Read Me

Srushti Gurav

UID-U00809171

Multiplicative Data Perturbation

Multiplicative data perturbation applies to multidimensional datasets (i.e., at least two dimensions) with a combination of linear transformation and random noise injection. The specific methods include rotation perturbation, random projection perturbation, **geometric data perturbation (GDP),** and random space perturbation (RASP). Some (e.g., RASP) are more resilient to *background-knowledge based attacks* than others (e.g., rotation perturbation). GDP is used to address the privacy concern for outsourced data mining. In GDP the original datasets are changed (Data perturbation) so that the public cloud service provider cannot access original value. The GDP equation -

$$G(X) = RX + \Psi + \Delta;$$

Where,

G(X) - perturbed vector

R - multiplicative transformation

Ψ - translation transformation,

and Δ - distance transformation.

Getting Started

These instructions will get you a copy of the project up and running on your local machine for development and testing purposes.

Prerequisites and Installations

Python (*Installation Guidelines*)	2.7:	(or	above)
Weka (*Installation Guidelines*)			
Datasets (*Source link*)			

Initial Setup

- Source code location - /PAC Final Project/multiplicativePerturbation.py
- Original datasets - /PAC Final Project/wdbc.csv

Built With

- Notepad++ - Used to develop code
- IDLE - Used to run the code
- WEKA - Used for classification

How to run the GDP code

- Open the file multiplicativePerturbation.py using IDLE
- Go to Run -> Run Module (F5)
- You can also run it using command prompt.

Output

- 4 output matrix files will be generated -
  -/PAC Final Project/wdbc_0.1.csv
  -/PAC Final Project/wdbc_0.2.csv
  -/PAC Final Project/wdbc_0.3.csv
  -/PAC Final Project/wdbc_0.4.csv

How to run the classifiers

- Click the Open file button to open a data set and double click on the data directory.
- Select wdbc.csv file to load the dataset
- After loading the dataset choose a filter (default is none)
- Click Classify tab
- Choose a classifier
- Click the Start button to run.

Output

- You can note the results in the Classifier Output section.
- Standard deviation is calculated in the Selected Attribute tab after loading file.
- Classification accuracy can be observed using the mean of correlation coefficient.

Experiment evaluation results

- Invariance property of the KNN and perceptron is verified with sigma values of noise distribution as 0.0, 0.1, 0.2, 0.3, 0.4 (WEKA is used to classify this data)
- Accuracy and standard deviation is calculated. By calculating the difference of model accuracy, between the classifier trained with the original data and those trained with the perturbed data we can observe that the model accuracy is fully preserved the model accuracy for all of the classifiers (perceptron could be sensitive for some data sets).

Conclusion

All the information cannot be preserved so this study shows that preserving task-specific information selectively in perturbation will guarantee privacy and data utility and to preserve this information GDP method can be used.

Authors

- **Keke Chen** –
    - Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining
    - Building Confidential and Efficient Query Services in the Cloud with RASP Data Perturbation

Acknowledgments

- To read the datasets, code provided by Prof. Keke Chen in this link was used.

Checklist

- ✓ Task 1 - Implement the GDP algorithm
- ✓ Task 2 - Experiments to evaluate the GDP approach
- ✓ Task 3 - Evaluate accuracy for GDP (will be refined in the report)
- ■ Task 4 - Analyze the inference attacks
- ✓ Task 5 - Conclusion
- ■ Task 6 - Submit project reports.

References

Keke Chen and Ling Liu, " Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining ", Journal of Knowledge and Information Systems (KAIS), 2011

Keke Chen and Ling Liu: "Towards Attack Resilient Geometric Data Perturbation ", *SIAM International Data Mining Conference, 2007 (SDM07)*

Keke Chen and Ling Liu: "Privacy-Preserving Data Classification with Rotation Perturbation ", *Proc. of IEEE Intl. Conf on Data Mining 2005 (ICDM05)*.