Detect the watermark through the training model

A watermarking scheme to protect numerical classification datasets

Author: Ruonan Li Supervisor: Devris Isler Responsible Professor: Dr Zekeriya Erkin

Introduction

- Traditional digital watermarking technology is mainly used to protect the intellectual property of multimedia data such as images, audio and video[1].
- To date, dataset watermarking is a relatively new topic in research.
- Sablayrolles et al. proposed the radioactive data method[2] to protect image datasets. We modify the radioactive data method to protect numerical machine-learning classification datasets.

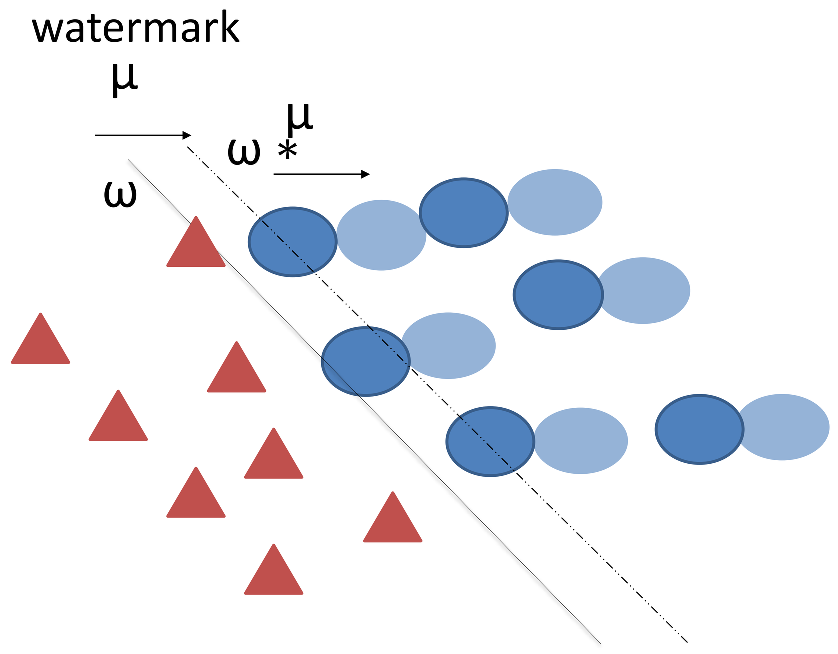
Research Questions

- How to make use of the watermarking technique to protect the intellectual property of numerical datasets used for machine learning?
- How much distortion will the watermark bring into the dataset?
- How robust is the watermark?

Method

- Watermark embedding:

Figure1: Illustration of embedding the watermark



μ is a unit vector (|μ| = 1) generated randomly and has the same dimension as the dataset. It is added in the feature space of data with the same label. After being trained with the watermarked dataset, the new classifier ω ∗ is likely to move in the direction of μ.

- Watermark detection:

The detection is based on statistical hypothesis testing.
H0: the model has been trained with the watermarked dataset. H1: the model has not been trained with the watermarked dataset.
The cosine similarity between two vectors v1 and v2 (c(v1,v2)) in high-dimensional space of dimension d follows an incomplete distribution[3]:

$$P(c(v_1,v_2)\geq\tau)=\frac{B_{1-\tau^2}(\frac{d-1}{2},\frac{1}{2})}{2B_1(\frac{d-1}{2},\frac{1}{2})}$$

with

$$B_x(\frac{d-1}{2},\frac{1}{2})=\int_0^x\frac{(\sqrt{t})^{d-3}}{\sqrt{1-t}}dt$$

We can calculate c(ω ∗ − ω, μ) and get the p-value.

Results

The effectiveness of our method:

Table 1: Accuracy and log_{10}(p) for each dataset with q of data modified

		q= 0	0.01	0.02	0.05	0.1
Iris	accuracy	0.97	0.967	0.967	0.9	0.833
	$\log_{10}(p)$	-0.30	-0.625	-0.876	-0.826	-1.07
Wine	accuracy	1	1	1	1	1
	$\log_{10}(p)$	-0.30	-2.2	-2.7	-3.3	-5.0
Breast	accuracy	0.982	0.982	0.982	0.976	0.976
	$\log_{10}(p)$	-0.30	-3.35	-3.45	-4.63	-4.74

The robustness of against data normalization:

Table 2: Cosine similarity between $\omega^* - \omega$ and μ after normalizing the data

	q= 0	0.01	0.02	0.05	0.1
Iris	0	-0.49	-0.49	-0.49	-0.49
Wine	0	-0.15	-0.13	-0.06	0.02
Breast	0	-0.53	-0.51	-0.42	-0.38

The robustness against column dropping:

Table 3: log_{10}(p) when dropping c columns with q of data modified, "/" means the cosine similarity between $\omega^* - \omega$ and μ is negative

		c = 1	2	3
q = 0.1	Wine	-0.65	-0.48	/
	Breast	-4.73	-4.61	-3.2
q = 0.05	Wine	/	/	/
	Breast	-1.92	-2.3	-1.78
q = 0.01	Wine	/	/	/
	Breast	-0.83	-0.62	-0.59

Conclusions

- With only 1% of data modified, we can detect if a linear classification model has been trained with the watermarked dataset with more than 99% of confidence.
- Our method is robust against column dropping when the dimension of the dataset is high.
- However, after performing data normalization, we cannot detect the watermark anymore.

Further Work

- The method can be improved so that it is effective for non-linear models.
- It is common to apply data normalization before training a machine learning model but our method is not robust against data normalization and this needs to be improved.

Contact

Ruonan Li

Email: R.Li-11@studelt.tudelft.nl

References

1. Arezou Soltani Panah, Ron G. van Schyndel, Timos K. Sellis, and Elisa Bertino. On the properties of nonmedia digital watermarking: A review of state of the art techniques. IEEE Access, 4:2670–2704, 2016.
2. Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Herve J ́egou. Radioactive data: tracing through training. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 8326–8335. PMLR, 2020.
3. Ahmet Iscen, Teddy Furon, Vincent Gripon, Michael G. Rabbat, and Herve J ́egou. Memory vectors for similarity search in high-dimensional spaces. IEEE Trans. Big Data, 4(1):65–77, 2018.