CERIAS

The Center for Education and Research in Information Assurance and Security

Statistical Learning Theory Approach for Data Classification with ℓ-diversity

Koray Mancuhan kmancuha@purdue.edu
Chris Clifton clifton@cs.purdue.edu

Research Problem

How well can we train a classifier on private data?

Assumptions

1. Training set must be anonymized
2. Learned model may be released for use

Real life problem for cloud services storing large amount of data.

Objective

- Show theoretically and practically that privacy and analytical capabilities are possible at the same time.
- **Binary classification task:** Predict an attribute in IT given the other attributes in IT and the attribute in ST (See Anatomy Model)

Anatomy Model (Xiao et al. 2006)

- Private data in this research problem
- Based on the concepts of table partition and ℓ-diversity
- Every instance in IT can be matched to ℓ instances in ST (See example, ℓ=2)

Age (A)	Zipcode (Z)	Job (J)	GID (G)
41	47905	Assoc. Prof	1
29	47906	Assist. Prof	1
30	47906	Assist. Prof	2
35	47907	Assoc. Prof	2
28	47906	Assist. Prof	3
47	47905	Prof.	3
45	47905	Prof.	4
31	47906	Assist. Prof	4

Identifier table (IT)

GID (G)	Income (I)
1	[100K-150K)
1	[50K-75K)
2	[75K-100K)
2	[50K-75K)
3	[75K-100K)
3	[100K-150K)
4	[100K-150K)
4	[75K-100K)

Sensitive table (ST)

ℓ =2

Possible Values

Support Vector Classifier

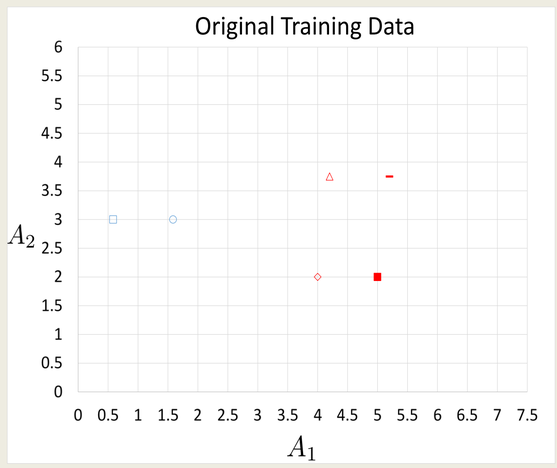
- Widespread usage
- Solid theory under some assumptions

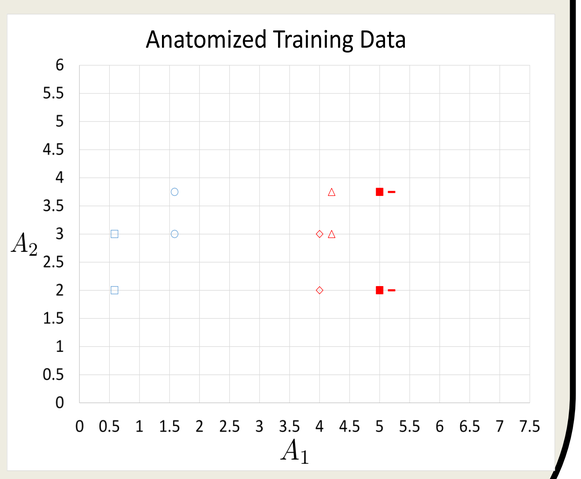
Challenge 1: How do we efficiently preprocess the training set to get the most effective information without violating ℓ-diversity?

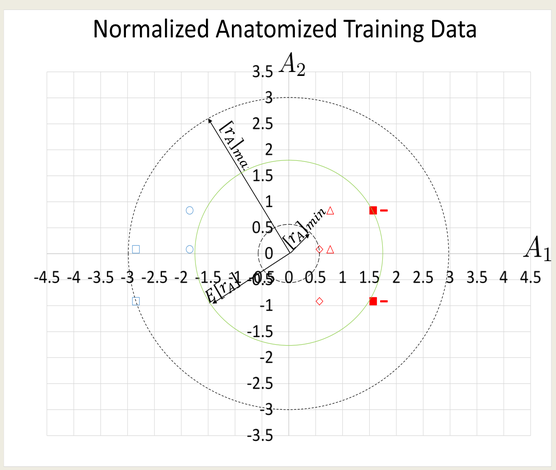
Challenge 2: How does the support vector classifier perform in practice after preprocessing?

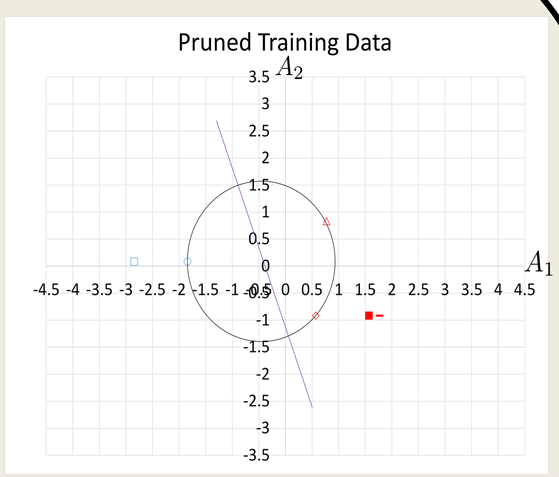
Training Set Pruning Algorithm (Mancuhan and Clifton 2017):

Idea: For every instance in IT, pick an instance in ST such that **the support vector classifier is expected to perform best (Best generalization). Prune out other $\ell - 1$ instances.** (See example below $\ell = 2$)









Privacy Preservation: Preserves ℓ-diversity as picking an instance in ST that is expected to perform best doesn't imply a correct matching between IT and ST

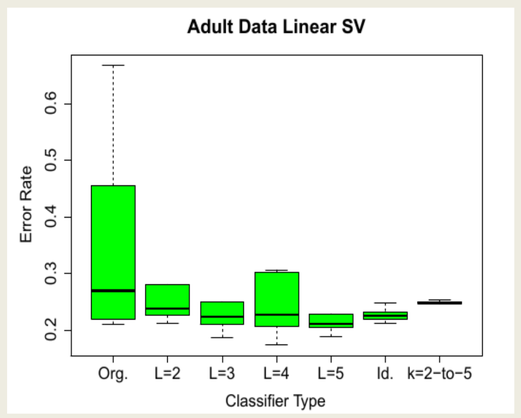
Implementation: Time complexity is linear in function of the training data size. Constant space complexity.

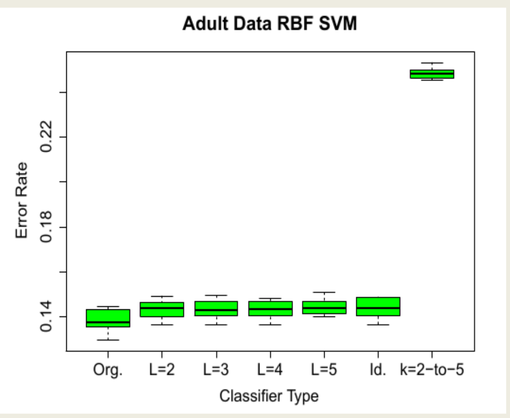
Theoretical Analysis: The support vector classification after pruning training set algorithm could be as accurate as it would be if the training data is not in anatomy (no ℓ-diversity) under two conditions:

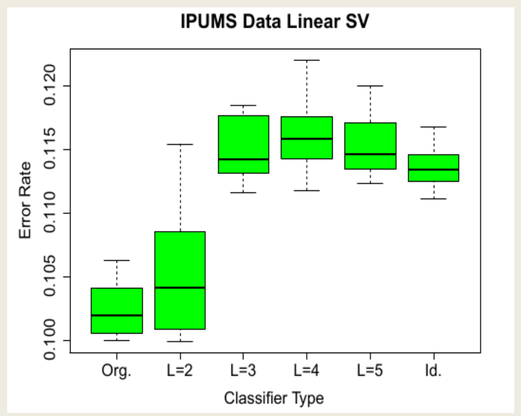
1. Training set is sufficiently large
2. Sensitive attribute domain size is small

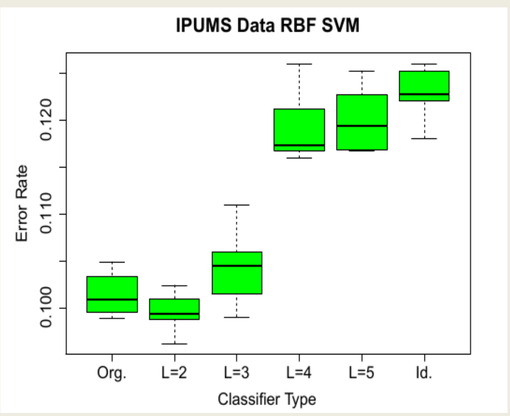
Experiments and Results

- Outperforms alternative methods that are developed for privacy standards such as k-anonymity.
- Accuracy performance comparable to support vector classification trained on the original data.
- 10 fold cross validation on multiple datasets available at UCI repository. Statistically significant according to t-test.









Conclusion

- Similar results to our previous work with K-Nearest Neighbor Classification (Mancuhan and Clifton 2016)
- Analytical capabilities and preserving privacy is possible (actually quite good)
- Publishing data with privacy constraints is a reasonable option!

Acknowledgments

This work was made possible by the support of the Northrop Grumman Cybersecurity Consortium

CERIAS

PURDUE UNIVERSITY