Privacy Preserving Data Sharing by Integrating Perturbed Distance Matrices

Hanten Chang
Graduate school of Systems and
Information Engineering,
University of Tsukuba, Japan
s1820554@s.tsukuba.ac.jp

Hiroyasu Ando
Faculty of Engineering,
Information and Systems,
University of Tsukuba, Japan
ando@sk.tsukuba.ac.jp

Abstract

Collecting more data is beneficial in machine learning to generate models that are less biased. There are many cases in which pieces of similar data are distributed among organizations and difficult to integrate these data due to issues involving privacy and cost. Integrating these distributed data without delivering the original data leads to the concept of data collaboration, which combines data held by different organizations in a secure manner. We propose a method in which a distance matrix of the original data obtained using common data among organizations is shared to learn neighbor information of the original data. Specifically, the proposed method robustly integrates distributed data of as good quality as connected raw data in cases where the amount of data in each organization is small and the data bias is large. In addition, the proposed method is applicable to data contaminated by noise. To demonstrate the effectiveness of the proposed method, we perform a classification task on open biological data divided into several pieces and show that the classification results for divided data are as precise as when all data is available. Finally, we show that the robustness of the method against noise improves the anonymity of the original data as a side effect.

Introduction

Due to recent developments in technology for data acquisition, a huge amount of data can be concentrated in a single organization. Therefore, a variety of analyses for those data can be done within each organization. However, there are limitations for analyzing data in a single organization. For example, as for clinical data in hospitals, there are rare diseases for which the data are imbalanced, and its amount is small, only in a single hospital. Moreover, there are data biases among hospitals. To overcome these difficulties, integrating such small, biased data from different hospitals should be useful. In reality, clinical data are not centralized but distributed among hospitals due to privacy concerns and communication costs, even if they have a similar data format. Thus, there is a need for a technique to perform integrated analysis on distributed data while securing privacy. One possible technique is secure computation (Chida et al. 2014; Du and Atallah 2001; Nikolaenko et al. 2013; Yao 1986) that uses cryptography. This method makes it possible to calculate target statistics and values while keeping data confidential. On the other hand, there is an overhead to be anonymized in the calculation, and the encryption key should be carefully managed. Another option is federated learning (Bonawitz et al. 2019; Konečný et al. 2016; McMahan et al. 2017), which is a method of learning an integrated model with distributed data. It has been shown that federated learning can work well even when there are many parameters to be learned, as in deep learning (McMahan et al. 2017). Because the integration of models is under the condition of limited tasks, it is necessary to learn other models for other tasks.

Recently, a method of data collaboration analysis has been proposed, where distributed data can be safely integrated without using encryption (Imakura and Sakurai 2019). Even if this method is not using cryptography, it can reduce the risk of estimating the original data. The method learns a transformation of data to a low-dimensional space in which important expressions of the original data, such as neighbor information, are preserved. This procedure is done independently in each organization. Furthermore, it is learned a projection from such low dimensional data to an intermediate representation in a shared space for all organizations. The intermediate representation is provided for all organizations so that various tasks for integrated data are performed by means of the representation.

As a useful property of the data collaboration analysis, it is impossible to restore the original data due to the dimensional reduction. Moreover, each organization independently reduces the dimension of data to preserve anonymity. In practice, Takahashi et al. showed that data collaboration analysis is effective for analyzing clinical data (Takahashi et al.). However, in the proposed method, dimensional reduction is performed in each organization, which means that the data is susceptible to bias and that a sufficient amount of data is required for learning compressed representation. These could be critical issues for distributed data in cases where the data is imbalanced and its size is small.

In this study, we propose a method that has properties desirable for practically integrating data, namely by sharing the

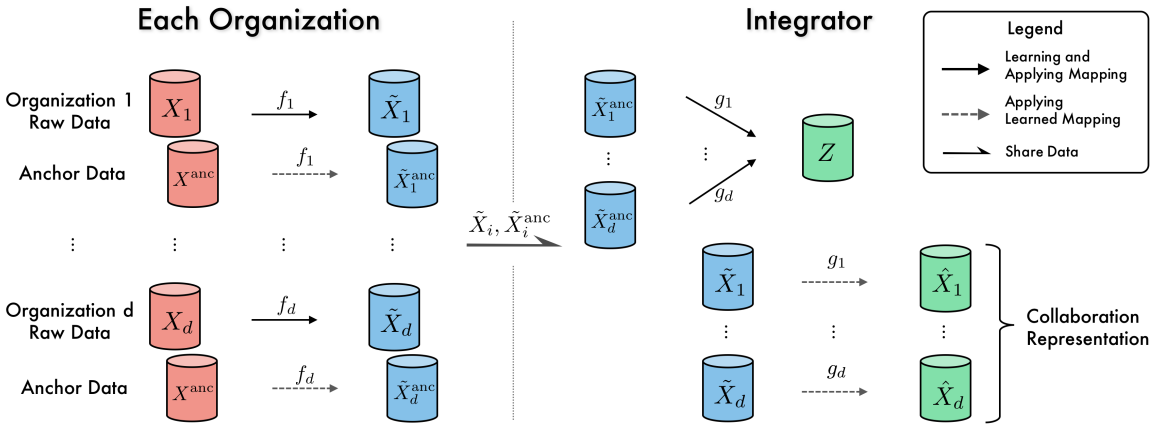

Figure 1: Data Collaboration Analysis Overview

distance matrix of an organization's data. By using anchor data, which is the data shared by all organizations, it is possible to estimate the distance between each organization's data without sharing the data itself. Then, the top-k nearest neighbor graph defined by these estimated distances is used to get the integrated features that express the relationships among all data. The obtained features, called the collaboration representation, can preserve important features of the relations among raw data even if the raw data is not shared directly in estimation. Furthermore, in the proposed method, the following properties can be obtained, which are discussed in the following sections.

1. Robust against biases in data among organizations.
2. The relation, such as k nearest neighbors among all organizations' data, is estimated with high accuracy.
3. Anonymity is improved by adding noise.
4. Robust against changes in the scale of raw data and noise. As a result, it is easy to improve data anonymity.

The rest of this paper proceeds as follows: the data collaboration analysis and one of the graph embedding methods, Node2vec (Grover and Leskovec 2016), are presented in Section 2. In Section 3, we proposed a novel method of data collaboration analysis that shares the distance matrix of each organization's data with the integrator, that is, an organization integrates distance matrices and estimates common features. In Section 4, we evaluate the accuracy of the proposed method for distributed data. Finally, we conclude and discuss the experimental results and the proposed method in Section 5.

Preliminaries

Data collaboration analysis with anchor data

In this section, we describe one type of data collaboration analysis proposed in (Imakura and Sakurai 2019). A schematic illustration is shown in Figure 1. The goal of data collaboration analysis is estimating a common representation of the original data, which are not centralized but distributed in d different organizations. The method assumes that the i-th organization has n_i samples with m-dimensional data as

$$X_i = [x_{i1}, x_{i2}, ..., x_{in_i}] \in \mathbb{R}^{m \times n_i}, \quad (1 \leq i \leq d).$$

The goal of this method is to compare X_i by sharing mapped data

$$\tilde{X}_i = f_i(X_i) \in \mathbb{R}^{\tilde{m}_i \times n_i}.$$

with the integrator where f_i is learned by using each organization's data without that of other organizations. Additionally, a common representation, called a collaboration representation, is described as follows.

$$\hat{X} = [\hat{X}_1, \hat{X}_2, ..., \hat{X}_d] = [g_1(\tilde{X}_1), g_2(\tilde{X}_2), ..., g_d(\tilde{X}_d)] \in \mathbb{R}^{\hat{m} \times n}.$$

The collaboration representation is estimated in a common space in which features in the original data, such as local neighbors of data in each organization, are preserved. Here, $\sum_i n_i = n$, and f_i is a map which extracts features of data as in dimensional reduction methods (Cunningham and Ghahramani 2015). Specifically, LLE (Roweis and Saul 2000) and LPP (He 2005) are used for preserving local neighbors of original data. Therefore, it often assumes $\hat{m} < m$. It also assumes that $f_i(x) \neq f_j(x)$ $(i \neq j)$, since the map f_i is learned in each organization independently. Under the assumption, g_i, satisfying $g_i(f_i(x)) \approx g_j(f_j(x))$ $(i \neq j)$, is estimated. Note that $\hat{X}_i$ is not an approximation of X_i.

In the remainder of this section, we describe how to estimate g_i. In general, estimating g_i is difficult, because there is no common information among the organizations. Therefore, the data collaboration analysis prepares pseudo-data called anchor data which is shared among all the organizations. The anchor data is generated by random numbers or open data for securing privacy. Let the dimension of the anchor data be the same as that of the original data, and the number of anchor data samples be r. The anchor data is described as follows.

$$X^{\text{anc}} = [x_1^{\text{anc}}, x_2^{\text{anc}}, ..., x_r^{\text{anc}}] \in \mathbb{R}^{m \times r}$$

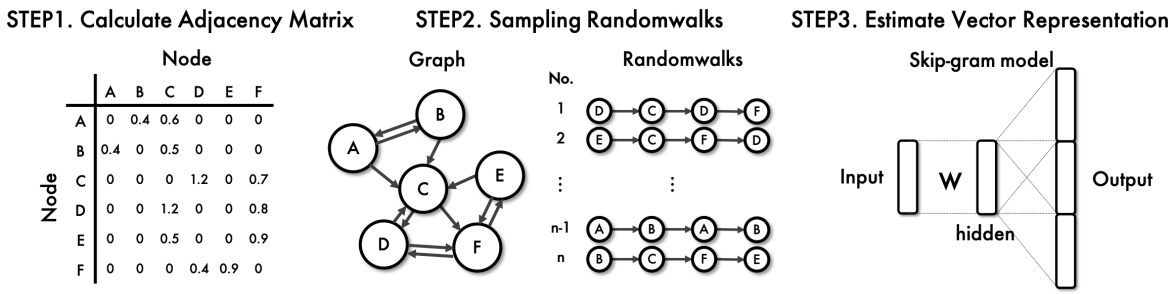

Figure 2: Skip-Gram Model Based Graph Embedding

The data X^{anc} is shared among all organizations. And, each organization applies f_i to the anchor data. Then, the following is obtained:

$$\tilde{X_i}^{\text{anc}} = f_i(X^{\text{anc}}) \in \mathbb{R}^{\tilde{m} \times r}.$$

Since $\tilde{X_i}^{\text{anc}}, (i = 1, ..., d)$ are the representations generated from the same data, both a common space $Z \in \mathbb{R}^{\tilde{m} \times r}$ for $\tilde{X_i}^{\text{anc}}$ and g_i to satisfy $Z \approx g_i(\tilde{X_i}^{\text{anc}})$ can be estimated.

In (Imakura and Sakurai 2019), a method to estimate Z and g by solving the minimal perturbation problem has been proposed as follows:

$$\min_{Z, g_1, g_2, ..., g_d} \sum_{i=1}^{d} \|Z - g_i(\tilde{X_i}^{\text{anc}})\|_{\text{F}}^2.$$

By assuming that g_i is linear, this problem can be solved by singular value decomposition (SVD), so that Z and g can be estimated simultaneously. Finally, we obtain $\hat{X}_i, \hat{X}$ by applying g_i to $\tilde{X}_i$. If f_i is a method preserving local neighbors such as LPP, local neighbors in the original data are expected to be preserved in $\hat{X}$. Estimating Z and g_i can be done by other methods. For example, Takahashi et al. proposed the method to estimate Z by a graph embedding method and g_i by feedforward neural network (Takahashi et al.).

In practice, each organization learns and shares $\tilde{X}_i$ and $\tilde{X_i}^{\text{anc}}$ with a data integrator. And, using these data, the integrator estimates g_i by the above procedure. After estimating g_i, the integrator applies g_i to $\tilde{X}_i$ to generate collaboration representation $\hat{X}$. Finally, this representation is provided to all organizations. In this way, the data collaboration analysis is able to learn the map of data to a common space without centralizing all original data by means of the anchor data.

Node2vec

Node2Vec is one of the methods called graph embedding (Cai, Zheng, and Chang 2017; Goyal and Ferrara 2017) and network embedding (Cui et al. 2017), which learns a vector representation of the graph nodes. Since this method is able to preserve local feature flexibly by generating a graph that expresses local features, such as local neighbors of data. Among the graph embedding methods, DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) and Node2Vec (Grover and Leskovec 2016) are methods based on the model for learning vector representation of words called Skip-Gram (Mikolov et al. 2013). DeepWalk and Node2Vec apply the Skip-Gram model to a sequence of graph nodes called a random walk, which is generated by a transition of nodes based on probability and weighted by graph edges. We explain this procedure as follows.

Let $\tilde{W} \in \mathbb{R}^{r \times r}$ be the matrix representation of a weighted graph that has r nodes: $\tilde{v}_i (i = 1, 2, ..., r)$. And, let the non-negative edge weight between the s-th node $\tilde{v}_s$ and the t-th node $\tilde{v}_t$ be $\tilde{W}_{st}$, which are the elements of $\tilde{W}$. The random walk on the graph is generated by following the distribution as:

$$\Pr(c_j = \tilde{v}_t | c_{j-1} = \tilde{v}_s) = \begin{cases} \frac{\pi_{st}}{C} & \text{if } \tilde{W}_{st} > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Here, c_0 is the initial node randomly selected from the graph nodes. c_j is j-th node in the random walk, C is a normalizing constant, and π_{st} is the probability of transition from node $\tilde{v}_s$ to $\tilde{v}_t$, defined as $\pi_{st} := \alpha_{pq}(c_{j-2}, t)\tilde{W}_{st}$. α_{pq} is defined as follows.

$$\alpha_{pq}(c_{j-2}, t) = \begin{cases} \frac{1}{p} & \text{if } d_{c_{j-2}t} = 0 \\ 1 & \text{if } d_{c_{j-2}t} = 1 \\ \frac{1}{q} & \text{if } d_{c_{j-2}t} = 2, \end{cases}$$

where $d_{c_{j-2}t}$ is the number of nodes in the shortest path from c_{j-2} to $\tilde{v}_t$. Note that $d_{c_{j-2}t}$ must be one of $\{0, 1, 2\}$. Here, p is called the return parameter representing how easy it is to return to the original node, and q is called the in-out parameter representing how easy it is to leave c_{j-2}. An overview of the Skip-Gram model based the graph embedding method is shown in Figure 2. Following this distribution, nodes are sampled iteratively l times to obtain the random walk $c_0, c_1, c_2, ..., c_l$. After generating a sufficient number of random walks, it can estimate the vector representation of each node by applying random walks to the Skip-Gram model. In the context of data collaboration analysis, Takahashi et al. proposed to use Node2Vec to preserve local neighbors by vectorizing a top-k nearest neighbor graph in which one node has edges to k nearest neighbor nodes (Takahashi et al.).

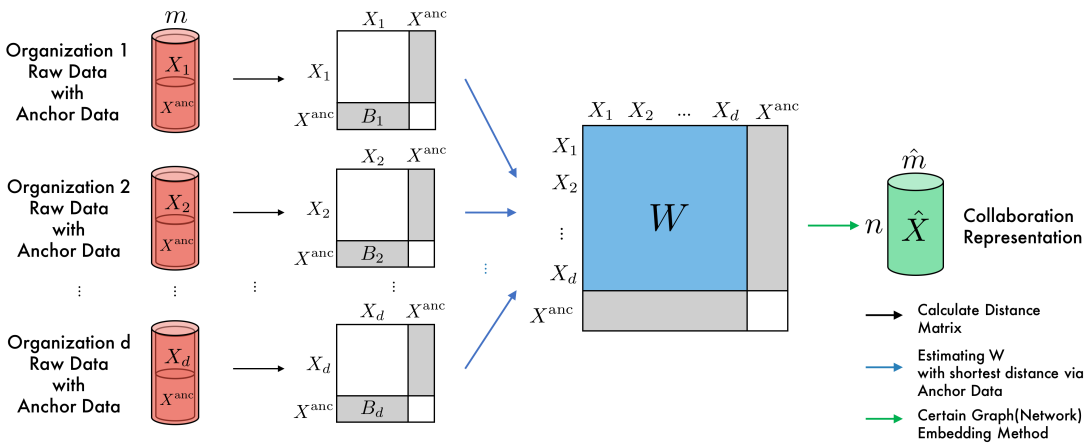

Figure 3: Proposed Method Overview

Proposed Method

Basic concept

In the data collaboration analysis, the anchor data, instead of the original data that is difficult to share, is used to learn the transformation to the representation in a common space, which preserves certain features of the original data, such as local neighbors. Following this idea, we propose an alternative method for data collaboration analysis. Instead of sharing $\tilde{X}$ and $\tilde{X}^{\text{anc}}$, our method requires each organization to share a distance matrix whose elements correspond to the distances between the original data in each organization and the anchor data with the integrator. According to the distance matrix, the integrator estimates an integrated distance matrix whose elements correspond to distances among the original data in all organizations via the anchor data and also estimate a top-k nearest neighbor graph among all organizations. Finally, applying the estimated graph to Node2Vec, we get a vector representation of all nodes. The overview is shown in Figure 3. In the framework of data collaboration analysis (Imakura and Sakurai 2019), our method can be regarded as using f for mapping from the original data to the distance matrix between the original data and the anchor data, and g for mapping from the integrated graph which is expected to have the top-k nearest neighbors of all nodes to the collaboration representation.

Since there is a risk of inverse estimation of the original data by preserving all distances between the original data and the anchor data, we expect to apply some privacy-preserving pre-processing (Aggarwal and Philip 2008; Wagner and Eckhoff 2018) to the original data, such as preserving k-anonymity (Sweeney 2002) or adding noise(Agrawal and Srikant 2000). In general, data with added noise is useless for further analysis, such as machine learning. However, in the case of the proposed method, it is useful even for the machine learning task because the proposed method uses information from the top-k nearest neighbors, which is determined only by the order of distances. That is, it is robust against scale and bias by noise. This is described later in detail.

Sharing B data collaboration analysis baseline

Let the merged d-distributed original data be:

$$X = [X_1, X_2, ..., X_d] = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n}, \quad (1 \leq i \leq d).$$

In the proposed method, we consider the graph with weights corresponding to the distance of n data in X. Therefore, we estimate the matrix representation of weighted graph $W \in \mathbb{R}^{n \times n}$ and estimate collaboration representation $\hat{X}$ by applying W to the graph embedding method. However, there is no connection among d data, since they are not able to be shared as the original data. In this paper, we integrate distance matrices whose elements correspond to the distances between the anchor data X^{anc} and the original data X_i in each organization,

$$B_i \in \mathbb{R}^{n_i \times r} \quad (1 \leq i \leq d)$$

instead of the original data and $\tilde{X}_i$. From this information, we estimate a matrix W. The element W, e.g., W_{st} represents a distance between the s-th data point of X and the t-th data point of X, in terms of the shortest conditioned distance via the anchor data. Since the anchor data is shared over all organizations, the matrix W can be estimated more precisely by using the anchor data. An overview of the proposed method regarding distance is shown in Figure 4a. Note that the distances among the same organization's data can be estimated without using anchor data. However, we use the distances via the anchor data to estimate the distance among the same organization's data to evaluate fairly with other organizational data. We assume that the distance takes a positive value, like the Euclidean distance and the Mahalanobis distance. Due to the triangle inequality, it is obvious that the shortest distance between the two data points via the anchor data includes only one anchor data point. We can determine the shortest distance via anchor data as follows:

$$W_{st} = \min_u \left(b_s(u) + b_t(u) \right),$$

where $b_s(u)$ is the distance between x_s and the u-th anchor data point. In this way, the shortest distance can be calculated by simple summation, and this calculation requires less computational cost than solving shortest path problems.

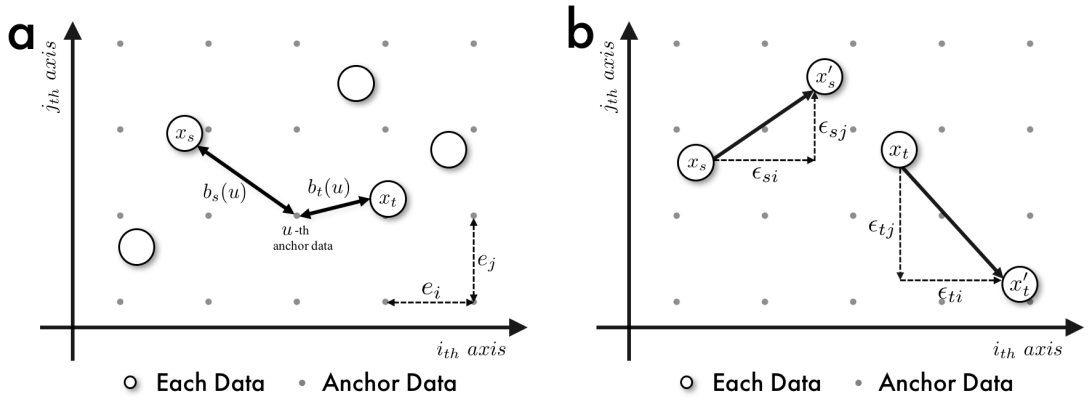

Figure 4: a: Distance via Anchor Data, b: Perturbation of Data

After W is estimated, we transform W to have a property that we want to preserve, such as local neighbors, and apply Node2Vec to the transformed matrix. In preserving local neighbors, we transform W to a W^k that has only k-NN elements of data and apply the Node2Vec function G to get $\hat{X} = G(W^k)$. The dimension of collaboration representation $\hat{m}$ is a hyperparameter. In this way, we can estimate and preserve local neighbors among each organization's data directly without integrating all original data.

This method has several desirable characteristics; namely, the proposed method is affected not by selection bias and data size in each organization but by the total data size in all organizations. However, there is a risk of inverse estimation of original data, since the method uses all distances among the original data and the anchor data. Therefore, we are supposed to add a sufficient amount of noise to the original data, as in Figure 4b, to reduce the risk. On the other hand, that noise can reduce the effectiveness of the collaboration representation. However, in the proposed method, it is necessary to estimate only the local neighbors. That is, it is possible to estimate the relative distance relationship robustly against adding noise, as described in the Experiments section.

Anchor Data

In the proposed method, the distribution of the anchor data is critical. The finer the anchor data is placed, the smaller the estimation error of the distance. Therefore, as in Figure 4a, we consider a lattice of anchor data in each dimension before adding noise. Let the interval of anchor data in the l-th dimension be e_l, such that anchor data in the l-th dimension is set at e_l interval. e_l is determined based on the smallest positive distance in the l-th dimension of the original data. Similarly, we set anchor data in all dimensions. Finally, we integrate all dimensions of anchor data using the Cartesian product. If the anchor data are normalized, taking values from 0 to 1 in each dimension, the number of data points is $r = \prod_{l=1}^{\hat{m}} \lfloor \frac{1}{e_l} \rfloor$. And, we set $e_l = t \cdot dist_l$, where $dist_l$ is the minimum positive difference in the l-th dimension. The parameter t determines the precision of the anchor data, and in the case of $t = 1$, there is at least one anchor data point between any pairs of sample data except for the identical data. To summarize the above, there are three parameters in the method: k, which means local neighbor range, $\hat{m}$, which means dimension of collaboration representation, and t, which means precision of anchor data.

Experiments

Experiments settings

We apply our method to two popular classification datasets in the UCI machine learning repository (Dua and Graff 2017), and compare the results with those by the method proposed in (Imakura and Sakurai 2019) (called the SVD-based method). In experiments, we use the Iris dataset, which contains 150 samples in a 4-dimensional dataset to classify into 3 classes, and the Breast Cancer Wisconsin (original) dataset, which contains 699 samples in a 9-dimensional dataset to classify into 2 classes. We normalize each feature in each dataset to have a value in the range from 0 to 1. We evaluate each method regarding pseudo-distributed datasets, which is given by dividing the original dataset into 5 datasets. For comparison, we also investigate the case that all data can be used to train the model, and the case that only one divided dataset in each organization can be used.

In the SVD-based method, we assume the first mapping f as LPP, which is to preserve one nearest neighbor to evaluate, even though the sample size is small. In addition, although the dimensionality of the original datasets should be reduced to protect the security of data by mapping f, the experiments are performed without reducing dimensionality, in order to investigate the ability of expression. In the proposed method, we consider the distance as the Euclidean distance, and Node2Vec preserves the 10 nearest neighbors by setting W^k to have the value one in 10-NN at each data point and zero for others. We also set $p = q = 1$, the node sequence length to 10, the window to 4, the number of random walks to 1000, and $\hat{m}$ to 64. We evaluated each method by average accuracy over 5-fold cross-validation using each collaboration representation as input to a logistic regression with L2 penalty.

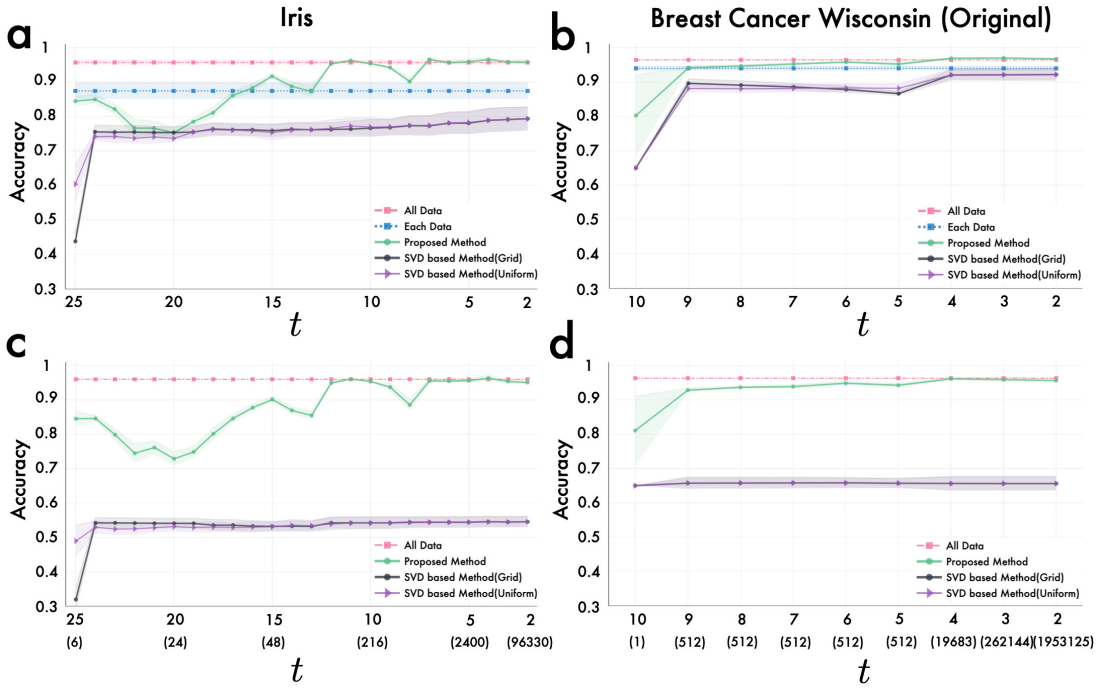

Figure 5: a, b: randomly split dataset result ($d = 5$), c, d: k-means split dataset result ($d = 5$), anchor data size is represented by (x) at the label. In a and c, b and d, the same anchor data is shared, respectively. All the results are averages over ten trials.

Results

Accuracy for Distributed Data Let the number of divisions be 5, and generate pseudo independent datasets in two ways: The first is to divide the original dataset at zrandom (random split case). The second is to divide the original dataset by having a bias at each organization. In the latter case, we use k-means ($k = 5$) to divide the original dataset into 5 datasets that have a bias at each dataset (k-means split case). This makes it possible to conduct experiments in situations where it is difficult to train general machine learning models, as the datasets of each organization are biased. Because in k-means split case, not all divided datasets have all class samples, we do not experiment in the case only one divided dataset in each organization can be used.

We set t to be from 25 to 2 by 1, but in the Breast Cancer Wisconsin (original) dataset, we set t to be from 10 to 2 by 1, because the size of the anchor data is same as 1 when $t \geq 10$. Furthermore, we also examine the case where anchor data is generated from the uniform distribution, which has a value in the range from 0 to 1 as in (Imakura and Sakurai 2019).

The result of the random split case is shown in Figure 5a, b, and the result of the k-means split case is shown in Figure 5c, d. As shown in Figure 5, the accuracy of the SVD-based method between the anchor data generated at random, and the anchor data placed in a lattice does not change much. The anchor data placed in a lattice can also be the anchor data of the data collaboration analysis.

Figure 5a and b show that the accuracy of both the proposed method and the SVD-based method tends to improve with an increase of anchor data size. Furthermore, in both datasets, it is clear that the accuracy of the proposed method is much higher than that of the SVD-based method and is as good as in the case where all data can be obtained.

The major difference between Figure 5a, b, and Figure 5c, d is that the SVD-based method has a smaller improvement in accuracy even if t becomes smaller, increasing the amount of anchor data. Because the SVD-based method estimates collaboration representation from all mapped datasets, which are mapped by f_i learned at each organization, it is difficult to estimate common manifolds if the dataset in each organization is largely biased.

On the other hand, in the proposed method, the relationship among each dataset via the anchor data is directly estimated, which means the bias of the dataset in each organization has no effect on the accuracy of the collaboration representation. Thus, in the case where each organization has a bias, it turns out that the proposed method can robustly estimate collaboration representation.

Accuracy for Distributed Noisy Data As a second experiment, noise is added to the original datasets, and the accuracy in the case where the raw data is more anonymized by noise perturbation is investigated.

The original data is divided into 5 datasets, and t is varied from 3 to 10 by 0.5 for the Iris dataset and from 3 to 9 by 0.5 for the Breast Cancer Wisconsin (original) dataset where the anchor data size is larger than 1. Noise is added to each dimension independently, and each noise is generated from $\mathcal{N}(0, \epsilon \cdot dist_l)$, where $dist_l$ is the minimum distance in the l-th dimension. In order to investigate the dependence of the accuracy on noise, we place the anchor data before adding

noise to the dataset. After placing the anchor data, we add the noise and renormalize the dataset from 0 to 1. As shown in Figure 4b, $\epsilon_{sj} \sim \mathcal{N}(0, \epsilon \cdot dist_j)$ is added to the s-th sample in the j-th dimensional direction. We evaluate accuracy by setting ϵ from 0 to 5 by 0.5 and show the results in Figure 6. Figure 6a and b are heat maps of the accuracy of the proposed method for the both datasets, and Figure6c and d show those of the accuracy of the SVD-based method.

In both the proposed method and the SVD-based method, it can be seen that the accuracy is better when t is smaller, as well as when ϵ is smaller in both datasets. Although in the Breast Cancer Wisconsin (Original) dataset, the number of anchor data points is unchanged when t is larger than 5, the accuracy becomes worse as t becomes smaller for both methods. This suggests that if the number of anchor data points is the same, it is preferable to have anchor data with wider intervals.

It can be seen that the accuracy of the proposed method is generally superior to that of the SVD-based method, even in the situation where noise is added. In addition, the proposed method tends to be more robust against greater noise than the SVD-based method. One of the reasons for the robustness is that converting the estimated W to a top-k nearest neighbor graph in the proposed method is not so much affected by additional noise in terms of preserving the relative distance. In addition, the estimated top-k nearest neighbor graph does not necessarily need to be close to that obtained from the original data due to the Node2Vec method, because there can exist random walks, which may result in two nodes which are neighbors in the original graph but not neighbors in the estimated graph. As a result, the neighboring relation between those nodes can be learned by the Node2Vec.

Furthermore, in the case of Breast Cancer Wisconsin data, adding a small noise enhances the accuracy of estimation, but then adding further noise degrades it, as shown in Figure 6b. This phenomenon occurs because adding an optimal amount of noise increases paths between neighboring nodes. It results in a sufficient amount of information about neighboring nodes in the original graph. However, further noise reduces that neighboring information so that learning fails.

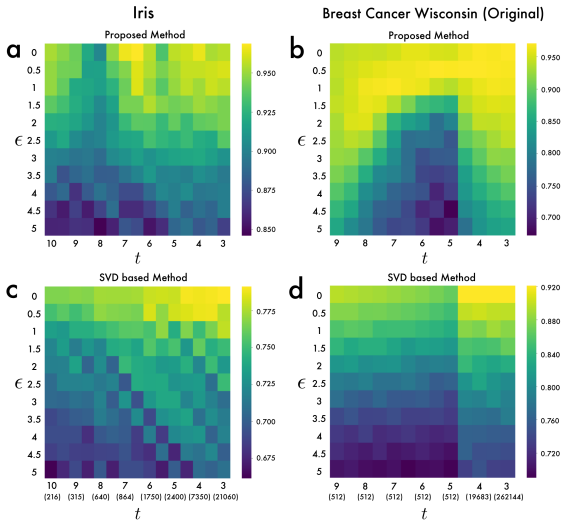

Figure 6: Heat map of accuracy with respect to both t and ϵ. a and b show the results of the proposed method, while c and d show results of the SVD-based method. All results are averaged over 10 trials.

Discussion and Conclusion

In this paper, we proposed an novel framework of data collaboration analysis by using the distance matrices via the anchor data, which makes it possible to learn a representation of neighboring relations among data in all organizations. In this method, each organization shares the distance matrix between its data and the anchor data with the integrator. Then, the integrator estimates the top-k nearest neighbors graph to estimate collaboration representation. In addition, the proposed method is insensitive to biases among data.

On the other hand, the method has a high computational cost for calculating the distance matrix, so that it is not suitable for large datasets. However, the data collaboration analysis is beneficial in the case where a small amount of data is distributed among different organizations, and sharing that data is not possible in terms of protecting privacy. Therefore, the proposed method is adequate for integrating small rare data while preserving privacy. Furthermore, in the case where the distributed data in each organization is biased, data collaboration analysis is helpful, but SVD-based analysis cannot easily learn a universal model due to the bias. The proposed method is robust against data bias and makes data collaboration analysis possible for that biased data.

Regarding the anonymity of the proposed method, the distance matrix removes all information for the data except distance so that it is suitable for guaranteeing anonymity. Although increasing the number of anchor data points improves the accuracy of estimation for classification, there is a risk for decryption of the distance matrix once the anchor data is leaked. To overcome this drawback, our method proposes adding noise. That noise does not degrade the accuracy of estimation and also protects privacy as shown in the experiments. Similarly, pre-processing raw data for guaranteeing anonymity, say preserving K-anonymity (Sweeney 2002) before calculating the distance matrix is effective and applicable for the proposed method. As a further improvement, it can be considered applying existing notions with some theoretical guarantees such as differential privacy (Dwork 2006; McSherry and Mironov 2009) to our method. In addition, it is possible to consider practical procedures such as random shuffling of the index of the distance matrix and sending the shuffled index with encryption, distinguishing a provider of the anchor data from the data integrator for avoiding estimation of raw data.

Acknowledgements

The present study is supported in part by the New Energy and Industrial Technology Development Organization (NEDO) and the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research No. 19K12198.

References

Aggarwal, C. C., and Philip, S. Y. 2008. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*. Springer. 11–52.

Agrawal, R., and Srikant, R. 2000. Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, 439–450. ACM.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konecný, J.; Mazzocchi, S.; McMahan, H. B.; Overveldt, T. V.; Petrou, D.; Ramage, D.; and Roselander, J. 2019. Towards federated learning at scale: System design.

Cai, H.; Zheng, V. W.; and Chang, K. C. 2017. A comprehensive survey of graph embedding: Problems, techniques and applications.

Chida, K.; Morohashi, G.; Fuji, H.; Magata, F.; Fujimura, A.; Hamada, K.; Ikarashi, D.; and Yamamoto, R. 2014. Implementation and evaluation of an efficient secure computation system using 'R' for healthcare statistics. *Journal of the American Medical Informatics Association* 21(e2):e326–e331.

Cui, P.; Wang, X.; Pei, J.; and Zhu, W. 2017. A survey on network embedding.

Cunningham, J. P., and Ghahramani, Z. 2015. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research* 16:2859–2900.

Du, W., and Atallah, M. J. 2001. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, 13–22. ACM.

Dua, D., and Graff, C. 2017. UCI machine learning repository.

Dwork, C. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, 1–12. Berlin, Heidelberg: Springer-Verlag.

Goyal, P., and Ferrara, E. 2017. Graph embedding techniques, applications, and performance: A survey. *CoRR* abs/1705.02801.

Grover, A., and Leskovec, J. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 855–864. New York, NY, USA: ACM.

He, X. 2005. *Locality Preserving Projections*. Ph.D. Dissertation, Chicago, IL, USA. AAI3195015.

Imakura, A., and Sakurai, T. 2019. Data collaboration analysis for distributed datasets.

Konečný, J.; McMahan, H. B.; Yu, F. X.; Richtarik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

McSherry, F., and Mironov, I. 2009. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 627–636. ACM.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Nikolaenko, V.; Weinsberg, U.; Ioannidis, S.; Joye, M.; Boneh, D.; and Taft, N. 2013. Privacy-preserving ridge regression on hundreds of millions of records. In *2013 IEEE Symposium on Security and Privacy*, 334–348. IEEE.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, 701–710. New York, NY, USA: ACM.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE* 290:2323–2326.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.

Takahashi, Y.; Kagawa, R.; ten Chang, H.; Ando, H.; Imakura, A.; Okada, Y.; Tsurushima, H.; Suzuki, K.; and Sakurai, T. Reality enhances machine-learning medical diagnosis ability: the role of artificial features for patients. in prep.

Wagner, I., and Eckhoff, D. 2018. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)* 51(3):57.

Yao, A. C.-C. 1986. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, 162–167. IEEE.