PROCEEDINGS OF SCIENCE

Privacy Preserving SVM with Different Kernel Functions for Multi-Classification Datasets

Zekun Li [1]
Shaanxi Normal University, Xi'an, China
E-mail: lizekun@snnu.edu.cn

Shuyu Li
Shaanxi Normal University, Xi'an, China
E-mail: lishuyu@snnu.edu.cn

Aiming at the mining problem of privacy preserving data, a SVM algorithm under differential privacy is presented for multi-classification datasets in the paper. The main idea of the proposed algorithm is to add Laplace noise for decision function, thus the privacy can be protected when the computation value of kernel function is changed and the normal vector is disturbed. Three different kernel functions including the linear kernel, the polynomial kernel and the Gaussian kernel respectively, are selected for classification comparison. Experiment results show that three kernel functions can achieve better classification accuracy rate to a certain degree. From the view of the computation time, the liner kernel is the fastest while that of Gaussian kernel is the slowest. From the perspective of classification accuracy rate after noise addition, the polynomial kernel is the most accurate.

CENet2017
22-23 July, 2017
Shanghai, China

[1]Speaker

[2] Supported by the National Natural Science Fund project (41271387) and the Fundamental Research Funds for the Central Universities (GK201703055)

PoS(CENet2017)018

1. Introduction

With the development of information society, more and more people start to pay attention to personal privacy. The privacy can be disclosed by various means, among which, the sensitive data concerning personal privacy can be achieved by data mining. It has resulted in research on the development of privacy preserved data mining. In order to achieve the corresponding knowledge while avoiding the leakage of personal privacy, the data mining under differential privacy [1] seems promising.

Recently, several differential privacies for SVM approaches have been designed. For example, for normal data, the privacy protection has been carried out from the protection of the minimum empirical risk [2] to the protection of decision function of algorithm for SVM itself [3]. Lin proposed the PPSVC algorithm [4], which destroyed the attribute value of support vector mainly with post-processing method and then used Taylor formula to compute and output an attribute similar to the original value. While Li et al. proposed a differential privacy algorithm for RFB kernel, in which the main decision function has been switched first and then the Laplace noise is added [5]. For partitioned data, Sun proposed a P3SVM for classification of vertically partitioned data, based on the concept of global random reduced kernel which is computed by using local reduced matrix with Gaussian perturbation [6]. Yu constructed the global SVM privacy preserving model from the horizontal partitioned data and vertical partitioned data without disclosing the data of each party to others [7] [8].

In this paper we derive the privacy-preserving SVM with different kernel functions and evaluate them on Multi-Classification Datasets. The different kernel functions including linear kernel, polynomial kernel and Gaussian kernel respectively are adopted by adding noise to the objective function to achieve privacy protection. As demonstrated, the liner kernel is the fastest while Gaussian kernel is the slowest. From the perspective of classification accuracy rate after adding the noise, polynomial kernel is the most accurate one.

The rest of the paper is structured as follows. Section 2 describes the fundamental knowledge. Section 3 proposes the differential privacy algorithm for SVM with different kernel functions (linear kernel, polynomial kernel and Gaussian kernel respectively), mainly adding Laplace noise for decision function to carry out privacy protection for multi-classification data sets. Section 4 gives the experiment result of the proposed algorithm. Section 5 concludes the paper and points out the future improvement of the algorithm.

2. Fundamental Knowledge

Definition 1 (the differential privacy): if algorithm M can be in sibling data sets D_1, D_2 (there is only one different data in two data sets) and the random output $S \subseteq Rang(K)$ meets

$$Pr[K(D_1 \in S)] \leq e^{\varepsilon} \times Pr[K(D_2 \in S)] \tag{2.1}$$

Algorithm M satisfies ε - differential privacy [1].

Definition 2 (the sensitivity and Laplace mechanism): define a numerical function f, so its sensitivity is indicated as $\triangle f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|$. In terms of the concept of sensitivity, the noise will follow the Laplace distribution with magnitude as $b = \Delta f / \varepsilon$, that

is, to meet the differential privacy of function *f* in data set *D*, the release form of function will be *f(D) + X* where *X* conforms to *Lap(b)* [9].

Definition 3 (Support Vector Machines): as one of the efficient classification algorithms, the form of SVM can be written as [10]:

$$\min_{w,b} \frac{1}{2}||w||^2 + c\sum_{i=1}^{m} l(y_i, f_w(x_i)) \tag{2.2}$$

The w is a normal vector and separates two different data in hyperplane, while *l* is a convex and differentiable loss function and $c > 0$ is the penalty factor.

Definition 4 (Reproducing Kernel Hilbert Space): let *X* be arbitrary set and *H* a Hilbert space of real-valued functions on *X*. The evaluation functional over the Hilbert space of functions *H* is a linear functional that evaluates each function at a point *x* [11],

$$L_x : f \to f(x) \forall\, f \in H \tag{2.3}$$

Assume *H* be a Reproducing Kernel Hilbert Space.

3. SVM and Differential Privacy

The basic idea for classification algorithm of SVM is that data sample can be dividable in Reproducing Kernel Hilbert Space (RKHS), while the kernel function has also privately defined this feature space. Hence, the selection of kernel function is extremely important to the entire classification algorithm.

For SVM algorithms, the privacy leakage may occur in the decision function, the leaked data is the value of the normal vector which is calculated by the kernel function; as a result, the privacy protection is implemented by adding noise to some part of the kernel function, namely, $w = min\, cw`$, where $w`$ is the result after noise addition.

In this paper, we use three kernel function realizations, specifically, the linear kernel, the polynomial kernel and the Gaussian kernel respectively. Aiming at the multi-classification data sets, the privacy model is respectively established and the privacy protection is carried out to support the classification algorithm of SVM.

3.1 Implementation of the Algorithm

In this paper, we use three kernel function realizations which are linear kernel, polynomial kernel and Gaussian kernel respectively. The first kernel function is the linear kernel in the concrete form of $k(x, y) = x_i^T x_j + c$. For large data sets with various categories, the classification speed of linear classifier will have obvious advantages. The second kernel function is polynomial kernel in the concrete form of $k(x, y) = (a\, x_i^T x_j + c)^d$. The third kernel function is Gaussian kernel. As a translation invariant kernel, the sampling data can be theoretically classified into infinite dimensions and the form of privacy preserving SVM can be written as

$$\min_{w,b} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \left(k(x_i, x_j) + Lap(x)\right) + c\sum_{i=1}^{m} l(y_i, f_w(x_i)) \tag{3.1}$$

The main idea of our algorithm is to add Laplace noise to three kernel functions respectively.

3.2 Privacy Analysis

Suppose that the values of the kernel functions on the data sets *D1* and *D2* be k_1 and k_2, respectively, where k refers to the value of the kernel function after the adding noise. Then the proof process is shown as follows:

For , so $Pr[k_1 \in k] \propto \exp(-\varepsilon((k)-(k_1)))$,Then

$$\frac{Pr[k_1 \in k]}{Pr[k_2 \in k]} \propto \frac{\exp\left(-\varepsilon\left((k)-(k_1)\right)\right)}{\exp\left(-\varepsilon\left((k)-(k_2)\right)\right)}$$

$$\exp(\varepsilon)(((k)-(k_1))-((k)-(k_2)))$$

Throug $((k)-(k_1))-((k)-(k_2)) < (k_1)-(k_2)$ and $(k_1)-(k_2) \le | k_1 \oplus k_2 |$, the above formulas can be summarized to prove that the kernel function satisfies ε -differential privacy.

4. Experiment

4.1 Dataset

The proposed algorithm is the improvement on the basis of Libsvm algorithm. The pendigits of UCI are used for the experiment dataset. Before starting the experiment, the dataset shall be firstly normalized to control that each property value is in the scope of [0,1]. To further improve the accuracy rate, the optimal value of parameter *gamma* and penalty factor *c* needs to be determined, which can be obtained only by the method of exhaustion through grid.py because of the property of SVM. And the optimal classification accuracy rate can be achieved based on the optimal value of parameter *gamma* and penalty factor *c*.

4.2 Analysis on the Experimental Results

Due to the randomness of noise, the average number of classification accuracy rate with 5 times is regarded as the final experimental results. Firstly, the optimal value of parameter *gamma* and penalty factor *c* is achieved through grid.py and then the optimal classification accuracy rate is obtained.

Then the classification accuracy rate of three selected kernel functions are compared under the condition that the fixed parameters *degree* = 3 and *gamma* = 0.5, the experiment is carried out by modifying the value of penalty factor *c* and privacy budgeting ε , the experimental results is shown in Fig. 1, Fig. 2 and Fig. 3 respectively.

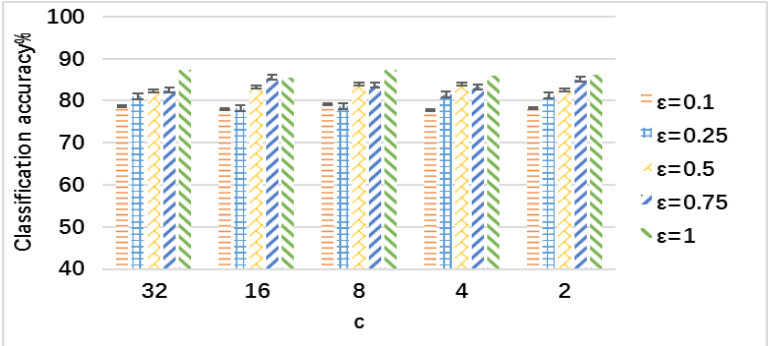


Figure 1: Classification Accuracy Rate of Liner Kernel in Different Privacy Budgeting

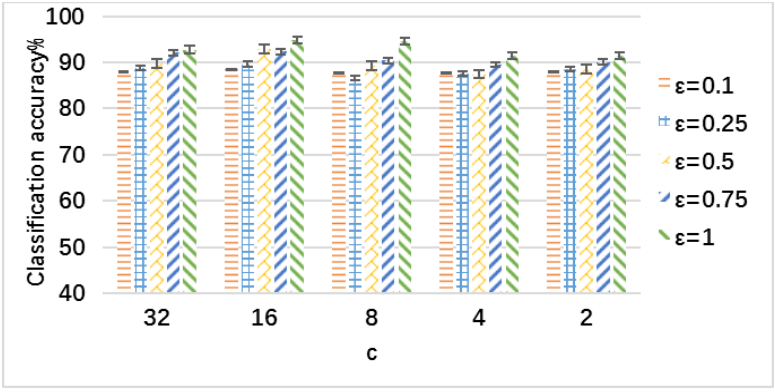


Figure 2:Classification Accuracy Rate of Polynomial Kernel in Different Privacy Budgeting

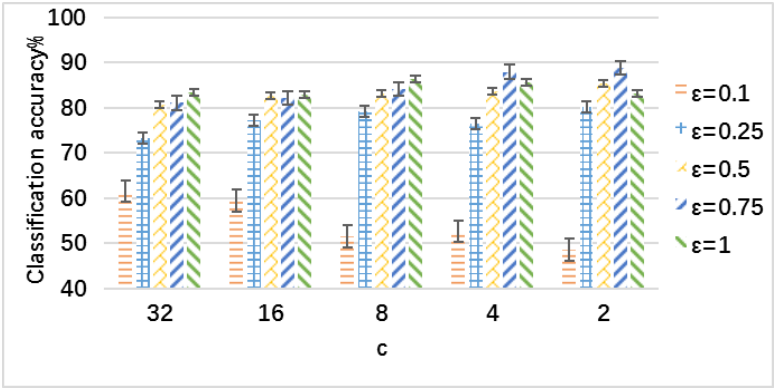
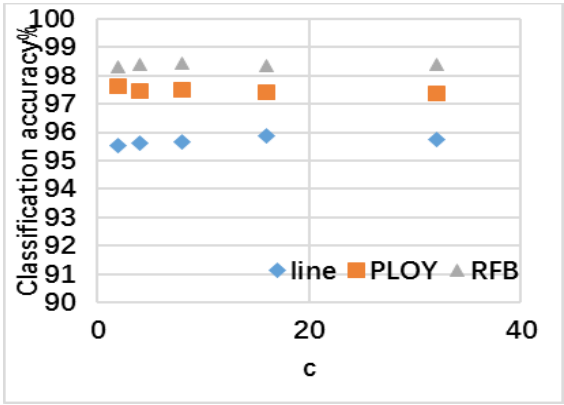


Figure 3: Classification Accuracy Rate of Gaussian Kernel in Different Privacy Budgeting

As shown in the above figures, the support vector machine model of three kernel functions can achieve better classification accuracy rate to a certain degree. With the increase of privacy budgeting, the classification accuracy rate of three kernel functions added with the Laplace noise increase in positive proportion. In the case of no privacy and under the condition of privacy budgeting for 0.1, 0.25, 0.5, 0.75 and 1, further study on the comparison of different classification accuracy rates of three kernel functions, are collected respectively. The experimental results are shown in the following figures (Fig. 4 to Fig. 9).



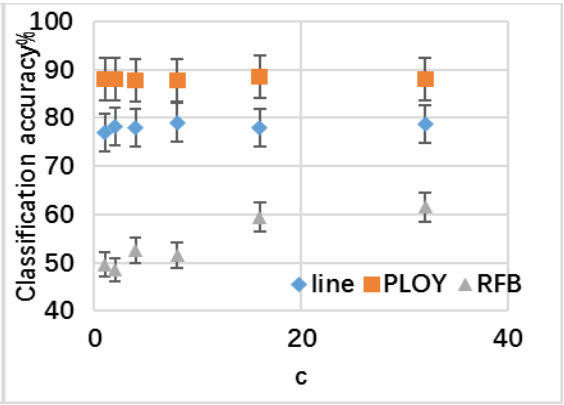


Figure 4: Classification Accuracy without Noise **Figure 5:** Classification Accuracy with *ε =0.1*

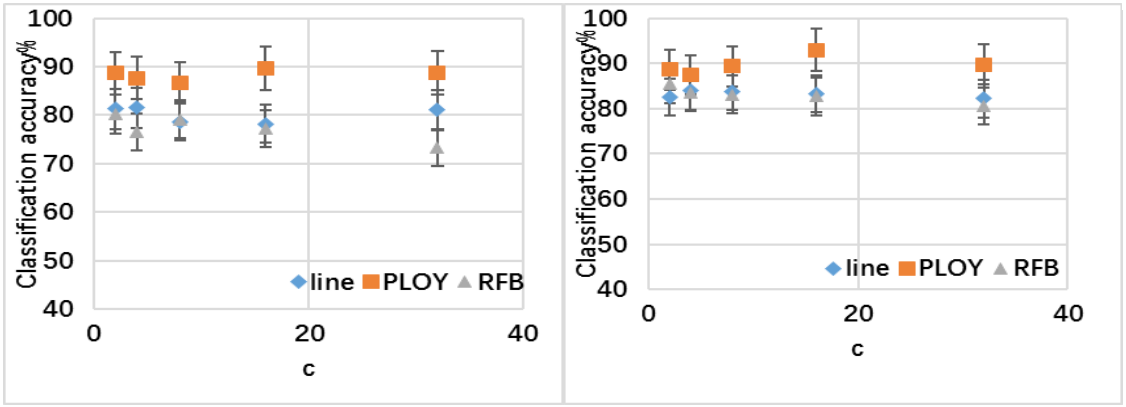


Figure 6: Classification Accuracy with $\varepsilon = 0.25$ **Figure 7:** Classification Accuracy with $\varepsilon = 0.5$

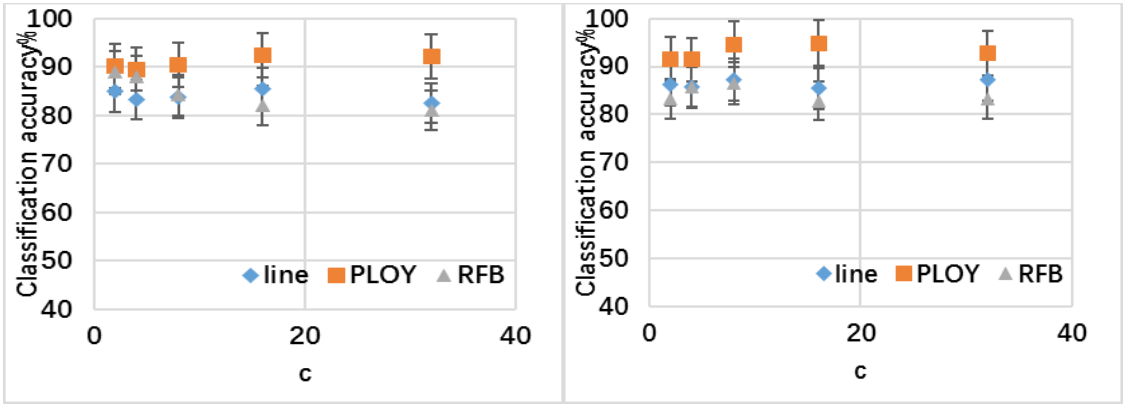


Figure 8: Classification Accuracy with$\varepsilon = 0.75$ **Figure 9:** Classification Accuracy with $\varepsilon = 1$

In above experiment process, the experimental computation time is analyzed and summarized at meanwhile. The mean value of computation time under a different c, is shown inFig. 10 as follows.

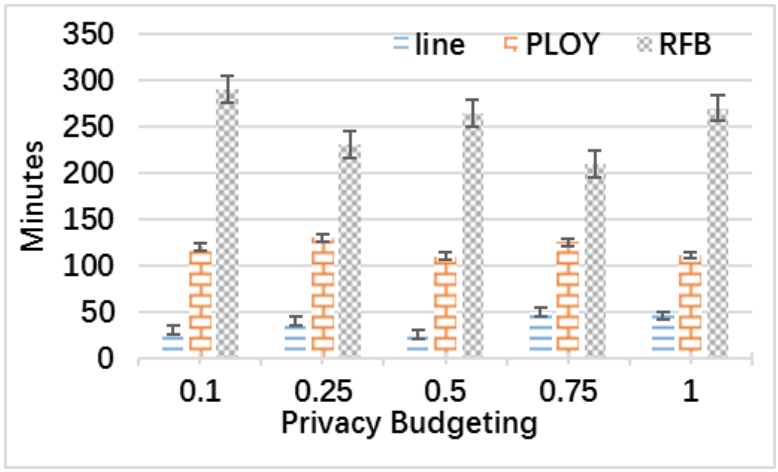


Figure 10: Computation Time of Different Kernel Function Models

Indicated by the above experimental results, the computation time of liner kernel is the fastest while that of Gaussian kernel is the slowest from the view of time. From the perspective of classification accuracy rate, Gaussian kernel is the best among these three kernel functions in the case of no noise. But after the noise is added, the classification accuracy rate of polynomial kernel is the most accurate, the liner kernel take second place and Gaussian kernel is the last.

The classification accuracy rate of polynomial kernel and liner kernel are higher than that of Gaussian kernel in a whole.

As the sensitivity of linear and polynomial kernel is small than that of Gaussian kernel, the smaller the sensitivity of the kernel function is, the smaller the added noise will be. While the kernel function of the Gaussian kernel has a more rigorous mathematical model, a slight change after noise added may result in big error.

5. Conclusion

For the multi-classification dataset, this paper analyzes the privacy protection mechanism of SVM by using differential privacy algorithms. Three different kernel functions, including the linear kernel, the polynomial kernel and Gaussian kernel respectively, are adopted by noise added to the objective function so as to achieve the privacy protection. Experiment results show that the proposed algorithm features a better effect of privacy protection to a certain degree for the three kernel functions. In the future, more efforts will be devoted to improve the classification accuracy of the proposed algorithm.

References

[1] Dwork C. Differential privacy//*Proceedings of the 33rd International Colloquium on Automata, Languages and Programming* (ICALP). Venice, Italy, 2006:1-12

[2] Chaudhuri K, Monteleoni C, Sarwate A D. *Differentially private empirical risk minimization*[J]. Journal of Machine Learning Research, 2011, 12(Mar): 1069-1109.

[3] Rubinstein B I P, Bartlett P L, Huang L, et al. *Learning in a large function space: Privacy-preserving mechanisms for SVM learning*[J]. arXiv preprint arXiv:0911.5708, 2009.

[4] Lin K P, Chen M S. *On the design and analysis of the privacy-preserving SVM classifier*[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(11): 1704-1717.

[5] Li H, Xiong L, Ohno-Machado L, et al. *Privacy preserving RBF kernel support vector machine*[J]. BioMed research international, 2014, 2014.

[6] Sun L, Mu W S, Qi B, et al. *A new privacy-preserving proximal support vector machine for classification of vertically partitioned data*[J]. International Journal of Machine Learning and Cybernetics, 2015, 6(1): 109-118.

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

[8] Yu H, Vaidya J, Jiang X. *Privacy-preserving svm classification on vertically partitioned data*[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 647-656.

[9] Dwork C, F.McSherry, K. Nissim, and A. Smith, ". *Calibratingnoise to sensitivity in private data analysis*," in Theory of Cryptography Conference—TCC, pp. 265–284, 2006.

[10] Hsu C W, Lin C J. *A comparison of methods for multiclass support vector machines*[J]. IEEE transactions on Neural Networks, 2002, 13(2): 415-425.

[11] Zhou S K, Chellappa R. *From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space*[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(6): 917-929.