Characterization of Differentially Private Logistic Regression

By: Shan Suthaharan

S. Suthaharan. 2018. Characterization of Differentially Private Logistic Regression, in Proc. of The ACMSE 2018 Conference, Richmond, KY. March 29-31, 2018. (p.15) 8-pages. ACM.

Abstract:

The purpose of this paper is to present an approach that can help data owners select suitable values for the privacy parameter of a differentially private logistic regression (DPLR), whose main intention is to achieve a balance between privacy strength and classification accuracy. The proposed approach implements a supervised learning technique and a feature extraction technique to address this challenging problem and generate solutions. The supervised learning technique selects subspaces from a training data set and generates DPLR classifiers for a range of values of the privacy parameter. The feature extraction technique transforms an original subspace to a differentially private subspace by querying the original subspace multiple times using the DPLR model and the privacy parameter values that were selected by the supervised learning module. The proposed approach then employs a signal processing technique called signal-interference-ratio as a measure to quantify the privacy level of the differentially private subspaces; hence, allows data owner learn the privacy level that the DPLR models can provide for a given subspace and a given classification accuracy.

Keywords: blind source separation | classification | differential privacy | logistic regression | privacy protections | random forest

Article:

1 INTRODUCTION

The differentially private logistic regression is a useful technique to facilitate secure data sharing - an essential task that enhances interdisciplinary collaborations that are needed for modern scientific applications. In data sharing practices, an owner of a data set generates models and allows users to adopt the models and query the data set. In this application, the users of a data set want to achieve the maximum utility (e.g., high classification accuracy) from the data and the owners of the data want the maximum privacy protection on the data set. For this purpose, a DPLR approach has been proposed by Chaudhuri and Monteleoni [5] by incorporating the fundamental concept of differential privacy proposed by Dwork *et. al.*[6]. Since its introduction, a significant amount of studies have been conducted using this model for achieving privacy strength and prediction/classification accuracy [8, 9, 12, 16]. This is a parametric approach and the selection of its privacy parameter ϵ is a challenging problem. The purpose of the privacy parameter is to achieve an acceptable balance between privacy strength and classification accuracy, and make the model ϵ-differentially private.

One of the observed characteristic of ϵ is that the privacy strength decreases and classification accuracy increases when ϵ increases. Another observed characteristics is that it is possible that the same classification accuracy and distinct privacy strengths can be achieved with multiple values of ϵ. Similarly, we can also observe that it is possible the domain or the range of ϵ is arbitrary which makes it difficult to find the right value for the privacy parameter. These characteristics of ϵ make the selection of a set of suitable values for ϵ much harder. A significant research has been performed to address this problem in various perspectives [1, 2]; however, two main contributions that are closely related to the proposed study are selected and discussed in this section. For example, the authors of [7] studied the contributions of the privacy parameter and proposed a model that provides a balance between the objectives of a data owner and a data user, and studied its effectiveness by selecting several values of privacy parameter.

The authors of [7] also stated that the privacy parameter has been studied using values from 0.01 to 7 in the literature; however, there is no clear explanation for the selection of such as range for ϵ. Similarly, the authors of [11] also studied the contributions of the privacy parameter, and proposed two-parameters approach which adds confidence interval w and confidence level p to estimate the true value of the privacy parameter. Although it provides a wider range for the privacy parameter ϵ, they only focused on the balance between the noise added and the privacy strength, rather than the balance between the utility (e.g., classification) and privacy strength. However, there is a relationship between the noise added and the utility, which was not analyzed in their study.

In the proposed study we introduce a machine learning approach that allows the owner of the data understand the characteristics (privacy strength and classification accuracy) of DPLR with respect to the privacy parameter ϵ and its classification performance. The proposed machine learning approach incorporates supervised learning and feature extraction modules [13]. Given a training data set for a binary classification with labels, the supervised learning module characterizes DPLR for its classification performance on the subspaces of the training data set with respect to a set of randomized values of the privacy parameter ϵ of DPLR. The feature extraction module transforms a subspace to a new ϵ-differentially private subspace by querying the subspace multiple times with the DPLR model and a privacy parameter values that are resulted from the supervised learning module. The well-known signal processing technique - called blind source separation - is then used and signal-interference-ratio [15] between the subspaces and their corresponding ϵ-differentially private subspaces are calculated to quantify the privacy levels. These results can allow a data owner learn the privacy level that a DPLR model can provide for a given subspace and a given classification accuracy. The randomized algorithm (Laplace noise) that is employed in DPLR helps one to perform multiple queries on a subspace and construct a new ϵ-differentially private subspace for privacy protection.

2 PROPOSED METHOD

The proposed method provides a framework that consists of two main modules: a supervised learning module, and a feature extraction module. The supervised learning module performs a

DPLR classification task on a subspace (i.e., a subset of a training data set) using a set of randomly generated privacy parameter values. The feature extraction module transforms a given subspace to a new differentially private subspace that has the same dimension as the input subspace using DPLR model constructed for a privacy parameter. The proposed framework then calculates signal interference ratio between a differentially private subspace and the input subspace so that a set of suitable values for the privacy parameters that give a balance between privacy strength and classification accuracy can be generated.

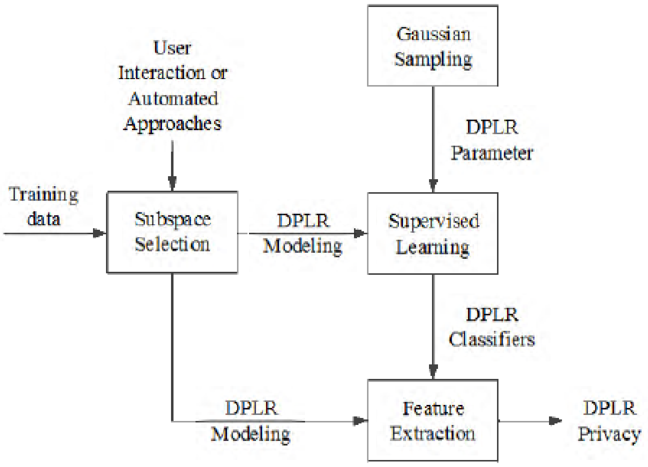

Figure 1: The proposed machine learning framework that can help data owners select the privacy parameter ϵ for DPLR models, analytically.

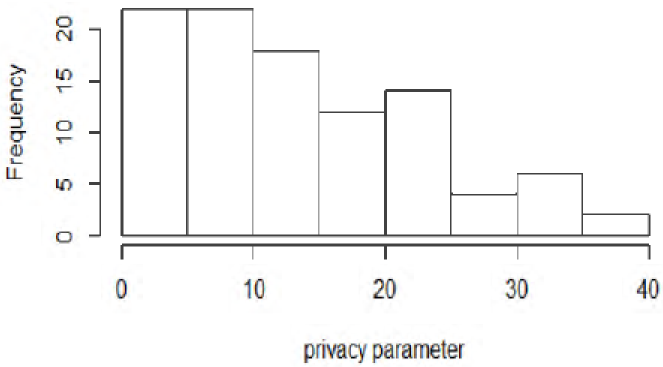

Figure 2: Distribution (in the form of a histogram) of the privacy parameter ϵ by considering it a random variable for supervised learning.

2.1 Supervised Learning Module

Suppose S(cc') is a subspace with class labels c and c', and $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_k\}$ is a set of random numbers drawn from a Gaussian Figure 2: Distribution (in the form of a histogram) of the privacy parameter ϵ by considering it a random variable for supervised learning. distribution, $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, k$. The supervised learning module then accepts an input subspace S(cc') from a training data set and the privacy parameter set ϵ, and applies DPLR modeling to classify the classes c and c'. The subspace is a subset of the training data set; hence, the labels are available for supervised learning. The classification accuracies for the elements of the set ϵ are the output of the module that uses the class labels for learning. Finally, the elements of ϵ that give the maximum classification accuracies are recorded by the supervised learning module. Let us denote this subset of ϵ by e, and assume there are d elements $e_1, e_2, \ldots, e_d$ in the set. Note that we can also find the elements of ϵ that give a given range of classification accuracies, rather than the maximum, requested by the users of the data and DPLR model. The theoretical model of DPLR suggests the smaller the privacy parameter the larger the privacy strength and the larger the privacy parameter the smaller the classification error (or larger the classification accuracy). Therefore, the distribution of the privacy parameter is assumed to follow the Gaussian distribution with mean 0 and variance σ^2. However, to capture the global characteristics of ϵ, it is recommended to select a large value for the variance (e.g., $\sigma = 16$ to satisfy 2σ-rule with a deviation of 32).

2.2 Feature Extraction Module

The purpose of feature extraction module is to construct a new subspace, called ϵ-differentially private subspace from the original subspace S(cc') by querying the original subspace (i.e., the training data set) multiple times using the DPLR model associated with each e_i, $i = 1, 2, . . . ,d$. Let us denote the ϵ-differentially private subspace constructed for e_i by Sei (cc'), where $i = 1, 2, . . . ,d$. It serves the feature extraction objective of the proposed framework. It also adopts the well-known signal processing technique, blind source separation, and constructs signal interference ratio between Si (cc') and S_{ei} (cc') to quantify the privacy strength of the new differentially private subspace for each e_i, $i = 1, 2, . . . ,d$:

$$\rho_i = SIR\left(\boldsymbol{S}(cc'), \boldsymbol{S}_{e_i}(cc')\right), \tag{1}$$

where $i = 1, 2, . . . ,d$ and the unit of this measure is decibel (dB). In this paper, the SIR function available in R-package JADE library, [10] is used. In signal processing [4], the SIR value is interpreted for signal recovery strength as follows: if the value is below 12dB then the recovery of the source signals cannot be recovered from their modulated signals and if the value is above 20dB then the source signal can be definitely recovered from the modulated signals. Similarly, if the value is between 12dB and 20dB then it maybe possible to recover. Hence, in this paper, it is used for privacy measure as: if the value is below or closer to 12dB then the privacy strength is strong, if the value is above or closer to 20dB then the privacy strength is weak, and if the value is between 12dB and 20dB then the privacy strength is moderate. Therefore, the SIR measure is useful to quantify the privacy strength by inputing subspaces.

3 EXPERIMENTAL RESULTS

In this section, the characteristics and the performance of DPLR models are studied and experimented with using a subset of NSLKDD data set that was downloaded and used previously in a research [14]. Since the prposed method is mainly focused on the patterns in the data - rather than in intrusion detection itself - the features are simply represented by variables $\alpha_1, \alpha_2, . . . , \alpha_{14}$, without describing their actual meaning in terms of network intrusion detection. In a previous study, we have used random forest classifiers to classify the same data set and extracted these (14) relevant features. Note that the NSL-KDD data set has altogether 41 features and we were able to extract 14 features as relevant features.

3.1 Single Subspace Analysis

To characterize the classification accuracy and the privacy strength of the DPRL model, a single three-dimensional subspace with binary classes is selected from the NSL-KDD data set. For simplicity, we can assume that the owner of the data wants to study the subspace that is formed by the three features ($\alpha_3, \alpha_4, \alpha_{11}$) with the binary class (9, 10). However, the subspaces can be selected using automated approaches, such as random forest [3], that can output relevance features as a part of its classification objective.

3.1.1 Supervised Learning. A set of random numbers from a Gaussian distribution with mean 0 and a large variance $\sigma^2 = 256$ is generated to define a wider domain so that it can capture a global characteristic of ϵ and its connection with classification accuracy of the DPLR model. The Gaussian random numbers are positive and negative; hence, their absolute values are assigned to the privacy parameter ϵ. Figure 2 shows the histogram of the values generated for the privacy parameter for supervised learning. It clearly provides more values within the range 0 and 7 to satisfy the testing range suggested in [7], and wider range up to 40 to capture the global characteristics with the Gaussian properties.

Figure 3 shows the relationship between the classification accuracy and the privacy parameter ϵ when DPLR is applied to the subspace (α_3, α_4, α_{11}) with the binary class (9, 10). We can clearly see the global characteristics and determine the maximum accuracy of 0.9876543 (99%) can be achieved; however, there are multiple values for ϵ that produce the maximum accuracy:

- e = {14.38466, 15.06771, 16.65124, 17.66038, 17.83741, 19.21488, 19.87036, 20.28188, 20.46009, 21.61631, 22.01760, 23.48987, 24.00954, 24.54474},

where d = 14. Hence, these values can satisfy the users' utility requirements (i.e., classification accuracy). The question now remains is that whether these values will provide privacy strength to satisfy owners' requirements. Figure 4 shows the privacy strengths calculated using SIR measure (which will be discussed later in detail). Figures 3 and 4 confirm that a smaller privacy parameter means a stronger privacy strength, and a larger privacy parameter means a higher classification accuracy.

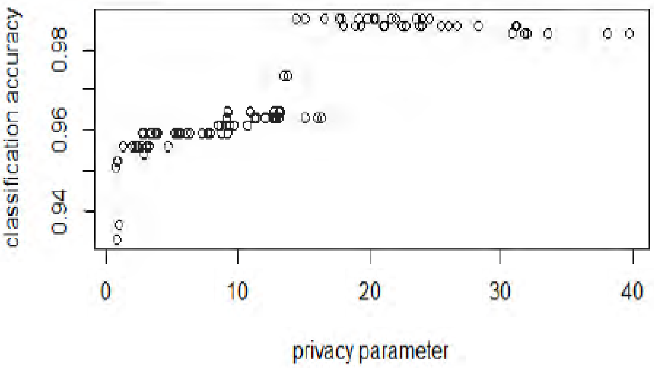

Figure 3: Relationship between DPLR classification and the privacy parameter ϵ for the subspace (α_3, α_4, α_{11}) of NSL-KDD training data set. The maximum accuracy at the mean ϵ value of 19.779333 with standard deviation of 3.200298.

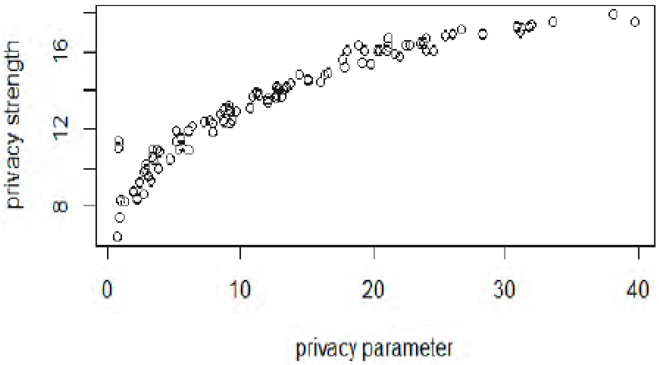

Figure 4: Relationship between DPLR privacy strength (SIR) and the privacy parameter ϵ for the subspace (α_3, α_4, α_{11}) of NSL-KDD training data set. Privacy weakness increases with respect to the increase of privacy parameter.

3.1.2 Feature Extraction. The experimental analysis is performed on three-dimensional subspaces; therefore, three queries are also performed on the same subspace (α_3, α_4, α_{11}) with the binary class (9, 10), using the same DPLR model for each e_i in the set e to extract three features to construct its three-dimensional ϵ-differentially private subspace. The mean of all the values in e is 19.79333, and the standard deviation is 3.200291. Hence, for each e_i, an ϵ-differentially private subspace is constructed using the DPLR model assuming the privacy parameter value of e_i, i = 1, . . . , 14. Figure 5 shows the privacy strength of DPLR around the values of the mean ϵ.

Hence, we have privacy strength for the ϵ values that recorded maximum classification accuracy for the subspace considered.

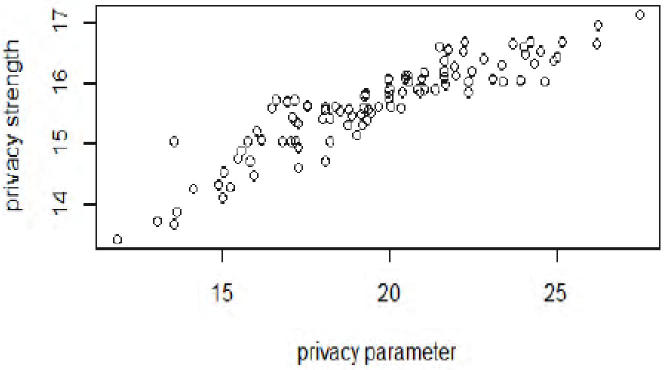

Figure 5: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the maximum classification accuracy - the statistical mean and the standard deviation of ϵ are 19.779333 and 3.200298, respectively.

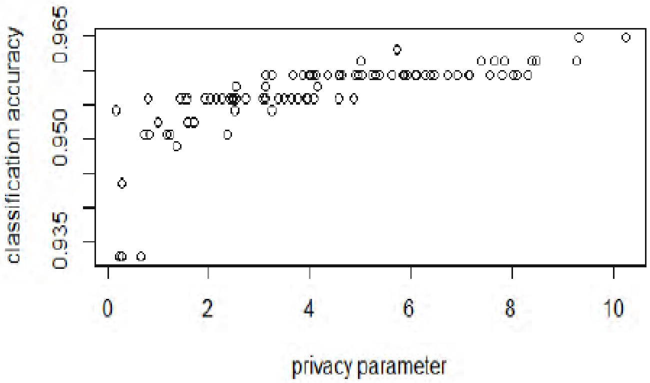

Figure 6: The privacy protection values that give the classification accuracies between 95% and 96% (assumed the user has requested this range) - the statistical mean and the standard deviation of ϵ are 4.45 and 2.43, respectively.

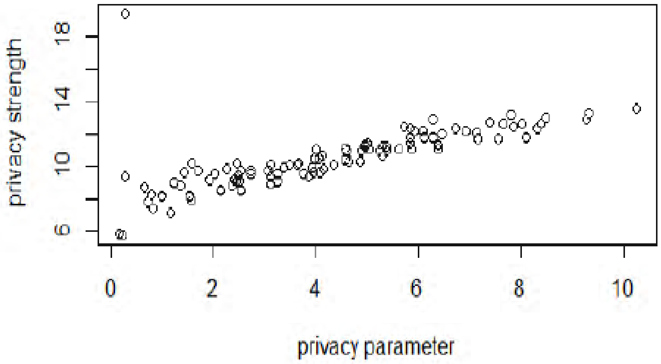

Figure 7: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the classification accuracy range requested by the user. The mean of ϵ is 4.45 and the standard deviation is 2.43.

However, if the user requested classification accuracies between 95% and 96% then the data owner can provide a stronger privacy, because it can be achieved as demonstrated in Figure 6 and Figure 7. The following privacy parameter values can provide DPLR classification accuracies between 95% and 96% as illustrated in Figure 6:

- *e* = {0.642902, 0.753830, 0.818717, 1.263481, 1.877766, 2.170097, 2.362652, 2.656176, 2.6786860, 2.812095, 2.857677, 3.072937, 3.232231, 3.288970, 3.340842, 3.384012, 3.713874, 3.838053, 3.888140, 4.609097, 5.115693, 5.157392, 5.278286, 5.551264, 5.631919, 6.056539, 6.079855, 6.329025, 7.201646, 7.703153, 7.890869, 7.918942, 8.701562, 8.711461, 9.197012}.

This subset of the ϵ values has 35 elements, i.e., $d = 35$. It also has the mean value of 4.45 with standard deviation of 2.43. Hence, it follows the range mentioned for ϵ in [7]. Figure 7 shows the SIR values between 6 and 14 with the center value of 10, which indicates a very strong privacy protection based on the SIR measure.

3.2 Multiple Subspace Analysis

In section 3.1, the supervised learning module and the feature extraction module of the proposed framework were evaluated using the subspace (α_3, α_4, α_{11}) of the NSL-KDD data set with binary classes (9,10) only. In a new experiment, the same subspace is again considered; however, the other classes (0,1), (0,5), (1,2), (1,9), (3,5), (3,9), and (6,8) are also studied. In addition, three other subspaces, (α_3, α_4, α_5), (α_3, α_4, α_7), and (α_4, α_7, α_{10}), are also included in the experiment to study the performance of the proposed analytical framework with DPLR. This experiment can allow data owners understand the characteristics of DPLR with respect to different feature and class characteristics using various values for ϵ of DPLR.

3.2.1 Supervised Learning. In this experiment, the subspace (α_3, α_4, α_{11}) with different binary classes are used to evaluate the performance of DPLR for binary classification. The binary classes considered in this performance analysis are listed in the first column of Table 1. The second column lists the maximum classification accuracies recorded for each classification task. For example, DPLR classifies the classes (1, 9) with about 99% accuracy, whereas it classifies the classes (3, 5) with about 53% accuracy. Since, DPLR's classification performance is poor for the binary class (3, 5), classification accuracies are obtained for other subspaces, (α_3, α_4, α_5), (α_3, α_4, α_7), and (α_4, α_7, α_{10}) other subspaces as well, and listed in the third, fourth, and fifth columns of the table. As we can observe, DPLR can improve it performance and achieve 92% accuracy within the subspace (α_4, α_7, α_{10}) for this binary class. Similarly, the maximum classification accuracy of 81% can be achieved for the binary classes (0, 5) within the subspace (α_4, α_7, α_{10}). Since, the maximum classification accuracies are recorded for the other class pairings within the subspace (α_3, α_4, α_{11}), the rest of the values are listed as dash (−) notation in this table, instead of leaving them blank.

Table 1. Maximum classification accuracies with multiple binary classes and three-dimensional subspaces

Binary Classes	**ACC(α_3, α_4, α_{11})**	**ACC(α_3, α_4, α_5)**	**ACC(α_3, α_4, α_7)**	**ACC(α_3, α_7, α_{10})**
(0,1)	0.977865	–	–	–
(0,5)	0.809062	–	–	0.811912
(1,2)	0.981599	–	–	–
(1,9)	0.991828	–	–	–
(3,5)	0.528645	0.528645	0.783854	0.924479
(3,9)	0.873239	–	–	–
(6,8)	0.870307	–	–	–
(9,10)	0.987654	–	–	–

Some of the classification results are presented in Figures 8 through 11 for performing visual analytics. Figure 8 illustrates the DPLR classification performance when the subspace (α_4, α_7, α_{10}) with the binary class (0,5) is used. It means that it explains the second row of the results presented in Table 1. Only a few values of the privacy parameter ϵ can give the highest classification accuracy 81%, but many values can give the accuracy closer to 80%. However, the maximum classification is achieved when much lower values are assigned to ϵ. Similar characteristics can be seen in Figure 9 when the same subspace with the binary class (3, 5) is used, but in this case the maximum accuracy is much higher (i.e., above 90%) with the lower value of ϵ. Figure 10 shows the classification results of the subspace (α_3, α_4, α_{11}) with the binary class (1, 9). It provides the best results with the classification accuracy of 99%; however, we can observe that this accuracy is achieved at much larger values of ϵ. Figure 11 shows the classification results of the same subspace with the binary class (3, 9). In this case, the accuracy is steadily increasing to about 87% but achieved at very large value of ϵ.

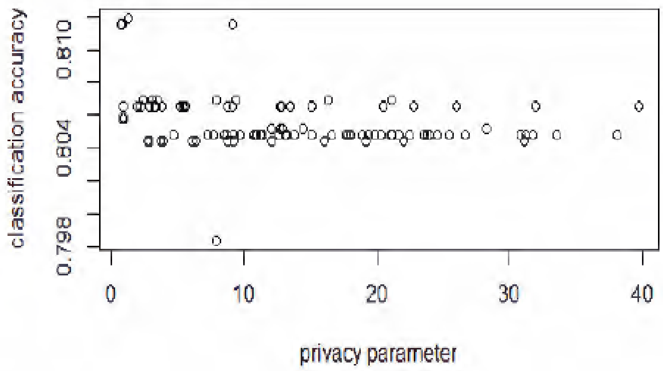

Figure 8: Relationship between DPLR classification and the privacy parameter ϵ for the subspace (α_4, α_7, α_{10}) of NSL-KDD training data set. The maximum accuracy at the mean ϵ value of 1.263481 with standard deviation of 1.

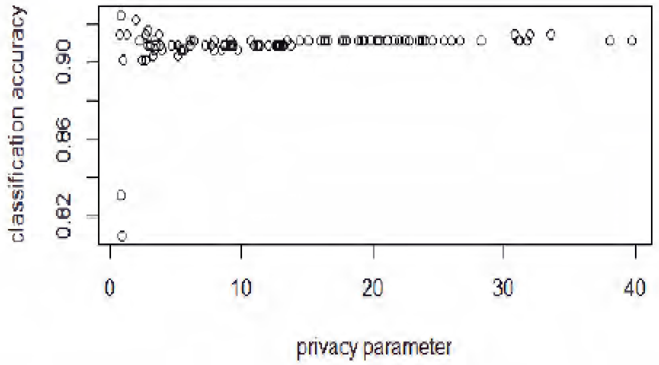

Figure 9: Relationship between DPLR classification and the privacy parameter ϵ for the subspace (α_4, α_7, α_{10}) of NSL-KDD training data set. The maximum accuracy at the mean ϵ value of 0.7538302 with standard deviation of 1.

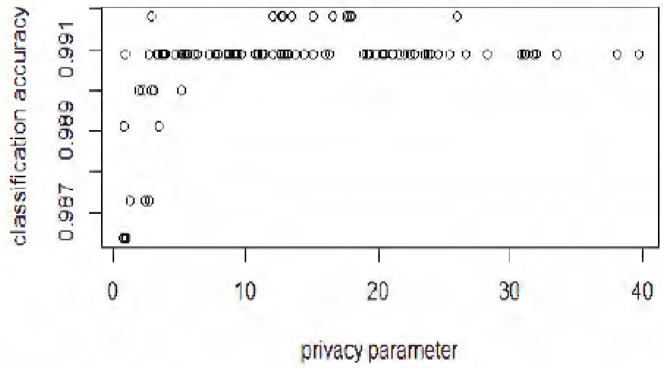

Figure 10: Relationship between DPLR classification and the privacy parameter ϵ for the subspace (α_3, α_4, α_{11}) of NSL-KDD training data set. The maximum accuracy at the mean ϵ value of 14.816290 with standard deviation of 5.400668.

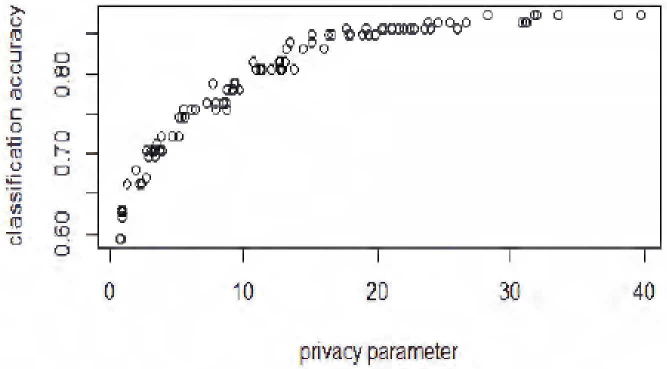

Figure 11: Relationship between DPLR classification and the privacy parameter ϵ for the subspace (α_3, α_4, α_{11}) of NSL-KDD training data set. The maximum accuracy at the mean ϵ value of 33.928860 with standard deviation of 4.283012.

3.2.2 Feature Extraction. Feature extraction is performed, as discussed previously, by running queries multiple times on the training data (subspaces) for each experiment that has distinct subspaces and distinct binary classes. In Table 2, the class labels, the maximum classification accuracies, and the subspaces where these accuracies are recorded are listed in the first three columns, respectively. As previously discussed, multiple ϵ values can provided the same classification accuracy levels; hence, the averages of the ϵ values that give the maximum classification accuracies for the class pairings are listed in fourth column. Similarly, their statistical standard deviations are also calculated and listed in the fifth column of the table. The standard deviation value of 0 indicates the maximum classification accuracy is produced by only a single value of ϵ for that specific binary classes, and it is replaced with 1 in the experimental analysis to obtain a range of values for ϵ. Figure 12 shows a very strong privacy strength for the subspace (α_4, α_7, α_{10}) with the binary class (0,5). It shows the average SIR value of 1.75dB while showing the classification accuracy of 81% in Figure 8. Hence, the users cannot reach more than 81% with using the DPLR that provides the strong privacy protection.

Table 2. Maximum classification accuracies versus SIR with mean and standard deviations of ϵ

Binary Classes	ACC	Subspace	mean(ϵ)	sd(ϵ)	SIR
(0,1)	0.977865	(α_3, α_4, α_{11})	30.80703	0	0.926556
(05,)	0.811912	(α_4, α_7, α_{10})	1.263481	0	1.754037
(1,2)	0.981599	(α_3, α_4, α_{11})	25.385810	4.137129	5.360779
(1,9)	0.991828	(α_3, α_4, α_{11})	14.816290	5.400668	5.314713
(3,5)	0.924479	(α_4, α_7, α_{10})	0.7538302	0	4.873076
(3,9)	0.873239	(α_3, α_4, α_{11})	33.928860	4.283012	36.480850
(6,8)	0.870307	(α_3, α_4, α_{11})	22.693740	1.545736	0.406017
(9,10)	0.987654	(α_3, α_4, α_{11})	19.704820	2.575838	15.648030

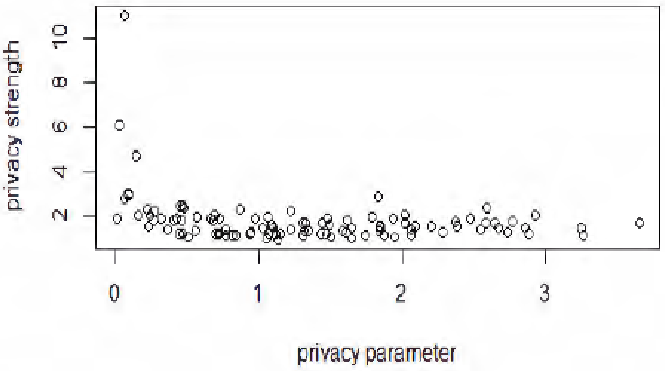

Figure 12: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the classification accuracy range requested by the user. The mean of ϵ is 1.263481 and the standard deviation is 1.

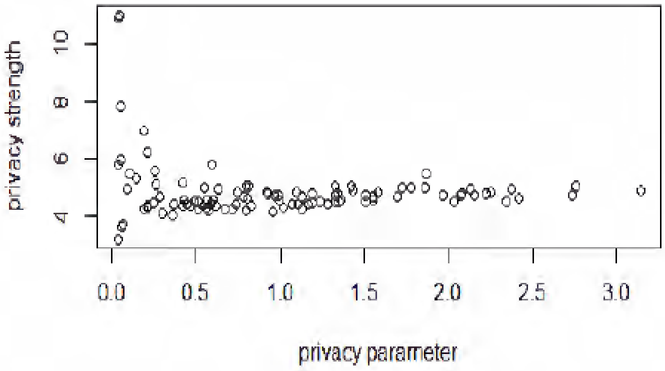

Figure 13: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the classification accuracy range requested by the user. The mean of ϵ is 0.7538302 and the standard deviation is 1.

Figure 13 shows the privacy strength of subspace (α_4, α_7, α_{10}) with the binary class (3,5). The lower average value 4.87dB of SIR indicates the privacy strength of the DPLR model in this subspace with the binary class. In addition, its associated classification results presented in Figure 9 indicate an acceptable maximum classification accuracy of 90%. Therefore, it characterizes the subspace (α_4, α_7, α_{10}) with the binary class (3,5) as the moderate subspace among the ones that are considered in this experiment.

Figure 14 shows the privacy strength of subspace (α_3, α_4, α_{11}) with the binary class (1,9). The low average value 5.31dB of SIR indicates the privacy strength of the DPLR model in this subspace with the binary class. In addition, its associated classification results presented in Figure 10 indicate the maximum classification accuracy of 99% can be easily achieved. Therefore, it characterizes the subspace (α_3, α_4, α_{11}) with the binary class (1,9) as the best subspace among the ones considered in this experiment in terms of both classification accuracy and privacy strength. It also illustrates that larger values (above 7) of the privacy parameter ϵ can give a better balance between privacy strength and classification accuracy.

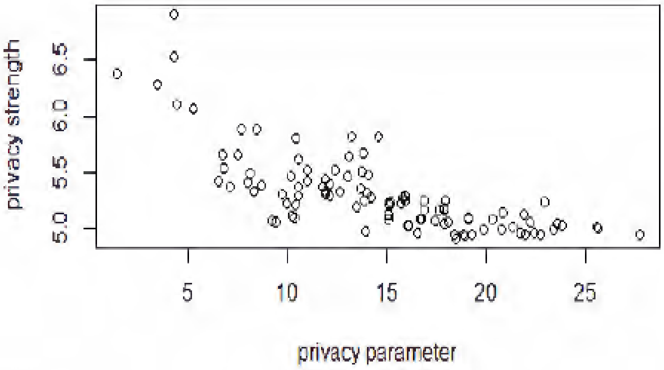

Figure 14: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the classification accuracy range requested by the user. The mean of ϵ is 14.816290 and the standard deviation is 5.400668.

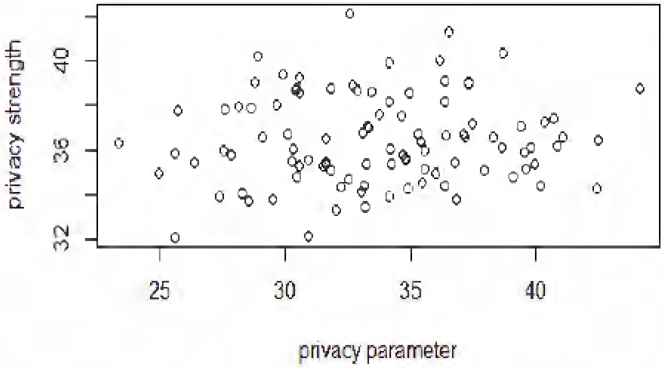

Figure 15: Privacy strength for the values of ϵ that are closer to the mean of ϵ that gives the classification accuracy range requested by the user. The mean of ϵ is 33.928860 and the standard deviation is 4.283012.

Figure 15 shows the privacy strength of subspace (α_3, α_4, α_{11}) with the binary class (3,9). The high average value 36dB of SIR indicates the privacy weakness of the DPLR model in this subspace with the binary class. In addition, its associated classification results presented in Figure 11 indicate the maximum classification accuracy cannot reach more than about 87%. Therefore, it characterizes the subspace (α_3, α_4, α_{11}) with the binary class (3,9) as the weakest subspace among the ones that are considered in this experiment.

4 CONCLUSION

The machine learning can help the data owners to generate DPLR models that can meet the users' expectation for the classification accuracy while satisfying their privacy protection expectation. The machine learning approach can also help them understand the relationship between the privacy parameter and the DPLR models; hence, they can share the data with greater peace of mind. The study also suggested a better privacy parameter values exist outside the regular interval used in the DPLR research, and they can provide a better balance between privacy protection and classification accuracy when DPLR is preferred for a data sharing application. It also provides a flexible approach that can be easily extended to larger-dimensional subspaces or the full feature space through multiple querying on the training data set. We can also extend this study by replacing the simple random number generator module that is used with more sophisticated approaches, such as the Monte Carlo simulation, Markov Chain Monte Carlo algorithm, Gibbs sampling, and Metropolis-Hastings algorithm.

REFERENCES

[1] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and LihuaWang. 2016. Privacy-preserving logistic regression with distributed data sources via homomorphic encryption. *IEICE TRANS. on Information and Systems* 99, 8 (2016), 2079–2089.

[2] Yoshinori Aono, Takuya Hayashi, Le Trieu Phong, and Lihua Wang. 2017. Input and Output Privacy-Preserving Linear Regression. *IEICE TRANSACTIONS on Information and Systems* 100, 10 (2017), 2339–2347.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Jean-François Cardoso and Antoine Souloumiac. 1993. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (radar and signal processing)*, Vol. 140. IET, 362–370.

[5] Kamalika Chaudhuri and Claire Monteleoni. 2009. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*. 289–296.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*, Vol. 3876. Springer, 265–284.

[7] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. 2014. Differential privacy: An economic method for choosing epsilon. In *Computer Security Foundations Symposium (CSF), 2014 IEEE 27th*. IEEE, 398–410.

[8] Zhanglong Ji and Charles Elkan. 2013. Differential privacy based on importance weighting. *Machine learning* 93, 1 (2013), 163–183.

[9] Zhanglong Ji, Xiaoqian Jiang, ShuangWang, Li Xiong, and Lucila Ohno-Machado. 2014. Differentially private distributed logistic regression using private and public data. *BMC medical genomics* 7, 1 (2014), S14.

[10] Jari Miettinen, Klaus Nordhausen, and Sara Taskinen. 2017. Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Journal of Statistical Software* 76 (2017).

[11] Maurizio Naldi and Giuseppe D'Acquisto. 2015. Differential Privacy: An Estimation Theory-Based Method for Choosing Epsilon. *arXiv preprint arXiv:1510.00917* (2015).

[12] Saeed Samet. 2015. Privacy-Preserving Logistic Regression. *Journal of Advances in Information Technology* Vol 6, 3 (2015).

[13] Shan Suthaharan. 2015. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. *Springer* 36 (2015).

[14] Yyyyyy Xxxxx and Zzzzzzz Xxx. yyyy. It anonymizes this paper to support double-blind reviewing process of the conference. *name of the conference* (yyyy).

[15] V Zarzoso and AK Nandi. 1999. Blind source separation. In *Blind Estimation Using Higher-Order Statistics*. Springer, 167–252.

[16] Xu Dong Zhu, Hui Li, and Feng Hua Li. 2013. Privacy-preserving logistic regression outsourcing in cloud computing. *International Journal of Grid and Utility Computing* 4, 2-3 (2013), 144–150.