature extraction and classification. SMS Spam Collection Data Set [25] is a collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site and a subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis. Finally, it has 1,002 SMS ham messages and 322 spam messages. Assuming the data of PAMAP2 Physical Activity Monitoring dataset is vertical distributed and held separately as 18 entities by 18 different physical activities. And the data of SMS Spam Collection Data Set is also vertical distributed and held separately as 2 entities by the ham messages and the spam messages. The classification accuracy result of the proposed optimized privacy protection SVM is shown in Table.1 and compared with the standard multi-class classification SVM [26]. The accuracy is improved by the proposed method under two kinds of datasets which has different dimension size. On the PAMAP2 Physical Activity Monitoring dataset, the proposed approach achieved better performance than on SMS Spam Collection Data Set.

TABLE III Classification Accuracy Results

Datasets (Dimension)	optimized privacy protection SVM	multi-class classification SVM
PAMAP2 (2500,18)	93.87%	90.05%
SMS Spam (1324,2)	96.44%	95.19%

VI. Conclusion

The validity and reliability of one novel optimized privacy protection SVM is proved as being greater than standard multi-class classification SVM on privacy-preserving by building and testing on two datasets. Compared with standard approach, the optimized privacy protection SVM has the advantage on the classification accuracy and the privacy protection. The present study only compared the two procedures on small databases. The testing on additional datasets and different types of data will be necessary to assess the generalisability of these results.

References

[1] L.M. Kaufman, "Data Security in the World of Cloud Computing," IEEE Security & Privacy, July-Aug. 2009, pp.61-64.

[2] R. Buyya, C.S. Yeo and S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities," Dalian: 10th IEEE International Conference on High Performance Computing and Communications (HPCC 2008), Sept. 25-27, 2008.

[3] https://cloudsecurityalliance.org/about/

[4] http://www.opencloudmanifesto.org/opencloudmanifesto1.htm

[5] http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf.

[6] F. Hu, M. Qiu, and J. Li, et al, "A review on cloud computing: Design challenges in architecture and security," Journal of Computing and Information Technology 2011:25-55.

[7] A. Shamir, "Identity-based cryptosystems and signature schemes," In: Blakely, G.R., Chaum,D. (eds.) CRYPTO 1984. LNCS, Springer, Heidelberg, vol. 196, pp. 47-53, 1985.

[8] D.Boneh and M. Franklin, "Identity-based encryption from the weil pairing," In Advances in Cryptology-CRYPTO, pp.213-229, 2001.

[9] J.Baek and Y. Zheng, "Identity-based threshold decryption," Public Key Cryptography – PKC 2004. 2947: pp. 262-276, 2004.

[10] N.M. Joseph, E. Daniel and N.A.Vasanthi, "A Scalable Privacy-Preserving Verification Correctness Protocol to Identify Corrupted Data in Cloud Storage," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2013, 2(3): pp: 0951-0956.

[11] M.S. Ackerman., "Privacy in pervasive environments: next generation labeling protocols," Personal and Ubiquitous Computing, 2004, 8(6): 430-439.

[12] L. Sweeney, "k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002: 557-570.

[13] http://www.w3.org/P3P/

[14] F. Olumofin and I. Goldberg, "Revisiting the computational practicality of private information retrieval," Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2012: 158-172.

[15] B.Chor, E. Kushilevitz and O. Goldreich, et al, "Private information retrieval," Journal of the ACM (JACM), 1998, 45(6): 965-981.

[16] A. Ambainis, "Upper bound on the communication complexity of private information retrieval," Automata, Languages and Programming. Springer Berlin Heidelberg, 1997: 401-407.

[17] D.Asonov, "Querying Databases Privately: A New Approach To Private Information Retrieval," SpringerVerlag, 2004

[18] O. Goldreich, "Secure multi-party computation," Manuscript. Preliminary version, 1998.

[19] H. Yu, X Jiang and J. Vaidya, "Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data," Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006: 603-610.

[20] V. Keeman, "Learning and soft computing, support vector machines," Neural Networks and Fuzzy Logic Models, The MIT Press, Cambridge, MA, 2001.

[21] H. Yu, J. Vaidya, X.Jiang, "Privacy-preserving svm classification on vertically partitioned data," Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 647-656.

[22] V. Vapnik, "The nature of statistical learning Theory," NY:Spring-Verlag,1995

[23] Reddy, B. Rama Sanjeeva, D. Vakula, and N. V. S. N. Sarma. , "Circular antenna array pattern analysis using radial basis function neural network," IOP Conference Series: Materials Science and Engineering. Vol. 44. No. 1. IOP Publishing, 2013.

[24] http://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring

[25] http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[26] http://www.csie.ntu.edu.tw/~cjlin/libsvm/