Privacy-Preserving Sharing of Horizontally-Distributed Private Data for Constructing Accurate Classifiers

Vincent Y. F. Tan [1] See-Kiong Ng [2]

[1]Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

[2]Institute for Infocomm Research (I^2R), Singapore

PinKDD 2007: 1st ACM SIGKDD Int. Workshop on Privacy, Security and Trust in KDD

Outline

1. Introduction
   - Motivation
   - Related Work
   - Problem Statement
2. KDE Resampling Algorithm
   - Resampling from Reconstructed PDF
   - Properties of Representative Samples
3. Numerical Results
   - Performance Metrics, Datasets, Classification Techniques
   - Results of Distributed Experiments
4. Conclusions
   - Summary
   - Further Work and Acknowledgments

Outline

1. Introduction
   - Motivation
   - Related Work
   - Problem Statement
2. KDE Resampling Algorithm
   - Resampling from Reconstructed PDF
   - Properties of Representative Samples
3. Numerical Results
   - Performance Metrics, Datasets, Classification Techniques
   - Results of Distributed Experiments
4. Conclusions
   - Summary
   - Further Work and Acknowledgments

Outline

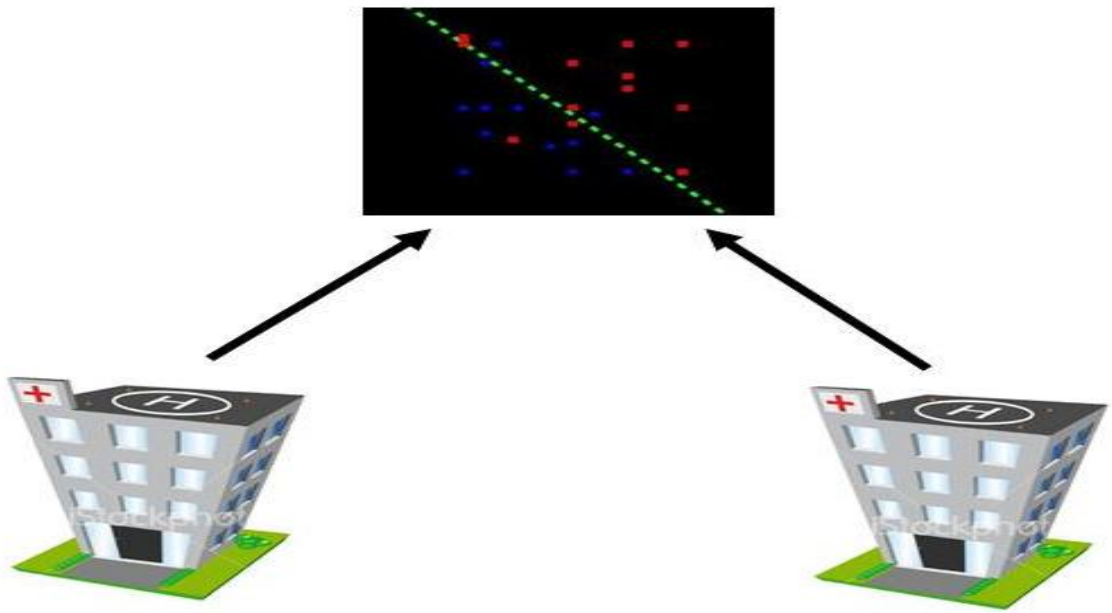
1. Introduction
   - Motivation
   - Related Work
   - Problem Statement
2. KDE Resampling Algorithm
   - Resampling from Reconstructed PDF
   - Properties of Representative Samples
3. Numerical Results
   - Performance Metrics, Datasets, Classification Techniques
   - Results of Distributed Experiments
4. Conclusions
   - Summary
   - Further Work and Acknowledgments

Outline

1. Introduction
   - Motivation
   - Related Work
   - Problem Statement
2. KDE Resampling Algorithm
   - Resampling from Reconstructed PDF
   - Properties of Representative Samples
3. Numerical Results
   - Performance Metrics, Datasets, Classification Techniques
   - Results of Distributed Experiments
4. Conclusions
   - Summary
   - Further Work and Acknowledgments

Motivation

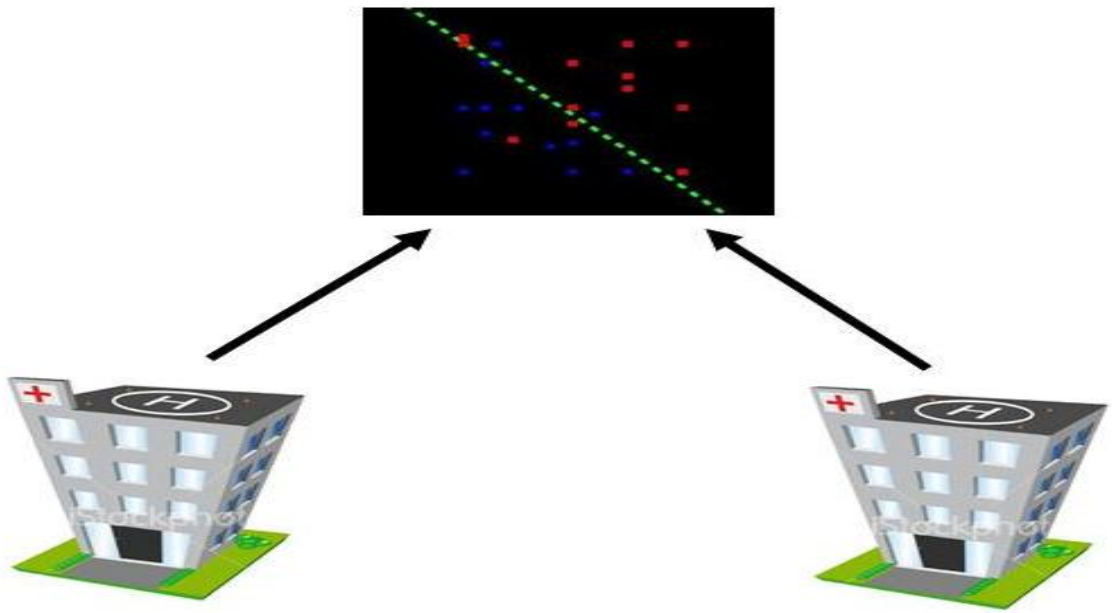
- Example: Two hospitals seek to construct a global classifier based on existing private patient data.



- Pooling data results in a more accurate global classifier.
- Instead of building local classifiers on limited data at each hospital site.

Motivation

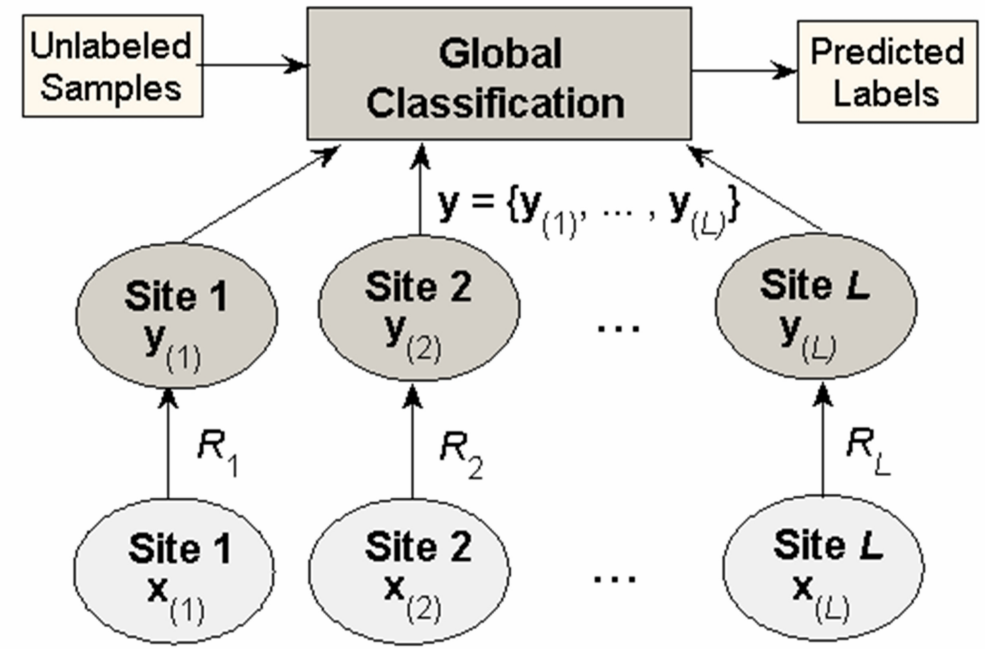
- Example: Two hospitals seek to construct a global classifier based on existing private patient data.



- Pooling data results in a more accurate global classifier.
- Instead of building local classifiers on limited data at each hospital site.

Motivation

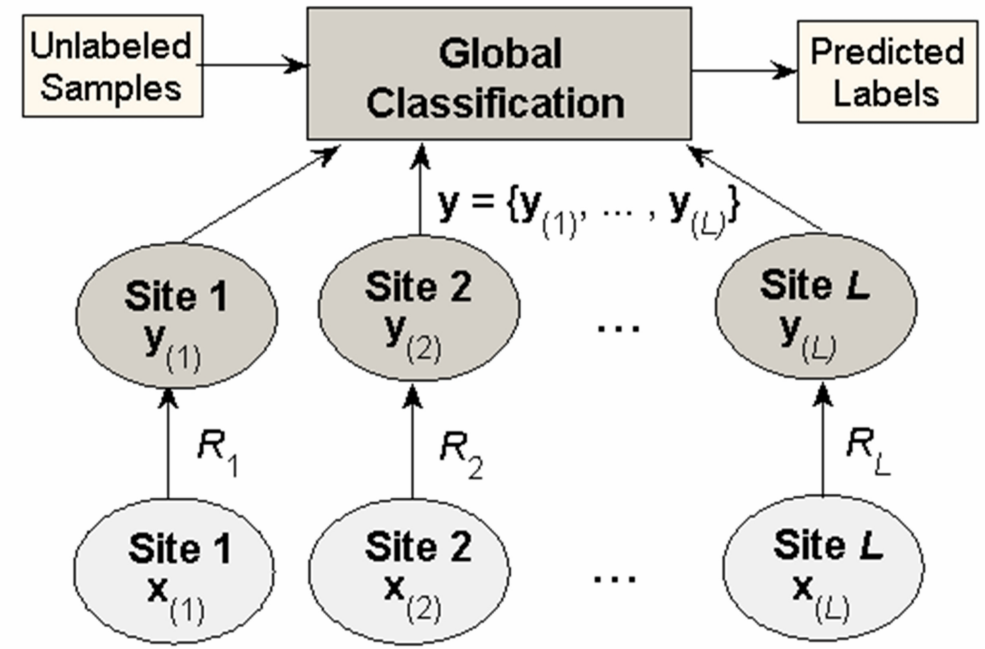
- We choose to randomize the data before transmitting the data to a centralized server.



- Can privacy + accuracy co-exist in a distributed scenario?

Motivation

- We choose to randomize the data before transmitting the data to a centralized server.



- Can privacy + accuracy co-exist in a distributed scenario?

Randomization in PPDM

- Randomization
  1. Randomize locally to give sanitized data for sharing.
  2. Similar to horizontally-partitioned scenario (Du et al. 2004).
  3. Pool data to form a larger training set to construct a global classifier.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.
- New randomization algorithm – Kernel Density Estimate (KDE) Resampling.

Randomization in PPDM

- Randomization
  1. Randomize locally to give sanitized data for sharing.
  2. Similar to horizontally-partitioned scenario (Du et al. 2004).
  3. Pool data to form a larger training set to construct a global classifier.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.
- New randomization algorithm – Kernel Density Estimate (KDE) Resampling.

Randomization in PPDM

- Randomization
  1. Randomize locally to give sanitized data for sharing.
  2. Similar to horizontally-partitioned scenario (Du et al. 2004).
  3. Pool data to form a larger training set to construct a global classifier.
- Challenges: Two conflicting concerns.
  1. Confidentiality of the private information
  2. Utility of the aggregate statistics.
- New randomization algorithm – Kernel Density Estimate (KDE) Resampling.

Randomization Techniques

- Atallah et al. (1999) first considered the data sanitization problem.
- Agrawal and Srikant (2000) suggested adding IID noise.
- Noise addition has since been shown to be insecure (Kargupta et al. 2003).
- Distance-Preserving approaches have been suggested (Liu et al. 2006, Chen et al. 2005, Olivera et al. 2007).
  - But we can derive bounds on the data.
- We suggest a Non-Distance-Preserving randomization approach, which also has classification accuracy.

Randomization Techniques

- Atallah et al. (1999) first considered the data sanitization problem.
- Agrawal and Srikant (2000) suggested adding IID noise.
- Noise addition has since been shown to be insecure (Kargupta et al. 2003).
- Distance-Preserving approaches have been suggested (Liu et al. 2006, Chen et al. 2005, Olivera et al. 2007).
  - But we can derive bounds on the data.
- We suggest a Non-Distance-Preserving randomization approach, which also has classification accuracy.

Randomization Techniques

- Atallah et al. (1999) first considered the data sanitization problem.
- Agrawal and Srikant (2000) suggested adding IID noise.
- Noise addition has since been shown to be insecure (Kargupta et al. 2003).
- Distance-Preserving approaches have been suggested (Liu et al. 2006, Chen et al. 2005, Olivera et al. 2007).
  - But we can derive bounds on the data.
- We suggest a Non-Distance-Preserving randomization approach, which also has classification accuracy.

Non-randomization Techniques

- k-anonymization used to generalize databases for preserving privacy (Sweeney, 1998).
- Secure Multi-Party Computation (SMC) techniques are more accurate but higher communication overhead.
- Distributed Clustering via optimization of information theoretic quantities (Merugu and Ghosh, 2005).

Problem Definition and Notation

- Private data is stored in d-dim row vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
- Associated with N targets (class labels) $t_1, \ldots t_N$.
- L distributed data sites (private) and 1 centralized (untrusted) server.
- $\mathbf{x}_{(l)} \in \mathbb{R}^{N_l \times d}$ contains the N_l data vectors at site l.
- Assume: Vectors in $\mathbf{x}_{(l)}$ are drawn from IID RVs with PDF $f_{\mathbf{x}_{(l)}}\left(\mathbf{x}_{(l)}\right)$.

Problem Definition and Notation

- Private data is stored in d-dim row vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
- Associated with N targets (class labels) $t_1, \ldots t_N$.
- L distributed data sites (private) and 1 centralized (untrusted) server.
- $\mathbf{x}_{(l)} \in \mathbb{R}^{N_l \times d}$ contains the N_l data vectors at site l.
- Assume: Vectors in $\mathbf{x}_{(l)}$ are drawn from IID RVs with PDF $f_{\mathbf{x}_{(l)}}\left(\mathbf{x}_{(l)}\right)$.

Problem Definition and Notation

- Private data is stored in d-dim row vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
- Associated with N targets (class labels) $t_1, \ldots t_N$.
- L distributed data sites (private) and 1 centralized (untrusted) server.
- $\mathbf{x}_{(l)} \in \mathbb{R}^{N_l \times d}$ contains the N_l data vectors at site l.
- Assume: Vectors in $\mathbf{x}_{(l)}$ are drawn from IID RVs with PDF $f_{\mathbf{x}_{(l)}}\left(\mathbf{x}_{(l)}\right)$.

Problem Definition and Notation

Task

Find a randomization scheme for site l, $R_l(\cdot)$ s.t.

$$\mathbf{y}_{(l)} = R_l(\mathbf{x}_{(l)}), \qquad 1 \leq l \leq L.$$

- $\mathbf{y} = \{\mathbf{y}_{(l)}\}_{l=1}^{L}$ is then sent to the centralized server for global classification.
- We will demonstrate empirically that for our choice of $R_l(\cdot)$,

$$P_{rand}(err) \approx P_{ori}(err).$$

- Why non-distance-preserving?

Problem Definition and Notation

Task

Find a randomization scheme for site l, $R_l(\cdot)$ s.t.

$$\mathbf{y}_{(l)} = R_l(\mathbf{x}_{(l)}), \qquad 1 \leq l \leq L.$$

- $\mathbf{y} = \{\mathbf{y}_{(l)}\}_{l=1}^{L}$ is then sent to the centralized server for global classification.
- We will demonstrate empirically that for our choice of $R_l(\cdot)$,

$$P_{rand}(err) \approx P_{ori}(err).$$

- Why non-distance-preserving?

Problem Definition and Notation

Task

Find a randomization scheme for site l, $R_l(\cdot)$ s.t.

$$\mathbf{y}_{(l)} = R_l(\mathbf{x}_{(l)}), \qquad 1 \leq l \leq L.$$

- $\mathbf{y} = \{\mathbf{y}_{(l)}\}_{l=1}^{L}$ is then sent to the centralized server for global classification.
- We will demonstrate empirically that for our choice of $R_l(\cdot)$,

$$P_{rand}(err) \approx P_{ori}(err).$$

- Why non-distance-preserving?

Why Non-Distance-Preserving Randomization?

If Distance-Preserving Randomization (Random Projection-Based Multiplication) was used:

- The randomized data can be vulnerable to disclosure (Caetano 2004)
  - True iff $(d+1)$ points do not lie in a $(d-1)$-dimensional vector subspace.
- Furthermore in a $L = 2$ site scenario (details in our paper):
  - One can lower bound the norm of all the columns of the data matrix $\mathbf{x}_{(1)}$ given an estimate of the norm of one of the columns of $\mathbf{x}_{(2)}$.

Why Non-Distance-Preserving Randomization?

If Distance-Preserving Randomization (Random Projection-Based Multiplication) was used:

- The randomized data can be vulnerable to disclosure (Caetano 2004)
  - True iff $(d+1)$ points do not lie in a $(d-1)$-dimensional vector subspace.
- Furthermore in a $L = 2$ site scenario (details in our paper):
  - One can lower bound the norm of all the columns of the data matrix $\mathbf{x}_{(1)}$ given an estimate of the norm of one of the columns of $\mathbf{x}_{(2)}$.

Why Non-Distance-Preserving Randomization?

If Distance-Preserving Randomization (Random Projection-Based Multiplication) was used:

- The randomized data can be vulnerable to disclosure (Caetano 2004)
  - True iff $(d+1)$ points do not lie in a $(d-1)$-dimensional vector subspace.
- Furthermore in a $L = 2$ site scenario (details in our paper):
  - One can lower bound the norm of all the columns of the data matrix $\mathbf{x}_{(1)}$ given an estimate of the norm of one of the columns of $\mathbf{x}_{(2)}$.

Why Non-Distance-Preserving Randomization?

Hence, if Distance-Preserving Randomization was used:

- Potential security breach.
- Intuition: Along with the preservation of distances, the order of the samples is also preserved.
- Hence, we propose a non-distance-preserving approach to circumvent the problem.

Why Non-Distance-Preserving Randomization?

Hence, if Distance-Preserving Randomization was used:

- Potential security breach.
- Intuition: Along with the preservation of distances, the order of the samples is also preserved.
- Hence, we propose a non-distance-preserving approach to circumvent the problem.

Why Non-Distance-Preserving Randomization?

Hence, if Distance-Preserving Randomization was used:

- Potential security breach.
- Intuition: Along with the preservation of distances, the order of the samples is also preserved.
- Hence, we propose a non-distance-preserving approach to circumvent the problem.

Kernel Density Estimation

- Main Idea: Sample from the reconstructed Kernel Density Estimate (Parzen, 1962) (Devroye, 1985).
- The KDE is a non-parameteric estimate of the PDF.
- Nice properties of the new, representative samples, including asymptotic independence and consistency.

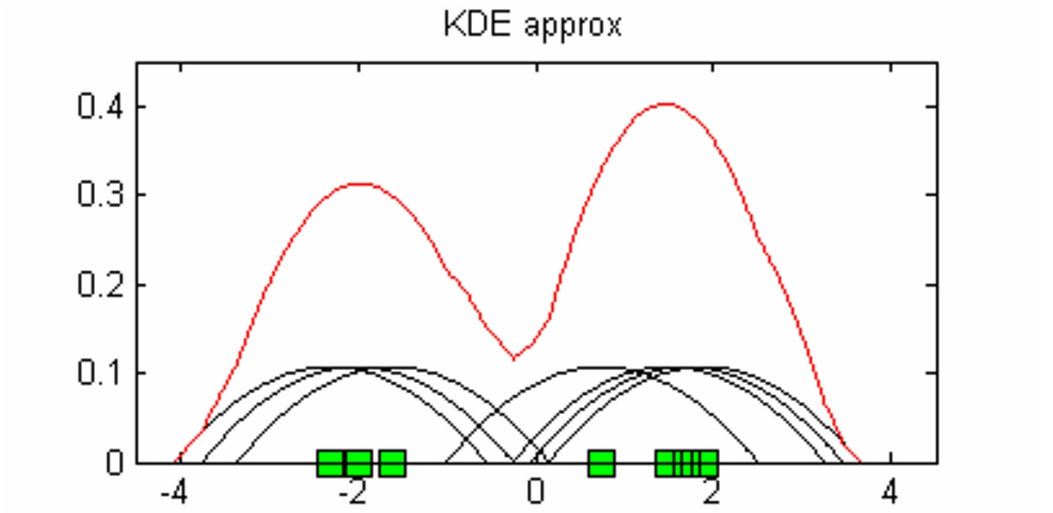
Kernel Density Estimation

- Main Idea: Sample from the reconstructed Kernel Density Estimate (Parzen, 1962) (Devroye, 1985).
- The KDE is a non-parameteric estimate of the PDF.
- Nice properties of the new, representative samples, including asymptotic independence and consistency.

Kernel Density Estimation

- For each data site l, we will construct and sample from an estimate of the PDF using $\mathbf{x}_{(l)}$.

$$\hat{f}_{\mathbf{x}_{(l)}}\left(\mathbf{x}_{(l)}; \mathbf{x}_{(l,1)}, \ldots, \mathbf{x}_{(l,N_l)}\right) = \frac{1}{N_l} \sum_{j=1}^{N_l} K\left(\mathbf{x}_{(l)} - \mathbf{x}_{(l,j)}; \mathbf{h}_l\right),$$



Resampling

- The random vector

$$\mathbf{x}_{(l)} = \frac{1}{N_l} \sum_{j=1}^{N_l} \mathbf{x}_{(l,j)}$$

  is a mixture density.

- KDE does not have to be explicitly constructed for sampling.

Resampling

- The random vector

$$\mathbf{x}_{(l)} = \frac{1}{N_l} \sum_{j=1}^{N_l} \mathbf{x}_{(l,j)}$$

  is a mixture density.

- KDE does not have to be explicitly constructed for sampling.

KDE Resampling Algorithm

for $l \leftarrow 1$ *to* L **do**
  for $i \leftarrow 1$ *to* d **do**
    $\hat{\sigma}_{l,i}$ = Standard deviation in dimension i;
    $h_{l,i}$ = Bandwidth in dimension i
  endFor
  for $j \leftarrow 1$ *to* M_l **do**
    r = Random integer from 1 to N_l inclusive;
    $\mathbf{y}_{(l,j)}$ = Random sample vector from r^{th} kernel;
  endFor
endFor

Asymptotic Independence

- Randomized samples are independent of the Original samples as the number of samples $N_l \to \infty$.
- Probabilistic inference cannot be performed based on the randomized samples $\mathbf{y}_{(l)}$ if N_l is sufficiently large.
- Dependent on how we select $\mathbf{h}_l$ (Scott's rule).

$$h_{l,i} = \left(\frac{4}{d+2}\right)^{1/(d+4)} N_l^{-1/(d+4)} \hat{\sigma}_{l,i}.$$

Asymptotic Independence

- Randomized samples are independent of the Original samples as the number of samples $N_l \to \infty$.
- Probabilistic inference cannot be performed based on the randomized samples $\mathbf{y}_{(l)}$ if N_l is sufficiently large.
- Dependent on how we select $\mathbf{h}_l$ (Scott's rule).

$$h_{l,i} = \left(\frac{4}{d+2}\right)^{1/(d+4)} N_l^{-1/(d+4)} \hat{\sigma}_{l,i}.$$

Asymptotic Independence

- Randomized samples are independent of the Original samples as the number of samples $N_l \to \infty$.
- Probabilistic inference cannot be performed based on the randomized samples $\mathbf{y}_{(l)}$ if N_l is sufficiently large.
- Dependent on how we select $\mathbf{h}_l$ (Scott's rule).

$$h_{l,i} = \left(\frac{4}{d+2}\right)^{1/(d+4)} N_l^{-1/(d+4)} \hat{\sigma}_{l,i}.$$

Consistency and Tractability

- Under mild assumptions (See our paper), KDE $\hat{f}_l$, is consistent i.e.

$$\lim_{N_l \to \infty} \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right] = 0, \quad 1 \leq l \leq L.$$

- As N_l becomes large, the KDE $\hat{f}_l(\cdot)$ becomes increasingly accurate.
- Treat $\{\mathbf{y}_{(l)}\}_{l=1}^{L}$ as the training data without compromising accuracy.
- KDE approximation algorithm is tractable.

Consistency and Tractability

- Under mild assumptions (See our paper), KDE $\hat{f}_l$, is consistent i.e.

$$\lim_{N_l \to \infty} \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right] = 0, \quad 1 \leq l \leq L.$$

- As N_l becomes large, the KDE $\hat{f}_l(\cdot)$ becomes increasingly accurate.
- Treat $\{\mathbf{y}_{(l)}\}_{l=1}^{L}$ as the training data without compromising accuracy.
- KDE approximation algorithm is tractable.

Consistency and Tractability

- Under mild assumptions (See our paper), KDE $\hat{f}_l$, is consistent i.e.

$$\lim_{N_l \to \infty} \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right] = 0, \quad 1 \leq l \leq L.$$

- As N_l becomes large, the KDE $\hat{f}_l(\cdot)$ becomes increasingly accurate.
- Treat $\{\mathbf{y}_{(l)}\}_{l=1}^{L}$ as the training data without compromising accuracy.
- KDE approximation algorithm is tractable.

Consistency and Tractability

- Under mild assumptions (See our paper), KDE $\hat{f}_l$, is consistent i.e.

$$\lim_{N_l \to \infty} \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right] = 0, \quad 1 \leq l \leq L.$$

- As N_l becomes large, the KDE $\hat{f}_l(\cdot)$ becomes increasingly accurate.
- Treat $\{\mathbf{y}_{(l)}\}_{l=1}^{L}$ as the training data without compromising accuracy.
- KDE approximation algorithm is tractable.

Distributed Aggregate Privacy Loss $\mathcal{DAPL}$

- We define privacy loss is a function of the degree of independence between new, randomized samples and original samples.
- A sufficient condition for asymptotic independence is:
- The expected l_1 distance between $\hat{f}_l$ and f_l tends to zero with $N_l \rightarrow \infty$ (Devroye, 1985).

Definition

The *Distributed Aggregate Privacy Loss* $\mathcal{DAPL}$ is defined as:

$$\mathcal{DAPL} \triangleq \frac{1}{2}\left(\sum_{l=1}^{L} c_l \, \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right]\right), \quad c_l \triangleq N_l/N.$$

Distributed Aggregate Privacy Loss $\mathcal{DAPL}$

- We define privacy loss is a function of the degree of independence between new, randomized samples and original samples.
- A sufficient condition for asymptotic independence is:
- The expected l_1 distance between $\hat{f}_l$ and f_l tends to zero with $N_l \rightarrow \infty$ (Devroye, 1985).

Definition

The *Distributed Aggregate Privacy Loss* $\mathcal{DAPL}$ is defined as:

$$\mathcal{DAPL} \triangleq \frac{1}{2}\left(\sum_{l=1}^{L} c_l \, \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right]\right), \quad c_l \triangleq N_l/N.$$

Distributed Aggregate Privacy Loss $\mathcal{DAPL}$

- We define privacy loss is a function of the degree of independence between new, randomized samples and original samples.
- A sufficient condition for asymptotic independence is:
- The expected l_1 distance between $\hat{f}_l$ and f_l tends to zero with $N_l \rightarrow \infty$ (Devroye, 1985).

Definition

The *Distributed Aggregate Privacy Loss* $\mathcal{DAPL}$ is defined as:

$$\mathcal{DAPL} \stackrel{\triangle}{=} \frac{1}{2}\left(\sum_{l=1}^{L} c_l \, \mathbb{E}\left[\int \left|\hat{f}_l - f_l\right|\right]\right), \quad c_l \stackrel{\triangle}{=} N_l/N.$$

Deterioration of Classification ϕ

- We define the classification error as:

$$P(err) \stackrel{\triangle}{=} 1 - \sum_{i=1}^{|\mathcal{C}|} \int_{\Omega_i} p\left(\boldsymbol{\xi}|\mathcal{C}_i\right) P(\mathcal{C}_i)\, d\boldsymbol{\xi}.$$

Definition

The *Deterioration of Classification* ϕ is defined as:

$$\phi \stackrel{\triangle}{=} P_{rand}(err) - P_{ori}(err),$$

- $P_{ori}(err)$: Error with original samples for training.
- $P_{rand}(err)$: Error with randomized samples for training.

Deterioration of Classification ϕ

- We define the classification error as:

$$P(err) \stackrel{\triangle}{=} 1 - \sum_{i=1}^{|\mathcal{C}|} \int_{\Omega_i} p\left(\boldsymbol{\xi}|\mathcal{C}_i\right) P(\mathcal{C}_i)\, d\boldsymbol{\xi}.$$

Definition

The *Deterioration of Classification* ϕ is defined as:

$$\phi \stackrel{\triangle}{=} P_{rand}(err) - P_{ori}(err),$$

- $P_{ori}(err)$: Error with original samples for training.
- $P_{rand}(err)$: Error with randomized samples for training.

Datasets

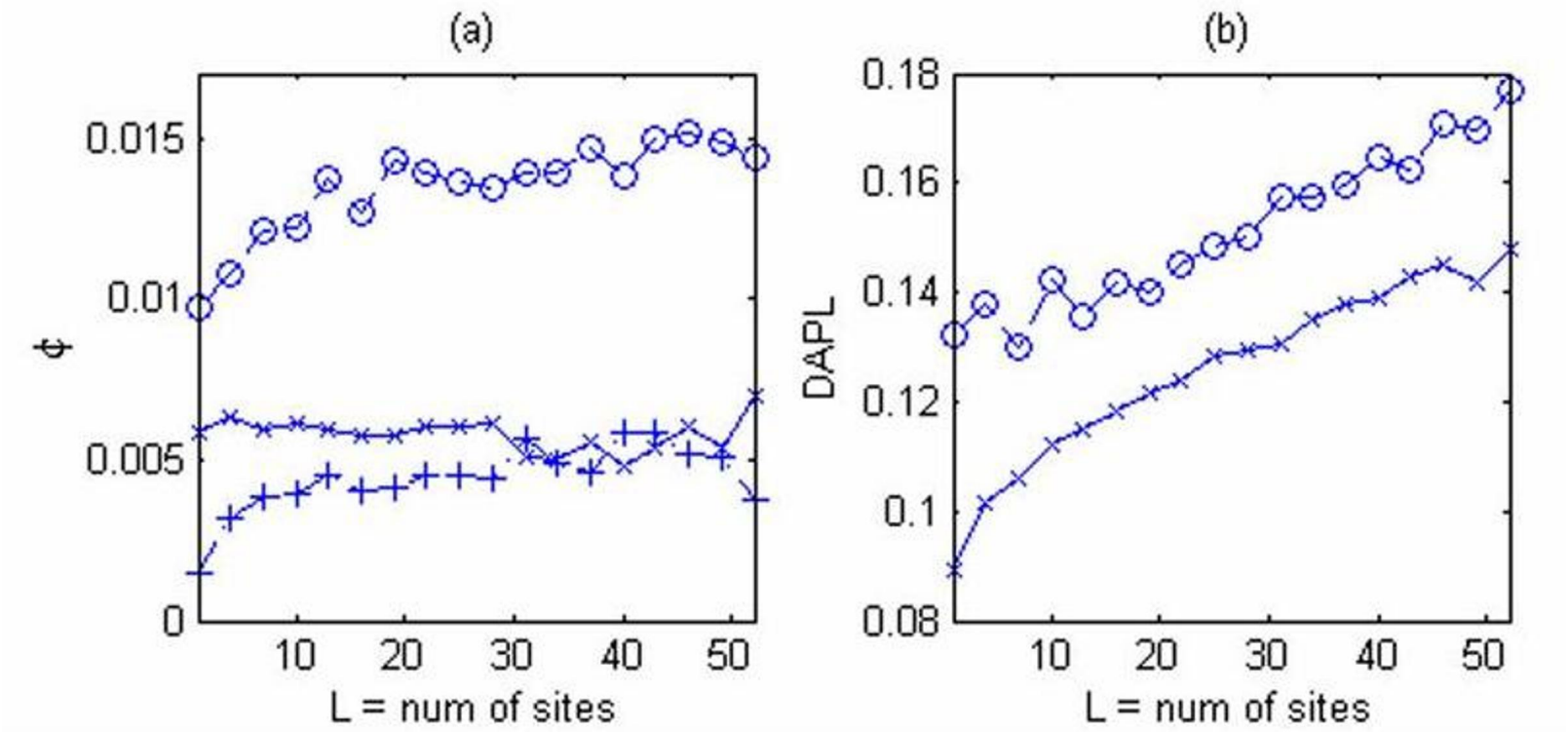
Dataset	#Class	#Dim(d)	#Trg(N)	#Test
Iris	2	4	120	30
SVMGuide1	2	4	3089	4000
Diabetes	2	8	576	192
Breast-Cancer	2	10	512	171
Ionosphere	2	34	263	88

Table: Our five datasets from LIBSVM and the UCI ML Repository.

Classification Techniques

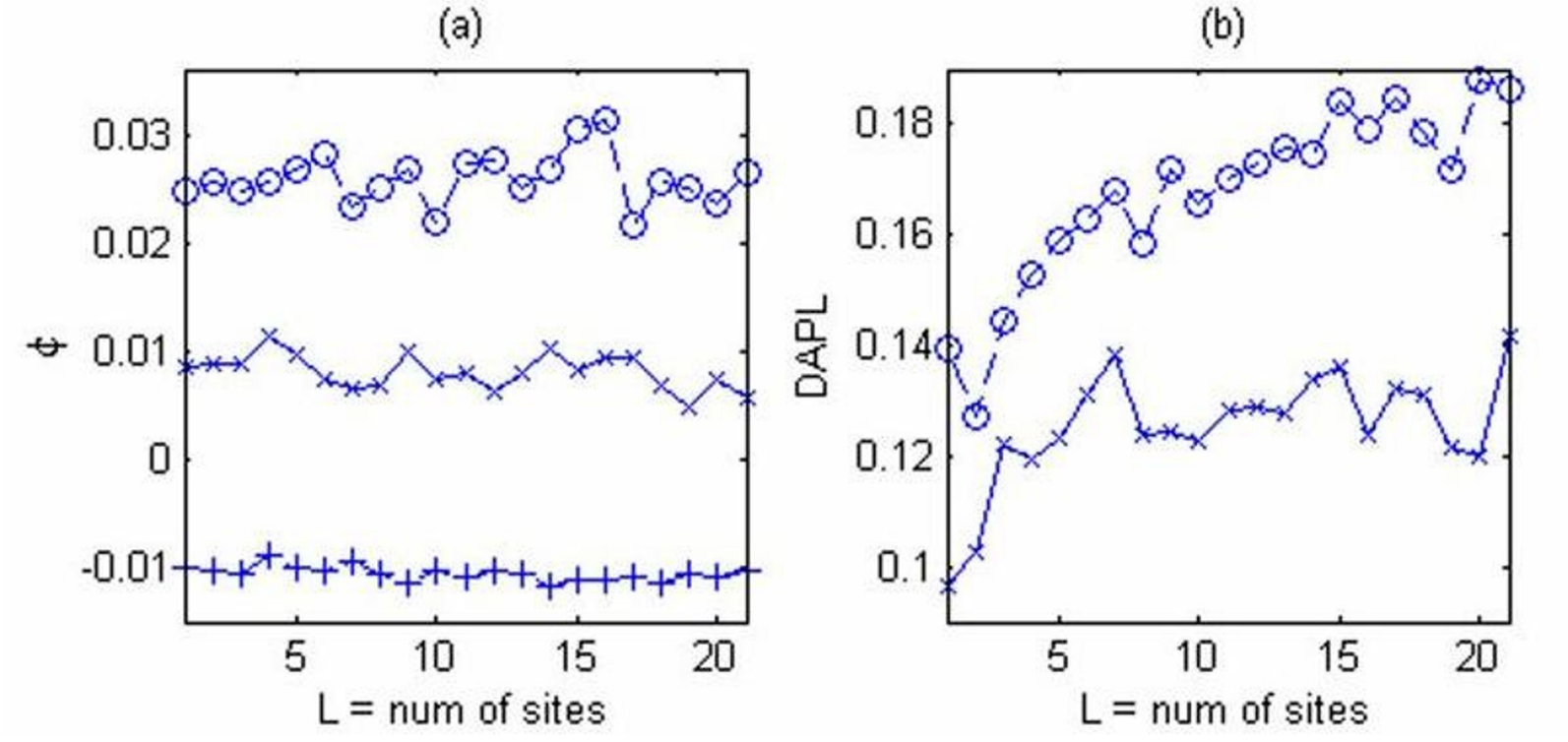
1. Artificial Neural Networks (ANN) by trained by error backpropagation.
2. k-Nearest Neighbors classifier (kNN).
3. Naïve Bayes classifier (NB) with each attribute assumed to be Gaussian.

SVMGuide1 Dataset



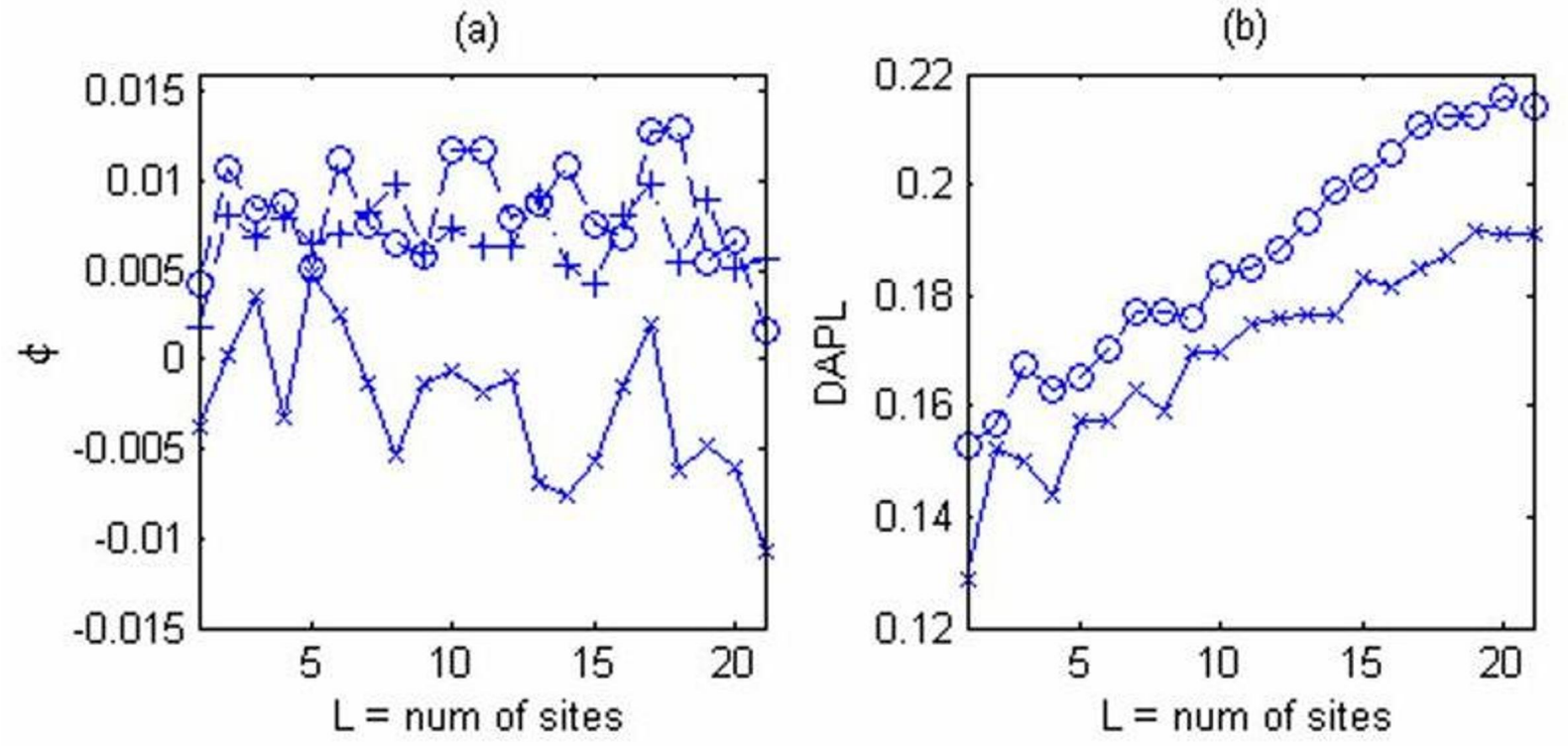
(a) (Key: x - ANN, o - *k*NN, + - NB); (b) DAPL (Key: x - C1, o - C2)

Breast-Cancer Dataset



(a) (Key: x - ANN, o - *k*NN, + - NB); (b) DAPL (Key: x - C1, o - C2)

Diabetes Dataset



(a) (Key: x - ANN, o - *k*NN, + - NB); (b) DAPL (Key: x - C1, o - C2)

Observations

- Accuracy:
  - No correlation between L and the Deterioration of Classification ϕ.
  - Randomized data is still amenable to data mining tasks regardless of L.
- Privacy:
  - Increasing trend for the $\mathcal{DAPL}$ because $N_l \downarrow$ with $L \uparrow$.
  - A compromise between L and $\mathcal{DAPL}$.
- Improving the sampling algorithm by selecting optimal $\mathbf{h}_l$.

Observations

- Accuracy:
  - No correlation between L and the Deterioration of Classification ϕ.
  - Randomized data is still amenable to data mining tasks regardless of L.
- Privacy:
  - Increasing trend for the $\mathcal{DAPL}$ because $N_l \downarrow$ with $L \uparrow$.
  - A compromise between L and $\mathcal{DAPL}$.
- Improving the sampling algorithm by selecting optimal $\mathbf{h}_l$.

Observations

- Accuracy:
  - No correlation between L and the Deterioration of Classification ϕ.
  - Randomized data is still amenable to data mining tasks regardless of L.
- Privacy:
  - Increasing trend for the $\mathcal{DAPL}$ because $N_l \downarrow$ with $L \uparrow$.
  - A compromise between L and $\mathcal{DAPL}$.
- Improving the sampling algorithm by selecting optimal $\mathbf{h}_l$.

Summary

- Proposed an algorithm (KDE Resampling) for data sanitization to share private data for distributed classification.
  - Asymptotically independent (Privacy).
  - Consistent (Accuracy).
- max $|P_{rand}(err) - P_{ori}(err)| < 3\%$ for all the datasets.
- A malicious intruder cannot establish bounds on the private data using KDE Resampling.
- Only one-way communication – tractable.

Summary

- Proposed an algorithm (KDE Resampling) for data sanitization to share private data for distributed classification.
  - Asymptotically independent (Privacy).
  - Consistent (Accuracy).
- max $|P_{rand}(err) - P_{ori}(err)| < 3\%$ for all the datasets.
- A malicious intruder cannot establish bounds on the private data using KDE Resampling.
- Only one-way communication – tractable.

Summary

- Proposed an algorithm (KDE Resampling) for data sanitization to share private data for distributed classification.
  - Asymptotically independent (Privacy).
  - Consistent (Accuracy).
- max $|P_{rand}(err) - P_{ori}(err)| < 3\%$ for all the datasets.
- A malicious intruder cannot establish bounds on the private data using KDE Resampling.
- Only one-way communication – tractable.

Summary

- Proposed an algorithm (KDE Resampling) for data sanitization to share private data for distributed classification.
  - Asymptotically independent (Privacy).
  - Consistent (Accuracy).
- max $|P_{rand}(err) - P_{ori}(err)| < 3\%$ for all the datasets.
- A malicious intruder cannot establish bounds on the private data using KDE Resampling.
- Only one-way communication – tractable.

Further Work

- Re-examine the issue of the privacy metric.
- $\mathcal{DAPL}$ only quantifies the l_1 distances between the 2 distributions.
- Assumption that the data records is IID may not be realistic.
  - But our algorithm requires this assumption.
  - This shortcoming will be addressed in our future work.

Further Work

- Re-examine the issue of the privacy metric.
- $\mathcal{DAPL}$ only quantifies the l_1 distances between the 2 distributions.
- Assumption that the data records is IID may not be realistic.
  - But our algorithm requires this assumption.
  - This shortcoming will be addressed in our future work.

Acknowledgments

- I would like to thank the support of the Agency for Science, Technology and Research (A*STAR), Singapore.

- Homepage: `http://web.mit.edu/vtan/www`.