CERIAS

The Center for Education and Research in Information Assurance and Security

Data Classification Using Anatomized Training Data

The Problem

Anatomized Learning Problem

Given some l-diverse data in the anatomy model, can we learn accurate data mining models?

- What is the anatomy model?
- What is l-diverse?
- Under which assumptions?

Anatomy Model

Separate data table (D) into two tables, identifier (IT) and sensitive table (ST) instead of generalizing records in the same group:

- Divide D in m groups G_j, group id (GID) j
- IT attributes (A_{id}): $A_1, \cdots, A_d$
- ST attribute: A_s
- Publish IT and ST instead of D
- L-diverse
- Xiao et al. (2006), Nergiz et al. (2011, 2013)
- Patient Data Example (HIPAA 2002)

Patient (P)	Age (A)	Address (AD)	GID (G)	SEQ (S)
Ike	41	Dayton	1	1
Eric	22	Richmond	1	2
Olga	30	Lafayette	2	3
Kelly	35	Lafayette	2	4
Faye	24	Richmond	3	5
Mike	47	Richmond	3	6
Jason	45	Lafayette	4	7
Max	31	Lafayette	4	8

Identifier table (IT)

H(SEQ)	GID (G)	Disease (D)
$H_{k2}(1)$	1	Cold
$H_{k2}(2)$	1	Fever
$H_{k2}(3)$	2	Flu
$H_{k2}(4)$	2	Cough
$H_{k2}(5)$	3	Flu
$H_{k2}(6)$	3	Fever
$H_{k2}(7)$	4	Cough
$H_{k2}(8)$	4	Flu

Sensitive table (ST)

L-diverse: Privacy Standard

Every instance in IT can be associated with L different instances in ST

- Patient Data Example: L=2
- $\forall\, G_j, v \in \pi_{A_s}(G_j), \frac{freq(v,G_j)}{|G_j|} \leq \frac{1}{l}$
- Machanavajjhala et al. (2007)

Approaches

Learning Problem Assumptions

- Training set of n_{tr} instances in anatomy model and test set of n_{te} instances without any anonymization (Inan et al. (2009))
- No background knowledge for IT
- **Can't predict the sensitive attribute A_s. If we could, we would be violating privacy!**
- Prediction task of $A_i$or C (binary!):
  - **Type 1:** $A_i \in \{A_1, \cdots, A_d\}$
  - **Type 2:** $C \notin \{A_1, \cdots, A_d\} \wedge C \neq A_s$
- No IT and ST linking
- Data must remain L-diverse
- No involvement of D's publisher
- Relaxed Assumptions (some models):
  - Minimal involvement of D's publisher, limited sources of D's publisher
  - Link IT and ST on small subsets
  - "Distributed data mining" between third party (server) and data publisher (client)

Collaborative Decision Tree Analysis:

- Type 1 prediction task with relaxed assumptions
- **Advantages:**
  1. Preserves privacy with reasonable accuracy
  2. Big part of the decision tree is learnt by the third party, a desired situation in cloud server/client architecture
- **Limitations:**
  1. Hard to give any bound on the model performance, in particular on the conditional risk (error rate) of the classification.
  2. What about the execution time guarantees in a cloud client/server architecture?
- ***Need of a more justified model with the conditional risk guarantees!***

Nearest Neighbor Rule in Anatomy Model

- Type 2 prediction task without relaxed assumptions.
- **Anatomized Training Data (D_A):** $IT \underset{IT.GID\,=\,ST.GID}{\bowtie} ST$
- **Augmentation of nearest neighbor rule (Cover and Hart 1967):** Expand the training set such that the expanded version has size $n_{tr}l$
  - For all fixed l, the conditional risk is the corollary of Cover and Hart when $n_{tr} \to \infty$
- **One Critical Question:** "*How does the Bayes Risk change?*"

Empirical Results

Collaborative Decision Tree Learning

1. Distributed Data Mining in the cloud (Client/server architecture)
2. On-the fly encrypted subtrees (Mancuhan and al. 2014)
3. Experiments with four datasets from the UCI collection: adult, vote, autos and Australian credit
4. 10 fold cross validation on each dataset measuring accuracy

Third Party (Server)

A	AD	G	S
41	Dayton	1	1
22	Richmond	1	2
30	Lafayette	2	3
35	Lafayette	2	4
24	Richmond	3	5
47	Richmond	3	6
45	Lafayette	4	7
31	Lafayette	4	8

AD=Dayton

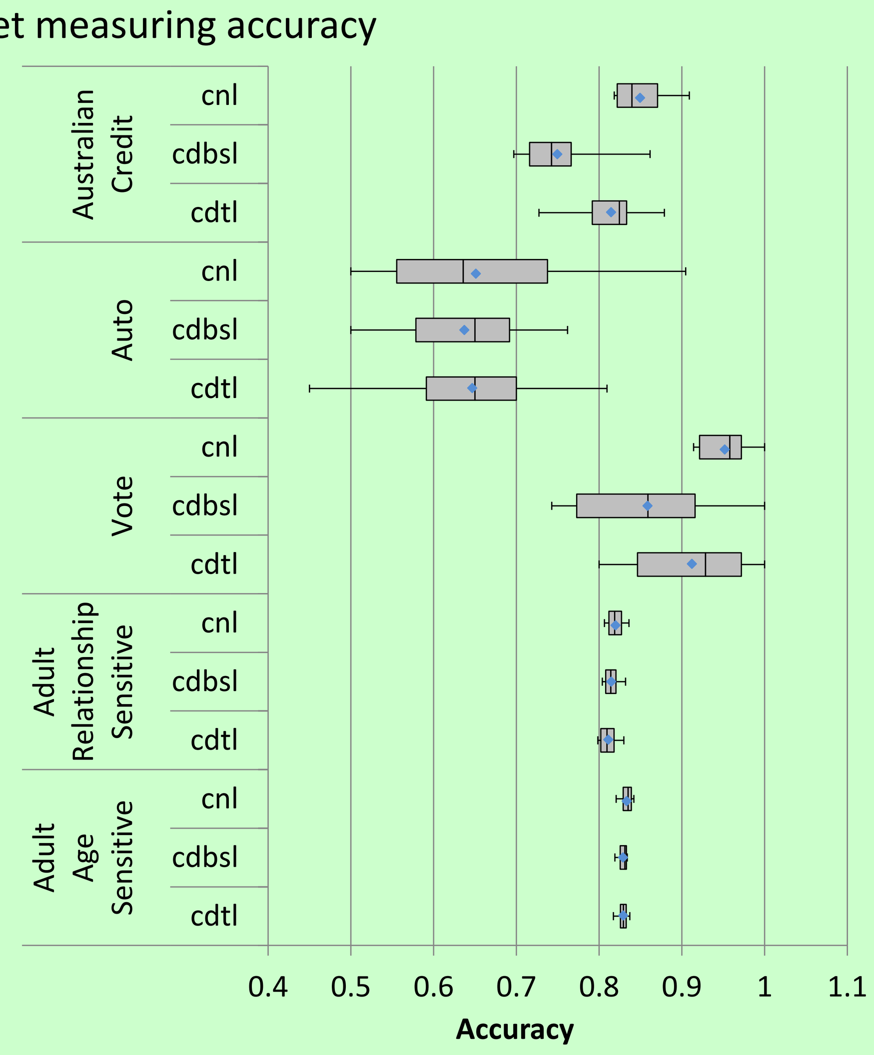
A	AD	G	S
41	Dayton	1	1

AD=Lafayette

A	AD	G	S
30	Lafayette	2	3
35	Lafayette	2	4
45	Lafayette	4	7
31	Lafayette	4	8

AD=Richmond

A	AD	G	S
22	Richmond	1	2
24	Richmond	3	5
47	Richmond	3	6



D's publisher (Client)

A	AD	D
30	Lafayette	Flu
35	Lafayette	Cough
45	Lafayette	Cough
31	Lafayette	Flu

D=Cough

A	AD	D
35	Lafayette	Cough
45	Lafayette	Cough

Label: >=35

D=Flu

A	AD	D
30	Lafayette	Flu
31	Lafayette	Flu

Label: <35

Encrypted Subtree!

H(SEQ)	G	D
$H_{k2}(3)$	2	Flu
$H_{k2}(4)$	2	Cough
$H_{k2}(7)$	4	Cough
$H_{k2}(8)$	4	Flu

Theoretical Results

<u>*Theorem:*</u> Let $M \in \mathbb{R}^{d+1}$ be a metric space, D be the training data and D_A be the anatomized training data. Let $P_{A_1}(X)$ and $P_{A_2}(X)$ be the smooth probability density functions of X. Let $P_{A_1}(X)$ and $P_{A_2}(X)$ be the class priors such that $P_A(X) = P_{A_1}P_{A_1}(X) + P_{A_2}P_{A_2}(X)$. Similarly, let $P_1(X)$ and $P_2(X)$ be the smooth probability density functions of X such that $P(X) = P_1P_1(X) + P_2P_2(X)$ with class priors P_1 and P_2. Let $h_A(X) = -\ln\left(P_{A_1}(X)/P_{A_2}(X)\right)$ and $h(X) = -\ln\left(P_1(X)/P_2(X)\right)$ be the classifiers with biases $\Delta h_A(X)$ and $\Delta h(X)$ respectively. Let $t = \ln(P_1/P_2)$ be the decision threshold with threshold bias Δt. Let $\epsilon_A > 0$ be the small changes on $P_1(X)$ and $P_2(X)$ resulting in $P_{A_1}(X)$ and $P_{A_2}(X)$; and R_A^*, R^* be the Bayesian error estimations with respective biases ΔR_A^*, ΔR^*. Let $\widehat{P_{A_i}}(X)$and $\widehat{P_i}(X)$ be the Parzen density estimations; and $K(*)$ be the kernel function for D with shape matrix A and size/volume parameter r. Last, let's assume that 1) A_{id} and A_s are independent in the training data D and the anatomized training data D_A 2) $R_A^* = R^*$ hold 3) $\Delta t < 1$. Therefore, the estimated Bayes risk is:

$$\widehat{R_A^*} \cong a_1r^2 + a_2r^4 + a_3\frac{r^{-d-1}}{N} + \epsilon_A a_4 r^2 + \epsilon_A a_5 r^4 - \epsilon_A a_6 \frac{r^{-d-1}}{N}$$

where $\epsilon_A a_6 \frac{r^{-d-1}}{N} > 0$ always holds.

- **Another critical question:** "*How does the convergence rate to the asymptotical conditional risk change?*"
  - $O(1/([Nl]^{d+1})$ versus $O(1/[N]^{d+1})$
  - Faster convergence to the asymptotical conditional risk using anatomized training data.
- How is the asymptotical conditional risk?
  - Depends on the Bayes risk (Theorem above)

This work supported by NPRP grant 09-256-1-046 from the Qatar National Research Foundation, 1415 PRF Research Grant from Purdue University

Current Work

- Experimentation of the nearest neighbor classifier using real data
- SVM classification generalization: How to adjust the right margin for the good generalization property when the training data is anatomized?
- Real-world case study: How this could inform data retention policies

Ongoing work supported by Northrup Grumman Cybersecurity Research Consortium

Koray Mancuhan kmancuha@purdue.edu

Chris Clifton clifton@cs.purdue.edu

PURDUE UNIVERSITY