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 Gj, group id (GID) j
- IT attributes (Ait): Ait,\.\.\.,Ait
- ST attribute: A
- Publish IT and ST instead of D
- L-diverse
- Xiao et al. (2006), Nengiz et al. (2011, 2013)
- Patient Data Example (HIPAA 2002)

L-diverse: Privacy Standard

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

- Patient Data Example: L=2
- \( V_{G_i, G_j} \subseteq N_{G_i} \leq \frac{1}{3} \)
- Machanavajhala et al. (2007)

Approaches

Learning Problem Assumptions

- Training set of \( n_r \) instances in anatomy model and test set of \( n_t \) instances without any anonymization (Inan et al. (2009))
- No background knowledge for IT
- Can’t predict the sensitive attribute A. If we could, we would be violating privacy!
- Prediction task of A | P or C (binary):
  - Type 1: \( A_i \in \{A_{i1}, \ldots, A_{in} \} \)
  - Type 2: \( C \in \{A_{i1}, \ldots, A_{in} \} \land C \neq A_{i1} \)
- 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 2 prediction task with relaxed assumptions
- Anatomized Training Data (D): IT, ST, GID = ST, GID
- Augmentation of nearest neighbor rule (Cover and Hart 1967): Expand the training set such that the expanded version has size \( n_r \cdot \frac{3}{1} \)
  - For all fixed i, the conditional risk is the corollary of Cover and Hart when \( n_r \to \infty \)
- One Critical Question: “How does the Bayes Risk change?”

Nearest Neighbor Rule in Anatomy Model

- Type 2 prediction task without relaxed assumptions
- Anatomized Training Data (D): IT, ST, GID = ST, GID
- Augmentation of nearest neighbor rule (Cover and Hart 1967): Expand the training set such that the expanded version has size \( n_r \cdot \frac{3}{1} \)
  - For all fixed i, the conditional risk is the corollary of Cover and Hart when \( n_r \to \infty \)
- 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

Theoretical Results

Theorem: Let \( M \in \mathbb{R}^{d+1} \) be a metric space, \( D \) be the training data and \( D \) be the anatomized training data. Let \( P_X \) and \( P_X \) be the smooth probability density functions of \( X \). Let \( P_X \) and \( P_X \) be the class priors such that \( P_X = P_X \cdot P_X \cdot P_X \). Similarly, let \( P_X \) and \( P_X \) be the smooth probability density functions of \( X \) such that \( P(x) = P_X(x) + P_X(x) \) with class priors \( P \) and \( P \). Let \( h_X = \ln(\frac{P_X}{P_X}) \) and \( h_X = \ln(P_X) \) be the classifiers with biases \( \Delta h_X \) and \( \Delta h_X \) respectively. Let \( t = \ln(P_X/P_X) \) be the decision threshold with threshold bias \( \Delta t \). Let \( \epsilon > 0 \) be the small changes on \( P_X \) and \( P_X \) resulting in \( P_X \) and \( P_X \) and \( P_X \) be the Bayesian error estimations with respective biases \( \Delta h_X \) and \( \Delta h_X \). Let \( \text{KL}(P_X \parallel P_X) \) be the Parzen density estimations; and \( K(\ast) \) be the kernel function for \( D \) with shape matrix \( A \) and size/volume parameter \( r \). Last, let’s assume that 1) \( A \) and \( A \) are independent in the training data \( D \) and the anatomized training data \( \frac{\text{KL}(P_X \parallel P_X)}{r^4} = \frac{\text{KL}(P_X \parallel P_X)}{r^4} = \frac{\text{KL}(P_X \parallel P_X)}{r^4} \). In this case, the estimated Bayes risk is:

- Another critical question: “How does the convergence rate to the asymptotic conditional risk change?”
- \( 0(1/(N^{d+1})) \) versus \( 0(1/(N^{d+1})) \)
- Faster convergence to the asymptotic conditional risk using anatomized training data.
- How is the asymptotic conditional risk?
- Depends on the Bayes risk (Theorem above)

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 anonymized?
- 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