Modeling Heterogeneous Statistical Patterns In High-dimensional Data By Adversarial Distributions: An Unsupervised Generative Framework

Han Zhang[1], Wenhao Zheng[2], Charley Chen[1], Kevin Gao[1], Yao Hu[2], Ling Huang[4], Wei Xu[1]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Alibaba Youku Cognitive and Intelligent Lab, [4]AHI Fintech

交叉信息研究院
Institute for Interdisciplinary Information Sciences

Motivation

- The intrinsic clusters in high-dimensional data may display heterogeneous statistical patterns.
- Specifically, different clusters display clustering patterns w.r.t. different features.
- For example, different fraud groups may share the IP address, phone number, ID card number etc.

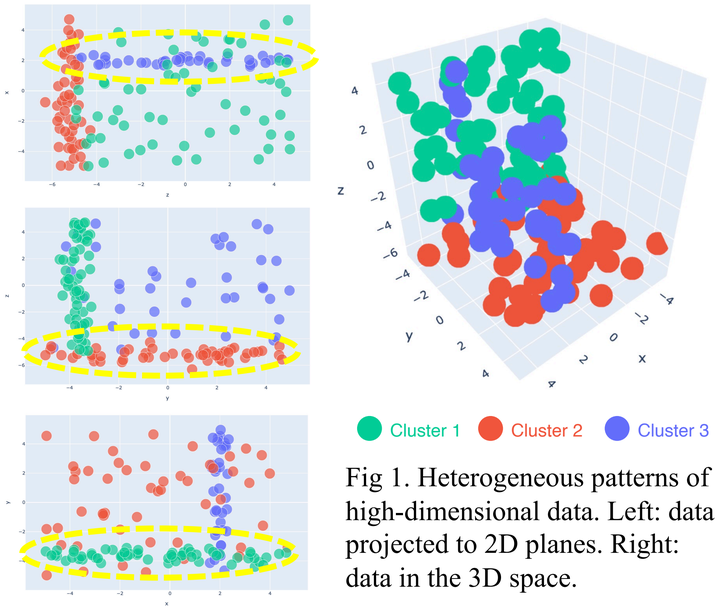


Fig 1. Heterogeneous patterns of high-dimensional data. Left: data projected to 2D planes. Right: data in the 3D space.

FIRD: a generative framework

- **F**eature **I**ndependent assumption: within each data cluster, the features are independent with each other.
- **AdveRsarial Distributions**: for each feature within a cluster, a pair of distribution **compete** with each other for generating the observations.

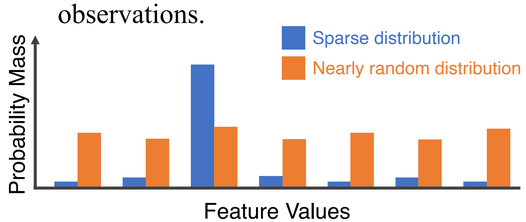


Fig 2. Example of discrete adversarial distributions. The sparse distribution models the fraud, and the random part models normal user.

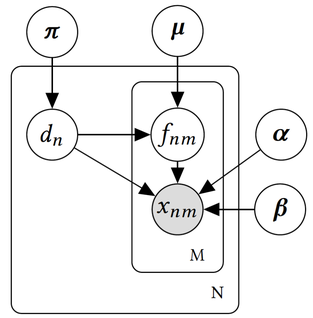


Fig 3. The plate representation of FIRD in discrete space. The parameter π is the mixture weight. μ balances the adversarial distribution pairs. The synchronization and randomness are captured by the adversarial distribution pairs, whose parameters are α and β, respectively.

- The generation process for an observation x_n:
  1. Choose cluster $d_n \sim Multinomial(\pi)$
  2. For each feature $\mathcal{F}_m$:
     a) Choose the indicator variable $f_{nm} \sim Bernoulli(\mu_{d_n m})$;
     b) If $f_{nm} = 1$, choose $x_{nm} \sim Multinomial(\boldsymbol{\alpha}_{d_n m})$;
     c) If $f_{nm} = 0$, choose $x_{nm} \sim Multinomial(\boldsymbol{\beta}_{d_n m})$;
- The objective function (log-likelihood):

$$\log \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}; \mathcal{D}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$$

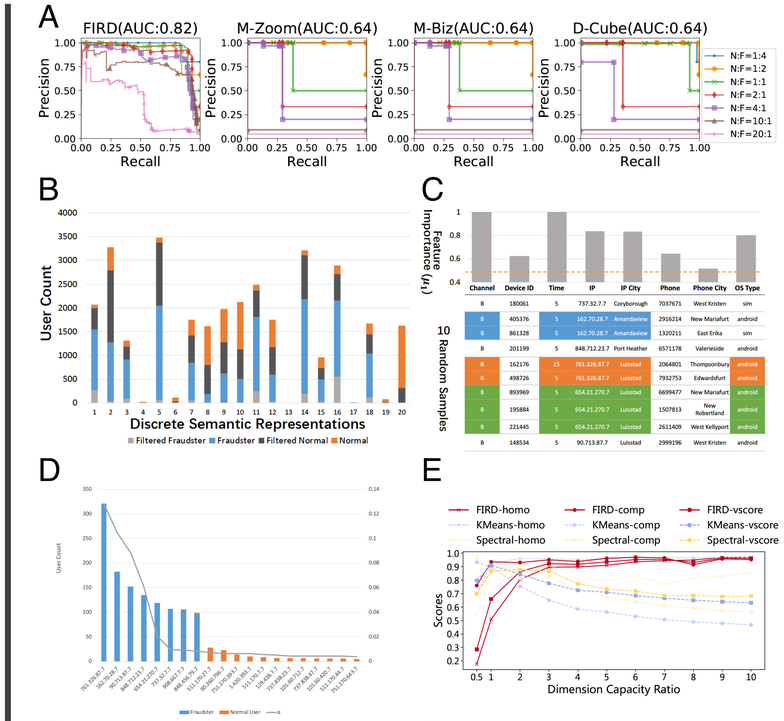
$$= \sum_{n=1}^{N} \log \left\{ \sum_{d_n, \mathbf{f}_n} p(\mathbf{x}_n | \mathbf{f}_n, d_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\mathbf{f}_n | d_n, \boldsymbol{\mu}) p(d_n | \boldsymbol{\pi}) \right\}$$

$$- \sum_{g=1}^{G} \lambda_g^{(1)} \log \pi_g - \sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{i=1}^{D_m} \lambda_{gmi}^{(2)} \left(\log \alpha_{gmi} - \log \beta_{gmi} \right)$$

Promote Sparsity — Promote Randomness

Results

- We evaluate FIRD on 3 types of datasets:
  1. E-commerce platform registration dataset;
  2. anomaly detection benchmark dataset;
  3. synthetic datasets according to FIRD's generation process.
- We show that FIRD is able to:
  1. detect fraud groups much better than comparison methods;
  2. work as a general anomaly detection method;
  3. provide significant performance with low time cost.



H

Dataset	FIRD	HBOS	IForest	OCSVM	LSCP
cardio	**0.949**	0.843.	0.924	0.938	0.901
musk	**1.000**	**1.000**	0.999	**1.000**	0.998
optdigits	**1.000**	0.865	0.714	0.500	-
satimage-2	**0.998**	0.977	0.993	0.997	0.9935
shuttle	0.990	0.986	**0.997**	0.992	0.5514
satellite	**0.900**	0.754	0.701	0.660	0.6015
ionosphere	**0.946**	0.5569	0.8529	0.8597	-
pendigits	**0.972**	0.9247	0.9435	0.931	0.8744
wbc	0.944	**0.954**	0.9325	0.9376	0.945