Shallow to Deep Non-IID Learning: Complex Systems, Behaviors and Data

Longbing Cao

University of Technology Sydney, Australia

Data Science Lab: www.datasciences.org

Acknowledgement

- Thanks to all past and present members at Prof Longbing Cao's team who made contributions to this slide and relevant research, including Dr Yanchang Zhao, Dr Huaifeng Zhang, Dr Can Wang, Dr Yuming Ou, Dr Jinjiu Li, Dr Chunming Liu, Dr Fangfang Li, Dr Bin Fu, Dr Xin Cheng, Dr Liang Hu, Dr Guansong Pang, Mr Chengzhang Zhu, Dr Trong Dinh Thac Do, and Ms Songlei Jian, Dr Shoujin Wang, etc.

Slides and info about non-IID learning

- http://noniid.datasciences.org/
- 2022 guest lecture on Shallow to deep non-IID learning: https://www.youtube.com/watch?v=ciBZFj1Jtn8
- KDD2017 tutorial on non-IID learning Youtube videos: https://www.youtube.com/watch?v=3RwyGoiYcLg
- IJCAI2019 tutorial Non-IID Learning of Complex Data an https://datasciences.org/publication/Non-IID%20Learni Full.pdf

Agenda

- IID Learning and issues
- Non-IIDness
- Non-IID similarity/metric learning
- Non-IID representation learning
- Coupling learning: complex interactions and relations
- Heterogeneity learning
- Non-IID learning tasks and applications:
  - Non-IID pattern mining
  - Non-IID statistical learning
  - Non-IID recommender systems
  - Non-IID behavior analytics
  - Non-IID vision learning
  - Non-IID outlier detection
  - Out-of-distribution detection
  - Non-IID document analysis
  - Non-IID ensemble learning
  - Non-IID federated learning
  - Domain adaptation

IID Learning and Issues

IID learning dominates classic analytics and learning in AI/KDD/ML/CVPR/Statistics research

Mathematically/statistically defined IID/i.i.d.

- Data set $D=\{\mathbf{X}, y\}$ is composed of N input & response tuples ($\boldsymbol{X}_i$, y_i) that are *independently drawn from the same joint distribution* P($\mathbf{X}$, y):

  ($\boldsymbol{X}_i$, y_i) ~ P($\mathbf{X}$, y)

- and a learning algorithm is built to learn

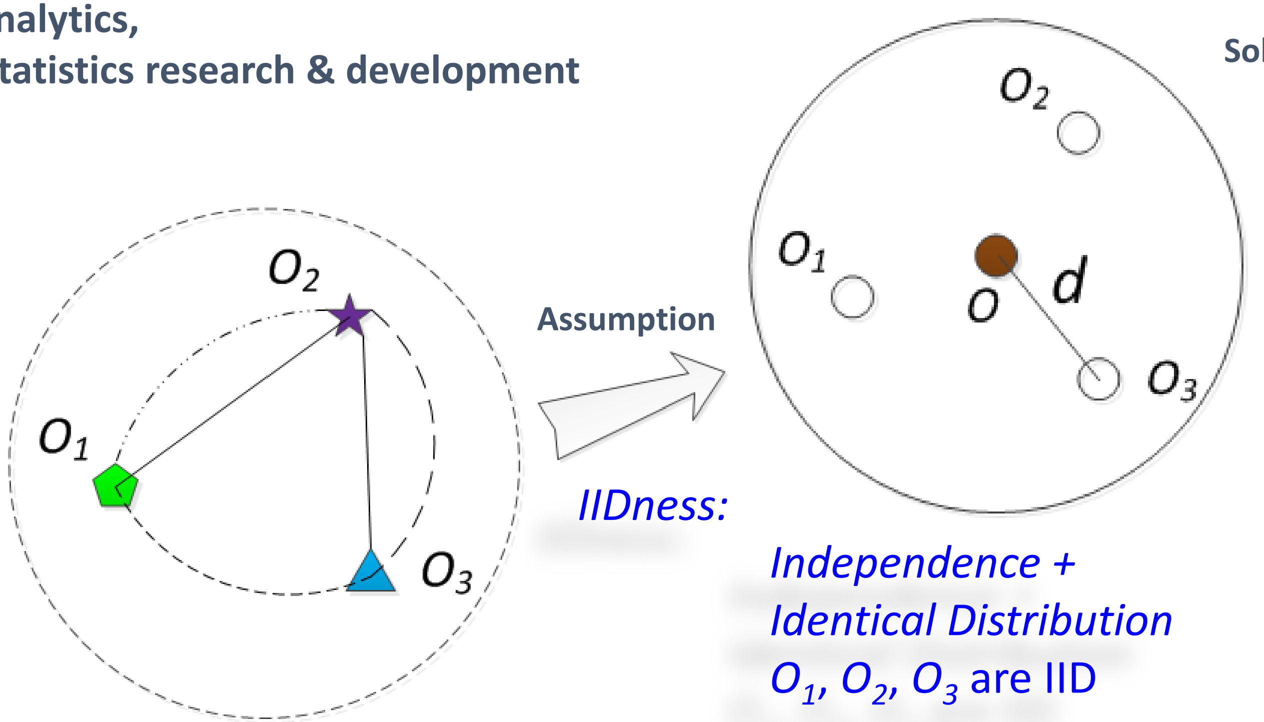
  $p(y|\mathbf{X}) = p(\mathbf{X}, y)/p(\mathbf{X})$

  where ($\boldsymbol{X}_i$, y_i) are independent of ($\boldsymbol{X}_j$, y_j)

Classic Assumption – IIDness & IID Learning

IID learning:
Dominates classic analytics,
AI/KDD/ML/CVPR/Statistics research & development

Solution



O_1, O_2, O_3 are IID:
An independent O:
$d_3 = ||O_3 - O||$

IIDness:

Independence +
Identical Distribution
O_1, O_2, O_3 are IID

Problem

Learning a Model of *y* Given X

- Discriminative learning
  - Learn a model p(y|X)
  - Model:
    - Supervised: e.g., neural networks, decision trees, random forest, etc.
    - Unsupervised: e.g., clustering, adversarial learning, autoencoder, contrastive learning

Assuming:

- Learn the model on each individual sample X_i in the set $\{X_i\}$: $p(y_i|X_i)$
- $p(y_i|X_i)$: each target y_i is conditionally independent given the independence of X_i
- No specific distributional assumption on each sample X_i (*i.e., i.d.)*

Learning A Model of *y* Given X

- Generative learning
  - Learn the joint probability p(X, *y*) of (X, *y*), i.e., by
    - Learning conditional probability p(X|*y*) with marginal distribution p(*y*)
    - Then learning p(*y*|X) (e.g., by Bayes' theorem)
  - Models:
    - Unsupervised: e.g., regressors, variational autoencoder
    - Pattern mining: e.g., associate rule mining, negative sequence analysis
    - Estimation: like linear discriminant analysis, Bayesian networks

Assuming:

- y_i and y_j are IID
- X_i and X_j are IID
- Learn p(X|*y*) from i IID samples: p(X|*y*) = $\prod_i p(X_i|y_i)$
- IID in transforming from p(X|*y*) to p(*y*|X)

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Distance measures and functions

- Objects/variables are IID
- Variables are random

- Euclidean distance: $d(x_1,x_2)$
- Hamming distance: $d(s_1,s_2)$
- Mahalanobis distance $d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})}.$

Questions & thinking:
- **What if objects or variables are dependent?**
- **What if they follow different distributions?**

Statistics of Data

- Variance of samples

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Questions & thinking:
- What if objects x_i and x_j are dependent?
- What if they follow different distributions?

- Covariance of variables

$$\mathrm{cov}(x, y) = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Questions & thinking:
- x and y are not with the same distribution and have diff means
- What if x and y are dependent?

- Cross entropy

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\,\log q(x)$$

$$H(p,q) = -\int_X P(x)\,\log Q(x)\,dr(x)$$

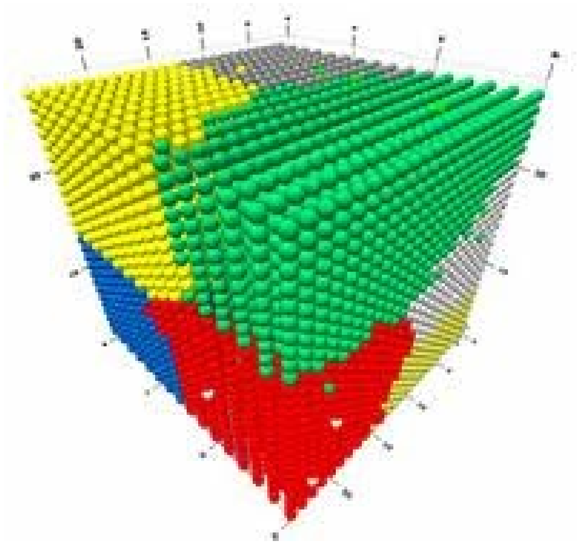
- KL-divergence/relative entropy

$$D_{KL}(p||q) = H(p,q) - H(p)$$

Questions & thinking:
- What if distributions p and q are dependent?

IID K-means

Clustering



Objective functions:

-K-means

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Note:
- x_j Individual objects only!
- S_i individually

-Fuzzy C-Means

$$J_{\mathrm{FCM}}(\boldsymbol{\mu}, \boldsymbol{A}) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m ||\boldsymbol{x}_j - \boldsymbol{a}_i||^2$$

$$\sum_{i=1}^{c} \mu_{ij} = 1 \quad \text{for all } j \in J.$$

Questions:
- What if x_{j1} and x_{j2} are dependent?
- What if clusters are not independent (overlappted etc.)?

What Makes K-means IID?

Objective functions:

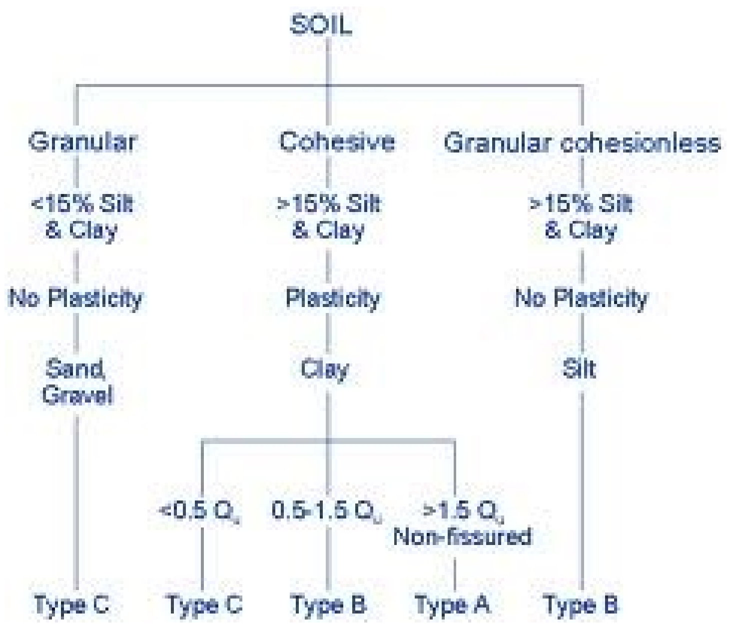
-K-means

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

- Object IIDness:
    - Object independence: X_j does not involve interactions with other objects/variables $\{X_k\}$
- Cluster IIDness:
    - Assume all clusters are independent
- Global to local:
    - Learning analytical goal: global task → local cluster
    - Global partition → local distribution (mean μ_i)

IID Decision Tree

Classification



Objective functions:

-Decision tree

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

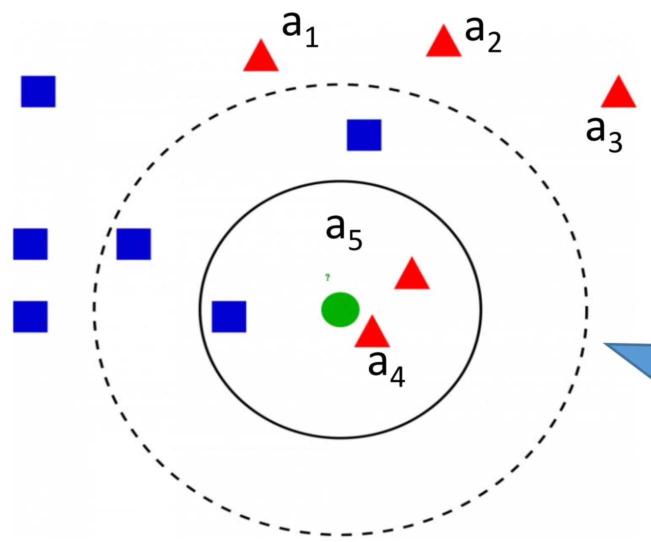
$$\overbrace{E_A\left(IG(T,a)\right)}^{\text{Expected Information Gain}} = \overbrace{I(T;A)}^{\text{Mutual Information between T and A}} = \overbrace{\mathrm{H}(T)}^{\text{Entropy (parent)}} - \overbrace{\mathrm{H}(T|A)}^{\text{Weighted Sum of Entropy (Children)}}$$

$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_a p(a) \sum_{i=1}^{J} -\Pr(i|a) \log_2 \Pr(i|a)$$

Questions & thinking:
- **T: The data set**
- **A: An attribute**
- **a: A value of A**
- **X: samples**
- **Y: The label set**
- **J: The number of classes**
- **p_i: the probability of class I**
- **p_a: the probability of value a**

Questions & thinking:
- **What if objects x_k and x_j are dependent?**
- **What if values a_1 and a_2 are dependent?**
- **What if classes i_1 and i_2 have different distributions?**

IID kNN

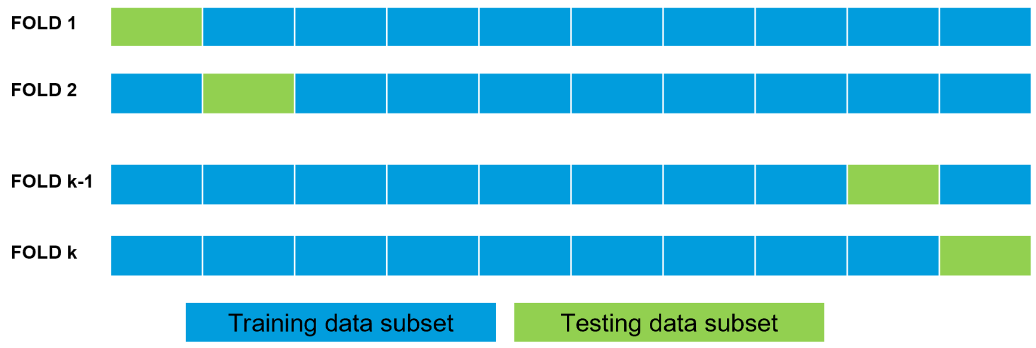


Questions & thinking:

- The label of c is determined by its k neighbors, which are IID
- What if objects x_i and x_j are dependent?
- What if neighbors are dependent?
- If all red triangles are coupled with each other, the same for the blue squares, what would be the label of green object?
- What if some of the red ones are coupled with some blue ones?
- What if the distributions of triangles and squares are different?

IID K-fold Cross Validation & Sampling, Batching

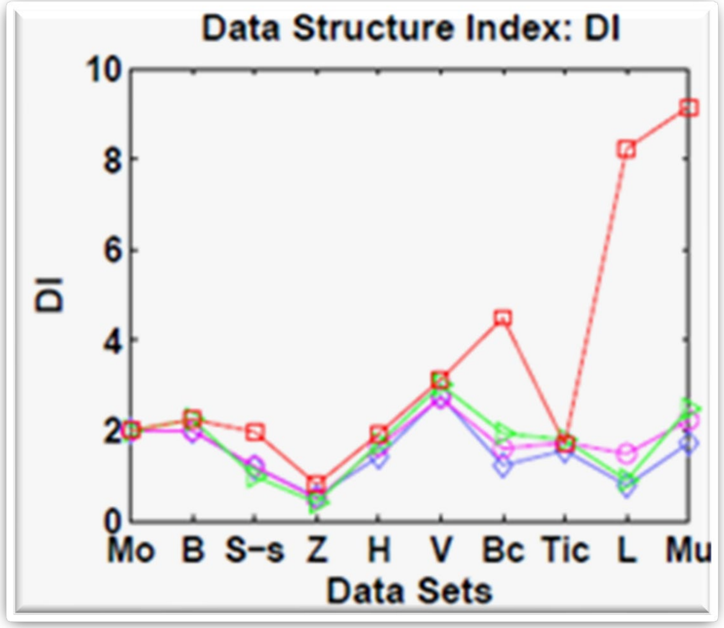
- Randomly sample k-folds



Questions & thinking:
- What if the samples in the data are non-IID?
- What if the samples in the training set are non-IID?
- What if the samples in training set and the test set are non-IID? ie OOD problem

Potential Risk of IID Assumption

- Results delivered by IID analytical/learning methods/algorithms on non-IID data could be:
  - incomplete
  - biased, or even
  - misleading
- Many ‘benchmarks’ may be unfair and wrong



Questions & thinking:
- Why learning bias exist?
- Beyond fitting issues, what other issues may have caused learning bias?

Non-IIDness

Longbing Cao. Non-IIDness Learning in Behavioral and Social Data, The Computer Journal, 57(9): 1358-1370 (2014).

Cao, Longbing. *Coupling Learning of Complex Interactions*, IP&M, 51(2): 167-186 (2015)

Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Applications, IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012)

Can Wang, Longbing Cao, Minchun Wang, Jinjiu Li, Wei Wei, Yuming Ou. Coupled Nominal Similarity in Unsupervised Learning, CIKM 2011, 973-978.

Mathematically/statistically defining IID/i.i.d.

- Data set $D=\{\mathbf{X}, y\}$ is composed of N input & response tuples $(\boldsymbol{X}_i, y_i)$ that are *independently drawn from the same joint distribution* $P(\mathbf{X}, y)$:

  $(\boldsymbol{X}_i, y_i) \sim P(\mathbf{X}, y)$

- A learning algorithm is built to learn

  $p(y|\mathbf{X}) = p(\mathbf{X}, y)/p(\mathbf{X})$

  where $(\boldsymbol{X}_i, y_i)$ are independent of $(\boldsymbol{X}_j, y_j)$

Question:
- Learning $p(y|\mathbf{X})$ in terms of $p(y_i|X_i)$ on each sample i
- What if $(\boldsymbol{X}_i, y_i)$ and $(\boldsymbol{X}_j, y_j)$ are coupled ($\overline{I}$)?
- What if $(\boldsymbol{X}_i, y_i) \sim P_i(\mathbf{X}, y)$ and $(\boldsymbol{X}_j, y_j) \sim P_j(\mathbf{X}, y)$ are heterogeneous ($\overline{ID}$)?

Mathematically/statistically defining IID/i.i.d.

- $\mathbf{X}_i$ is *d*-dimensional, i.e., *d*-variate vector/variable

  $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})$

  What if features X_m and X_n are not independent?
- What if features X_m and X_n are not identically distributed?

  $p(X_m)$ and $p(X_n)$ are different
- What if label classes y_i and y_j are dependent?
- What if label classes y_i and y_j follow different distributions $P_i(y)$ and $P_j(y)$?

Non-IIDness in Big and Small Data

- **Heterogeneity:**
  - Data types, attributes, sources, aspects, …
  - Formats, structures, distributions, relations, …
  - Learning objectives, learning results/targets

non-identically distributed.

- **Coupling and interaction:**
  - Within and between values, attributes, objects, sources, aspects, …
  - Structures, distributions, relations, …
  - Methods, models, …
  - Results, targets, impact, …

Non-independent.

Non-IIDness

L. Cao. Non-IIDness Learning in Behavioral and Social Data, The Computer Journal, 57(9): 1358-1370 (2014).

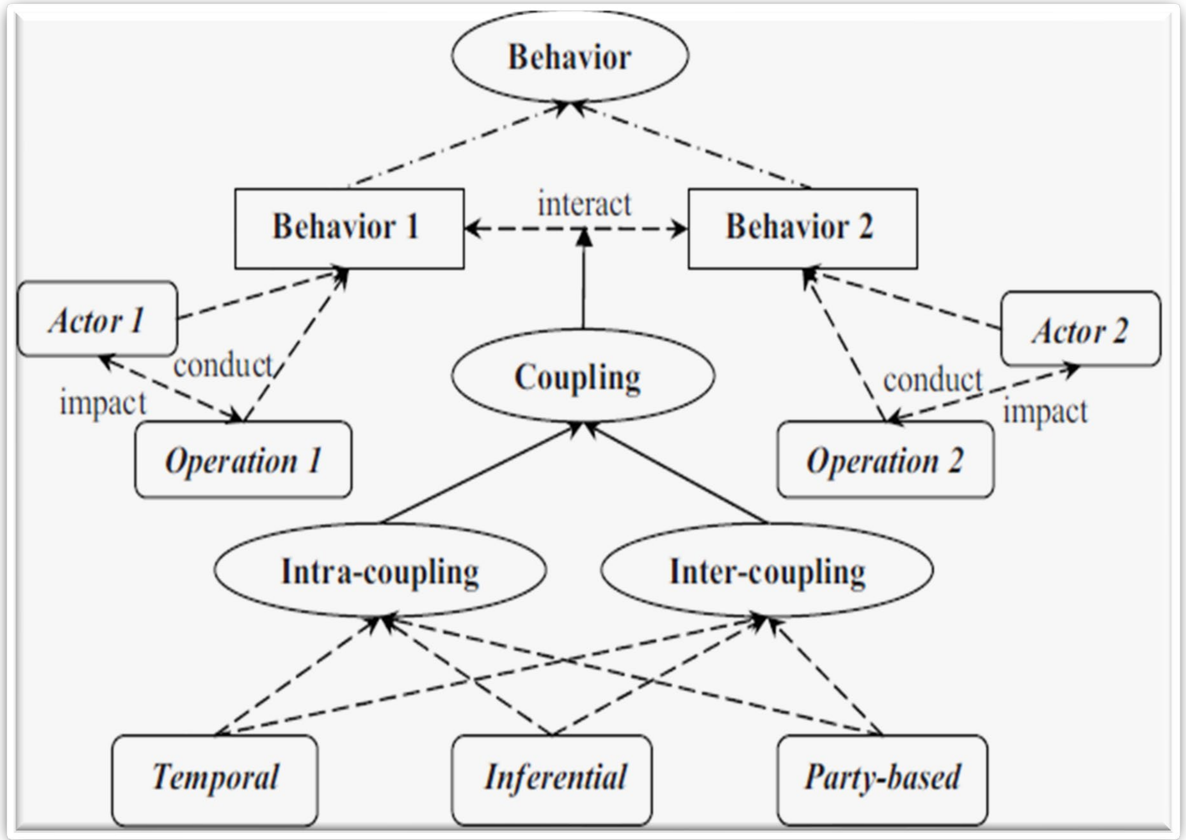
L. Cao. *Coupling Learning of Complex Interactions*, IP&M, 51(2): 167-186 (2015)

Couplings/Interactions vs. Common Relations

- Couplings and interactions: numerical, categorical, textual, mixed-structure, syntactic, semantic, organizational, social, cultural, economic, uncertain, unknown/latent relation etc.
- Coupling and interaction go beyond existing relations including Dependence, Correlation, Association and Causality
- Mathematically, Association, Causality, Correlation, and Dependence are specific, descriptive, explicit, etc.
- Couplings: explicit + implicit, qualitative + quantitative, descriptive + deep, specific + comprehensive, local + global, etc.

Can Wang, Fosca Giannotti, Longbing Cao. Learning Complex Couplings and Interactions. IEEE Intell. Syst. 36(1): 3-5, 2021.
L. Cao. Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning, IEEE Intelligent Systems, 37:4, 3-15, 2022

Example: Behavior Couplings



- **Instance Of** Connecting instances (in Rectangle) to their corresponding classes
- **Subclass Of** Linking a subclass (in Oval) to its parent class
- **Object Property** Denoting the relationships between instances, between an object and its properties (in Rounded Rectangle), or between properties.

Can Wang, Longbing Cao, Chi-Hung Chi. [Formalization and Verification of Group Behavior Interactions](). IEEE T. Systems, Man, and Cybernetics: Systems 45(8): 1109-1124 (2015)

Example: Couplings in Behaviors

Behavioral Couplings

Perspectives

Temporal

- *Serial Coupling*
- *Parallel coupling*
  - *Synchronous relationship*
  - *Asynchronous coupling*
    - *Interleaving*
    - *Shared-variable*
    - *Channel system*

Inferential

- *Causal Coupling*
- *Conjunction Coupling*
- *Disjunction Coupling*
- *Exclusive Coupling*

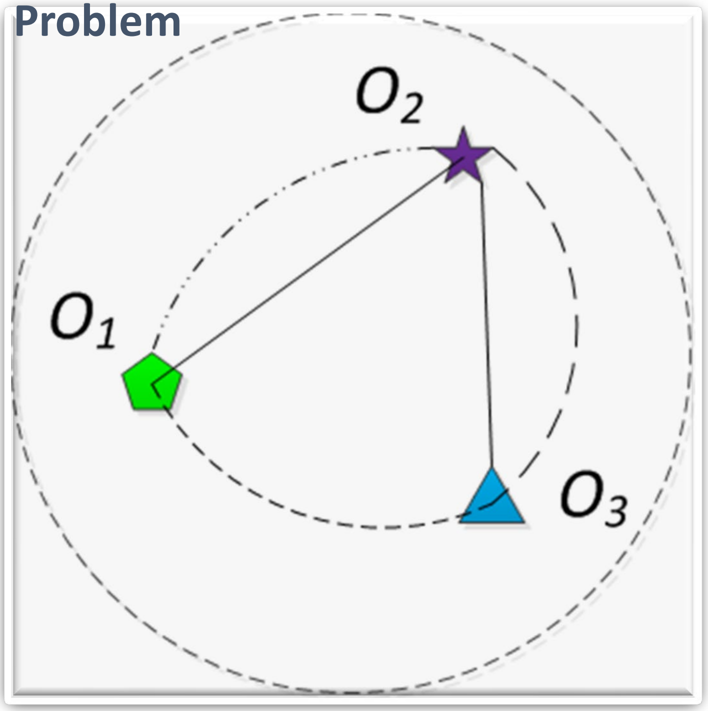
Party-based

- *One-Party-Multiple-Operation*
- *Multiple-Party-One-Operation*
- *Multiple-Party-Multiple-Operation*

Ex:
- Robotic games
- Customer interactions
- Transport
- Cyberattack

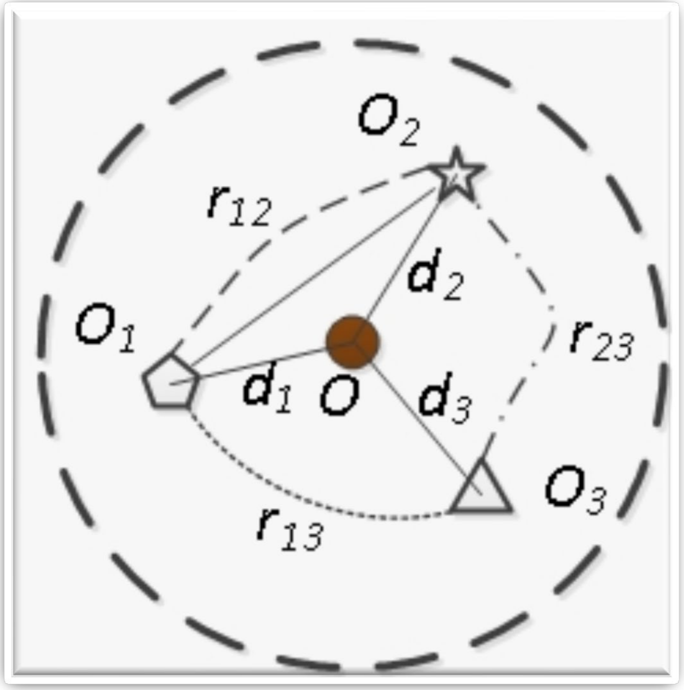
Can Wang, Longbing Cao, Chi-Hung Chi. Formalization and Verification of Group Behavior Interactions. IEEE T. Systems, Man, and Cybernetics: Systems 45(8): 1109-1124 (2015)

Beyond IID: Non-IID Learning

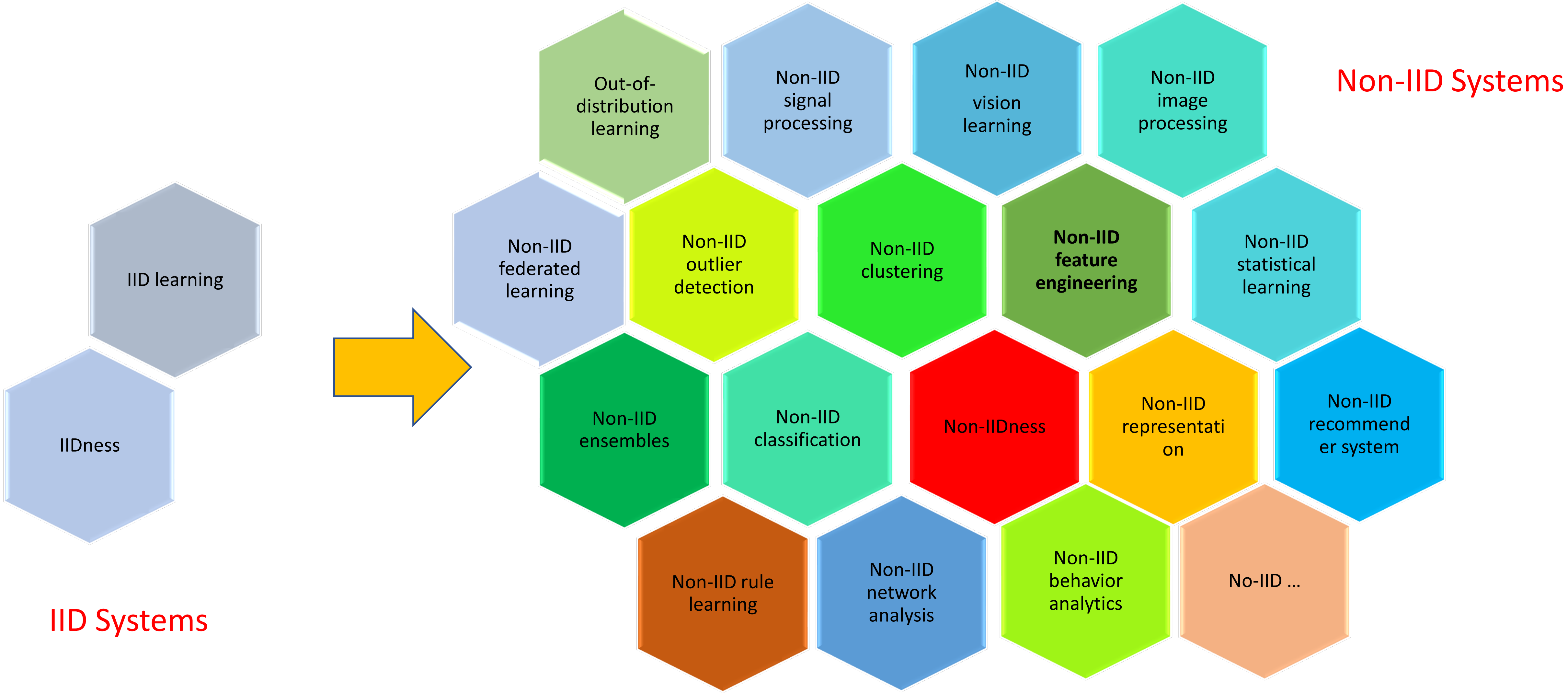


O_1, O_2, O_3 share different distributions

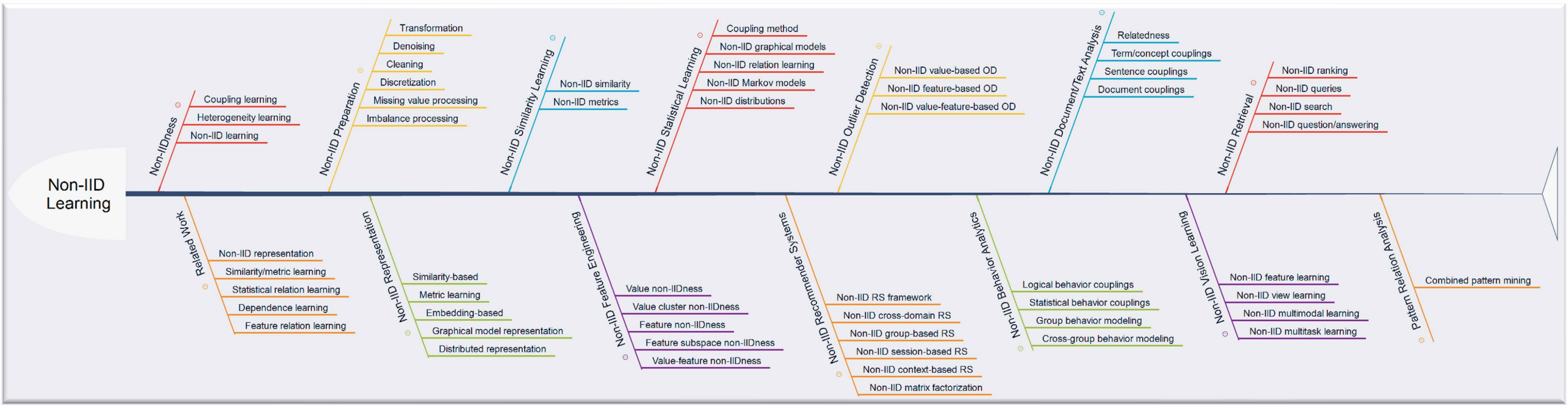
$$d_3 = ||O_3 - O||$$
$$= || O_3(r_{13}, r_{23}) - O(d_1, d_2) ||$$



IID to Non-IID Learning Systems



Landscape on non-IID Learning



Beyond IID in Information Theory

Beyond IID 4

- Home
- Organizers
- Accommodation
- Programme
- Posters
- Venue and Travel
- Participants
- Photos
- Beyond IID 1
- Beyond IID 2
- Beyond IID 3
- Beyond IID 5
- Beyond IID 6
- BIID Conference Series

Sponsors

erc

giQ

Site owners

Andreas Winter

Krishnakumar Sabapathy

Beyond IID in Information Theory 4

"Beyond IID in Information Theory" started as a workshop in Cambridge three years ago, organised by Nilanjana Datta and Renato Renner as a forum for the growing interest in information theoretic problems and techniques beyond the strict asymptotic limit, and aimed at bringing together researchers from a range of different backgrounds, ranging from coding theory, Shannon theory in the finite block length regime, one-shot information theory, cryptography, quantum information, all the way to quantum thermodynamics and other resource theories.

Quantum Shannon theory is arguably the core of the new “physics of information,” which has revolutionised our understanding of information processing by demonstrating new possibilities that cannot occur in a classical theory of information. It is also a very elegant generalisation, indeed extension, of Shannon's theory of classical communication. The origins of quantum Shannon theory lie in the 1960s, with a slow development until the 1990s when the subject exploded; the last 10-15 years have seen a plethora of new results and methods. Two of the most striking recent discoveries are that entanglement between inputs to successive channel uses can enhance the capacity of a quantum channel for transmitting classical data, and that it is possible for two quantum communication channels to have a non-zero capacity for transmitting quantum data, even if each channel on its own has no such quantum capacity.

In recent years, both in classical and quantum Shannon theory, attention has shifted from the strictly asymptotic point of view towards questions of finite block length. For this reason, and fundamentally, there is a strong drive to establish the basic protocols and performance limits in the one-shot setting. This one-shot information theory requires the development of new tools, in particular non-standard entropies and relative entropies (min-, Rényi-, hypothesis testing), both in the classical and quantum setting. These tools have found numerous applications, ranging from cryptography to strong converses, to second and third order asymptotics of various source and channel coding problems. A particularly exciting set of applications links back to physics, with the development of a resource theory of thermodynamic work extraction and more generally of state transformations. Physicists have furthermore found other resource theories, for instance that of coherence and that of asymmetry, which are both relevant to the thermodynamics of quantum systems and interesting in their own right.

The whole area is extremely dynamic, as the success of three previous "Beyond IID" workshops has shown.

Dates: 18-22 July 2016 (following ISIT 2016)

Venue: Institut d'Estudis Catalans - C/ del Carme, 47, 08001 Barcelona

Description:
The present workshop, the fourth in a series that started in 2013 in Cambridge, will bring together specialists and students of classical and quantum Shannon theory, of cryptography, mathematical physics, thermodynamics, etc, in the hope to foster collaboration in this exciting field of one-shot information theory and its applications. The plan is to have a modest number of talks over the course of the week. Participation is open to all, but the organisers request that everyone interested in attending does register.

Topics:
The topics covered under "Beyond IID" include but are not limited to the following:

-Finite block length coding
-Second, third and fourth order analysis
-Strong converses
-Quantum Shannon theory
-Cryptography and quantum cryptography
-New information tasks
-One-shot information theory and unstructured channels
-Information spectrum method
-Entropy inequalities
-Non-standard entropies (e.g. Rényi entropies, min-entropy, ...)
-Matrix analysis
-Thermodynamics
-Resource theories of asymmetry
-Generalised resource theories
-Physics of information

Non-IID Similarity/Metric Learning

Chengzhang Zhu, Longbing Cao and Jianpin Yin. Unsupervised Heterogeneous Coupling Learning for Categorical Representation. IEEE Transaction on Pattern Recognition and Machine Intelligence, 44(1): 533-549, 2022

Songlei Jian, Guansong Pang, Longbing Cao, Kai Lu and Hang Gao. CURE: Flexible Categorical Data Representation by Hierarchical Coupling Learning. IEEE Transactions on Knowledge and Data Engineering, 31(5): 853-866, 2019

Songlei Jian, Longbing Cao, Kai Lu, Hang Gao. Unsupervised Coupled Metric Similarity for Non-IID Categorical Data. IEEE Transactions on Knowledge and Data Engineering, 30(9): 1810 – 1823, 2018

Can Wang, Dong, Xiangjun; Zhou, Fei; Longbing Cao, Chi, Chi-Hung. Coupled Attribute Similarity Learning on Categorical Data, IEEE Transactions on Neural Networks and Learning Systems, 26(4): 781-797 (2015)

Similarity-based Representation

Can Wang, Longbing Cao, Minchun Wang, Jinjiu Li, Wei Wei, Yuming Ou. Coupled Nominal Similarity in Unsupervised Learning, CIKM 2011, 973-978.

Can Wang, Dong, Xiangjun; Zhou, Fei; Longbing Cao, Chi, Chi-Hung. Coupled Attribute Similarity Learning on Categorical Data (extension of the CIKM2011 paper), IEEE Transactions on Neural Networks and Learning Systems.

Motivation

Why these two people sit together at that place at that particular time?

Coupling Learning with feature interactions

TABLE 1. The Extended Information Table

$O \backslash A$	A_1	A_2	...	A_J	M_1	...	M_Q
O_1	$\mathcal{V}_{11}$	$\mathcal{V}_{12}$	...	$\mathcal{V}_{1J}$	C_{11}	...	C_{1Q}
O_2	$\mathcal{V}_{21}$	$\mathcal{V}_{22}$	...	$\mathcal{V}_{2J}$	C_{21}	...	C_{2Q}
...	...	...	...	...	...	...	...
O_n	$\mathcal{V}_{n1}$	$\mathcal{V}_{n2}$	...	$\mathcal{V}_{nJ}$	C_{n1}	...	C_{nQ}
...	...	...	...	...	...	...	...
O_N	$\mathcal{V}_{N1}$	$\mathcal{V}_{N2}$	...	$\mathcal{V}_{NJ}$	C_{N1}	...	C_{NQ}

- Feature interactions
- Feature-label couplings
- Object-feature-label couplings

Name	Gender	Performance	Commitment	Class
John	M	A	H	c1
Mary	F	B	H	c1
Sarah	F	B	I	c1
David	M	C	L	c1
Alice	F	C	I	c2
Edward	M	D	L	c2

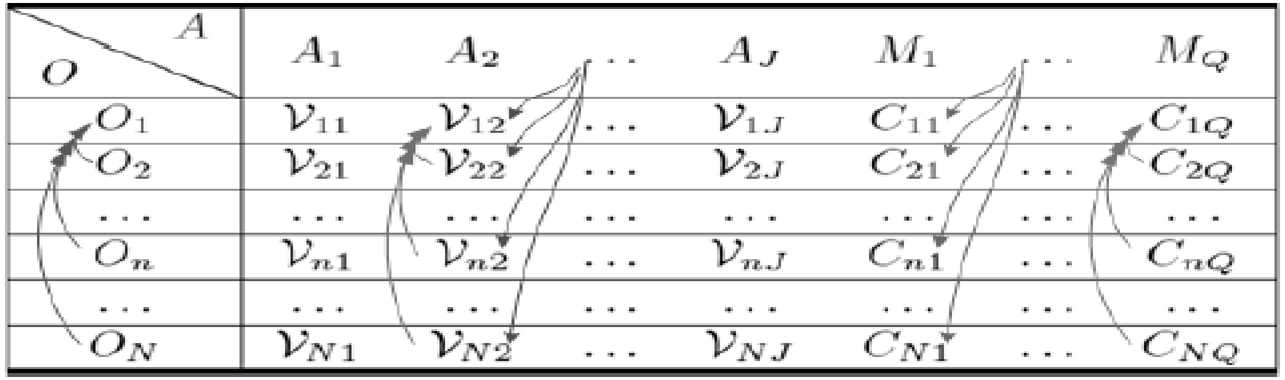


FIGURE 3. Extended information table and non-IIDness learning.

Longbing Cao. Coupling Learning of Complex Interactions, Journal of Information Processing and Management, 51(2): 167-186 (2015).

Pairwise Feature Couplings

- Intra-attribute couplings
  - For example, attribute value occurrence frequency within one attribute
  - how often the values co-occur or how do they depend on each other
- Inter-attribute couplings
  - the interactions between an attribute and other attributes
  - the extent of the value difference brought by other attributes

Hierarchical Coupling Relationships

- U/u: objects
- A/a: attributes, labels, models

$U \backslash A$	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

intra-attribute coupling

inter-attribute coupling

Set Information Functions

Obtain value information: assigns a particular value of attribute a_j to every object.

Obtain value sets: assigns the associated value set of attribute a_j *to the* object set

Obtain object: relates each value of attribute a_j *to the corresponding object set*

$$f = \bigcup_{j=1}^{n} f_j,\ f_j : U \rightarrow V_j (1 \leq j \leq n)$$

$$f_j^*(\{u_{k_1}, \cdots, u_{k_t}\}) = \{f_j(u_{k_1}), \cdots, f_j(u_{k_t})\}, \qquad (3.1)$$

$$g_j(v_j^x) = \{u_i | f_j(u_i) = v_j^x, 1 \leq j \leq n, 1 \leq i \leq m\}, \qquad (3.2)$$

$$g_j^*(V_j') = \{u_i | f_j(u_i) \in V_j', 1 \leq j \leq n, 1 \leq i \leq m\}, \qquad (3.3)$$

where $u_i, u_{k_1}, \cdots, u_{k_t} \in U$, and $V_j' \subseteq V_j$.

Obtain object set: maps the value set of attribute a_j to the dependent object set

Measuring Couplings

$U \backslash A$	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

$$f_2^*(\{u_1, u_2, u_3\}) = \{\mathcal{B}_1, \mathcal{B}_2\}$$

$$g_2(\mathcal{B}_1) = \{u_1, u_2\}$$

$$g_2^*(\{\mathcal{B}_1, \mathcal{B}_2\}) = \{u_1, u_2, u_3, u_6\}$$

Coupled Attribute Value Similarity

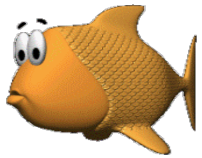
DEFINITION 4.1. *Given an information table S, the Coupled Attribute Value Similarity (CAVS) between attribute values x and y of feature a_j is:*

$$\delta_j^A(x,y) = \delta_j^{Ia}(x,y) \cdot \delta_j^{Ie}(x,y) \tag{4.1}$$

where δ_j^{Ia} and δ_j^{Ie} are IaAVS and IeAVS, respectively.

Intra-attribute couplings: $\delta_j^{Ia}(x,y)$

Inter-attributed couplings: $\delta_j^{Ie}(x,y)$



Intra-attribute (Value) Similarity

DEFINITION 4.2. *Given an information table S, the* ***Intra-coupled Attribute Value Similarity (IaAVS)*** *between attribute values x and y of feature a_j is:*

$$\delta_j^{Ia}(x, y) = \frac{|g_j(x)| \cdot |g_j(y)|}{|g_j(x)| + |g_j(y)| + |g_j(x)| \cdot |g_j(y)|}. \quad (4.2)$$

Rationale:

The Greater similarity is assigned to the pairwise attribute values which own approximately equal frequency.

The higher these frequencies are, the closer such two values are.

IaAVS **has been captured to characterize the value similarity in terms of attribute value occurrence times.**

Measuring Intra-attribute Couplings

$U \backslash A$	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

$$\delta_2^{I_a}(B1, B2) = \frac{|B1| * |B2|}{|B1| + |B2| + |B1| * |B2|} = \frac{2 * 2}{2 + 2 + 2 * 2} = 0.5$$

Inter-attribute Similarity

Modified Value Distance Matrix:

$$D_{j|c}(x,y) = \sum_{g \in L} |P_{c|j}(\{g\}|x) - P_{c|j}(\{g\}|y)|$$

Object Co-occurrence Probability

Inter-attribute coupled Relative Similarity based on Power Set (IRSP), Universal Set (IRSU), Join Set (IRSJ), and Intersection Set (IRSI).

$$\delta^P_{j|k} = \min_{V_k' \subseteq V_k} \{2 - P_{k|j}(V_k'|v_j^x) - P_{k|j}(\overline{V_k'}|v_j^y)\}, \quad (4.5)$$

$$\delta^U_{j|k} = 2 - \sum_{v_k \in V_k} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (4.6)$$

$$\delta^J_{j|k} = 2 - \sum_{v_k \in \bigcup} \max\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (4.7)$$

$$\delta^I_{j|k} = \sum_{v_k \in \bigcap} \min\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (4.8)$$

Inter-attribute Similarity

DEFINITION 4.5. *Given an information table S, the* ***Inter-coupled Attribute Value Similarity (IeAVS)*** *between attribute values x and y of feature a_j is:*

$$\delta_j^{Ie}(x, y) = \sum_{k=1, k \neq j}^{n} \alpha_k \delta_{j|k}(x, y), \tag{4.7}$$

where α_k is the weight parameter for feature a_k, $\sum_{k=1}^{n} \alpha_k = 1$, $\alpha_k \in [0, 1]$, and $\delta_{j|k}(x, y)$ is one of the inter-coupled relative similarity candidates.

IeAVS focuses on the object co-occurrence comparisons with four inter-attribute coupled relative similarity options.

Coupled Attribute Similarity for Values

Definition 5.5 (CASV): The **Coupled Attribute Similarity for Values (CASV)** between attribute values v_j^x and v_j^y of attribute a_j is:

$$\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) = \delta_j^{Ia}(v_j^x, v_j^y) \cdot \delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k \neq j}), \tag{5.10}$$

Coupled Object Similarity

Coupled Object Similarity (COS) between objects:

Definition 7.1 (CASO): Given an information table S, the **Coupled Attribute Similarity for Objects (CASO)** between objects u_x and u_y is $CASO(u_x, u_y)$:

$$CASO(u_x, u_y) = \sum_{j=1}^{n} \delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^{n}), \tag{7.1}$$

Multi-kernel learning of hierarchical, heterogeneous multiple couplings:

Chengzhang Zhu, Longbing Cao, Qiang Liu, Jianpin Yin and Vipin Kumar. Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings. IEEE Transactions on Knowledge and Data Engineering, DOI: 10.1109/TKDE.2018.2791525, 2018

Examples: Measuring Hierarchical Couplings

TABLE 4
Example of Computing Similarity Using *IRSP*

V_1'	$\overline{V_1'}$	$P_{1\|2}(V_1'\|\mathcal{B}_1)$	$P_{1\|2}(\overline{V_1'}\|\mathcal{B}_2)$	$2 - P_{1\|2}(V_1'\|\mathcal{B}_1) - P_{1\|2}(\overline{V_1'}\|\mathcal{B}_2)$
$\varnothing$	$\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$	0	1	1
$\{\mathcal{A}_1\}$	$\{\mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$	0.5	1	0.5
...	...	...	...	...
$\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$	$\varnothing$	1	0	1

TABLE 5
Computing Similarity Using *IRSU*

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	max
$\mathcal{A}_1$	0.5	0	0.5
$\mathcal{A}_2$	0.5	0.5	0.5
$\mathcal{A}_3$	0	0	0
$\mathcal{A}_4$	0	0.5	0.5

U \\ A	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

TABLE 6
Computing Similarity Using *IRSJ*

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	max
$\mathcal{A}_1$	0.5	0	0.5
$\mathcal{A}_2$	0.5	0.5	0.5
$\mathcal{A}_4$	0	0.5	0.5

$$CASO(u_2, u_3) = \sum_{j=1}^{3} \delta_j^A(v_j^2, v_j^3, \{V_k\}_{k=1}^3) = 0.5 + 0.125 + 0.125 = 0.75.$$

TABLE 7
Computing Similarity Using *IRSI*

v_k	$P_{1\|2}(\{v_k\}\|\mathcal{B}_1)$	$P_{1\|2}(\{v_k\}\|\mathcal{B}_2)$	min
$\mathcal{A}_2$	0.5	0.5	0.5

Algorithm 1: Coupled Attribute Similarity for Objects

Data: Data set $S_{m \times n}$ with m objects and n attributes, object $u_x, u_y (x, y \in [1, m])$, and weight $\alpha = (\alpha_k)_{1 \times n}$.

Result: Coupled Similarity for objects $CASO(u_x, u_y)$.

begin

  `// Compute pairwise similarity for any two values of the same attribute.`

  for *attribute* $a_j, j = 1 : n$ **do**

    for *every value pair* $(v_j^x, v_j^y \in [1, |V_j|])$ **do**

      $U_1 \longleftarrow \{i | v_j^i == v_j^x\}$, $U_2 \longleftarrow \{i | v_j^i == v_j^y\}$;

      `// Compute intra-coupled similarity for two values` v_j^x `and` v_j^y.

      $\delta_j^{Ia}(v_j^x, v_j^y) = (|U_1| + |U_2|)/(|U_1||U_2|)$;

      `// Compute coupled similarity for two attribute values` v_j^x `and` v_j^y.

      $\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) \longleftarrow$
      $\delta_j^{Ia}(v_j^x, v_j^y) \cdot IeASV(v_j^x, v_j^y, \{V_k\}_{k \neq j})$;

  `// Compute coupled similarity between two objects` u_x `and` u_y.

  $CASO(u_x, u_y) \longleftarrow sum(\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n))$;

  end

Function $IeASV(v_j^x, v_j^y, \{V_k\}_{k \neq j})$

begin

  `// Compute inter-coupled similarity for two attribute values` v_j^x `and` v_j^y.

  for *attribute* $(k = 1 : n) \wedge (k \neq j)$ **do**

    $\{v_k^z\}_{z \in U_3} \longleftarrow \{v_k^x\}_{x \in U_1} \bigcap \{v_k^y\}_{y \in U_2}$;

    for *intersection* $z = U_3(1) : U_3(|U_3|)$ **do**

      $U_0 \longleftarrow \{i | v_k^i == v_k^z\}$;

      $ICP_x \longleftarrow |U_0 \bigcap U_1| / |U_1|$;

      $ICP_y \longleftarrow |U_0 \bigcap U_2| / |U_2|$;

      $Min_{(x,y)} \longleftarrow min(ICP_x, ICP_y)$;

    `// Compute IRSI for` v_j^x `and` v_j^y.

    $\delta_{j|k}^I(v_j^x, v_j^y, V_k) = sum(Min_{(x,y)})$;

  $\delta_j^{Ie}(x, y) = sum[\alpha(k) \times \delta_{j|k}^I(v_j^x, v_j^y, V_k)]$;

  return $\delta_j^{Ie}(v_j^x, v_j^y, \{V_k\}_{k \neq j})$;

Experiment and Evaluation

- Several experiments are performed on extensive UCI data sets to show the **effectiveness** and **efficiency.**
  - Coupled Similarity Comparison
    - The goal is to show the obvious superiority of IRSI, compared with the most time-consuming one IRSP.
  - COS Application (COD)
    - Four groups of experiments are conducted on the same data sets by k-modes (KM) with ADD (existing methods), KM with COD, spectral clustering (SC) with ADD, and SC with COD.

Different Similarity Metrics

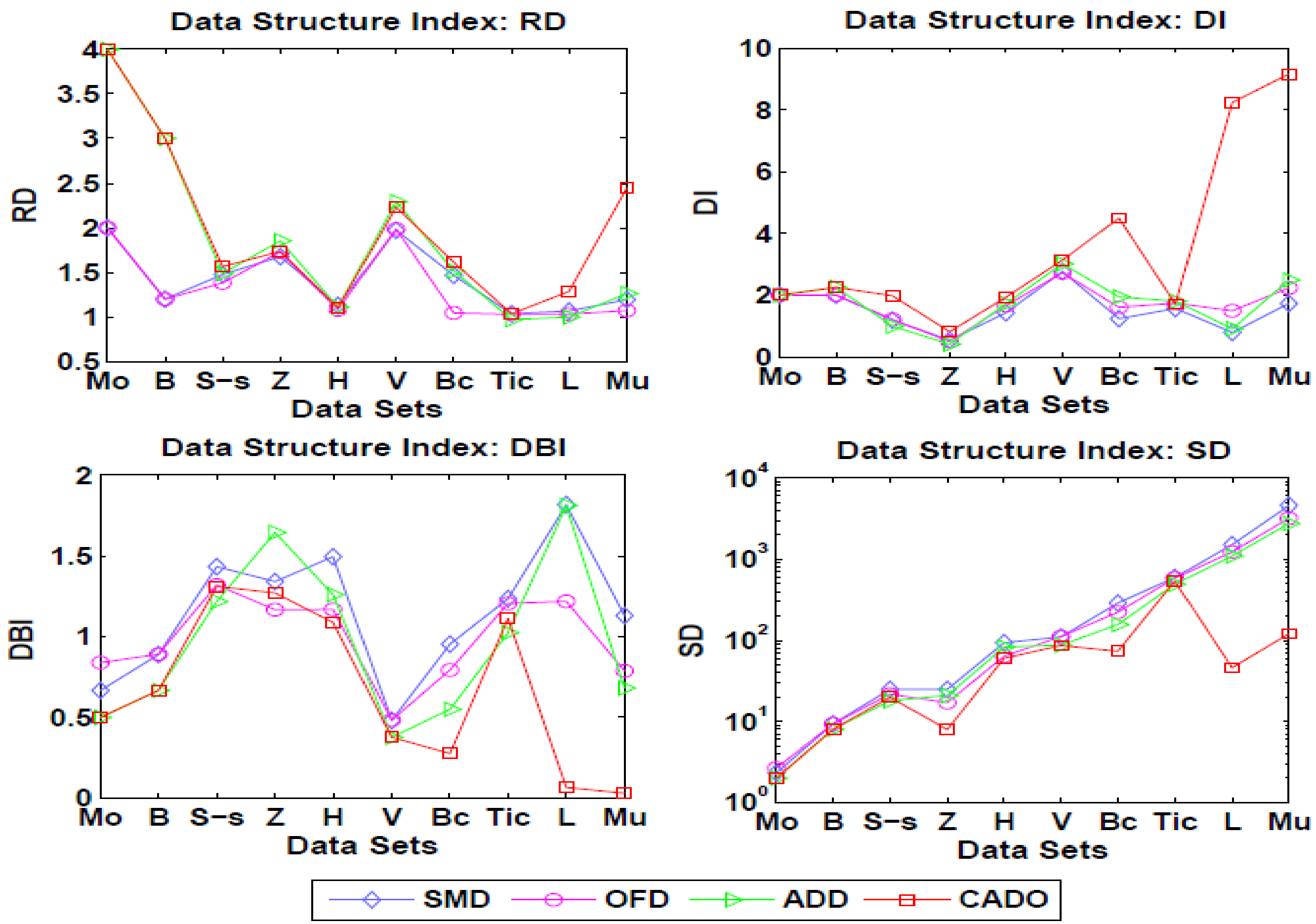


Fig. 3. Data structure index comparison.

Clustering performance indicator:

- Increasing
  - Relative Dissimilarity (RD)
  - Dunn Index (DI) [21]
- Decreasing:
  - Davies-Bouldin Index (DBI) [20],
  - Sum-Dissimilarity (SD)

Applications – Clustering Performance

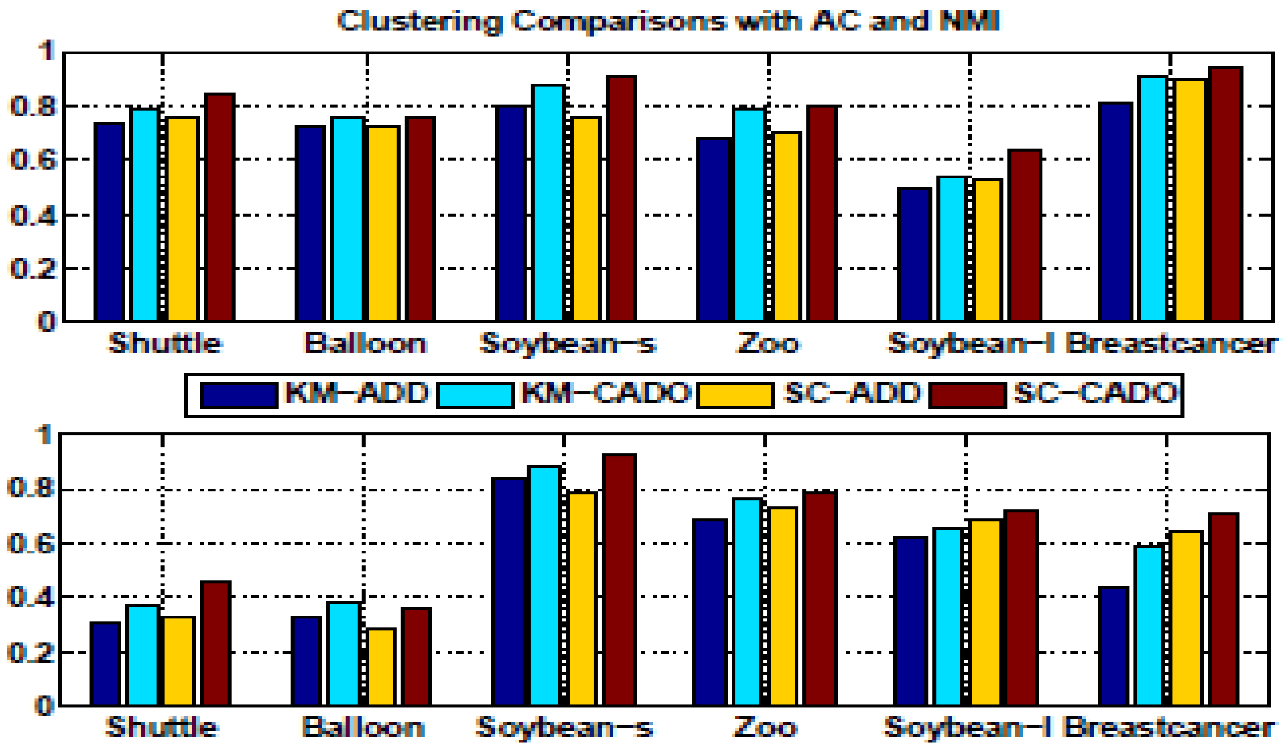


Fig. 4. Clustering evaluation on six data sets.

- k-modes (KM) with ADD (existing methods),
- KM with COS,
- spectral clustering (SC) with ADD
- SC with COS

Non-IID Metric Learning

Chengzhang Zhu, Longbing Cao, Qiang Liu, Jianpin Yin and Vipin Kumar. Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings. IEEE Transactions on Knowledge and Data Engineering, DOI: 10.1109/TKDE.2018.2791525, 2018

Motivation

Name	Gender	Performance	Commitment	Class
John	M	A	H	c1
Mary	F	B	H	c1
Sarah	F	B	I	c1
David	M	C	L	c1
Alice	F	C	I	c2
Edward	M	D	L	c2

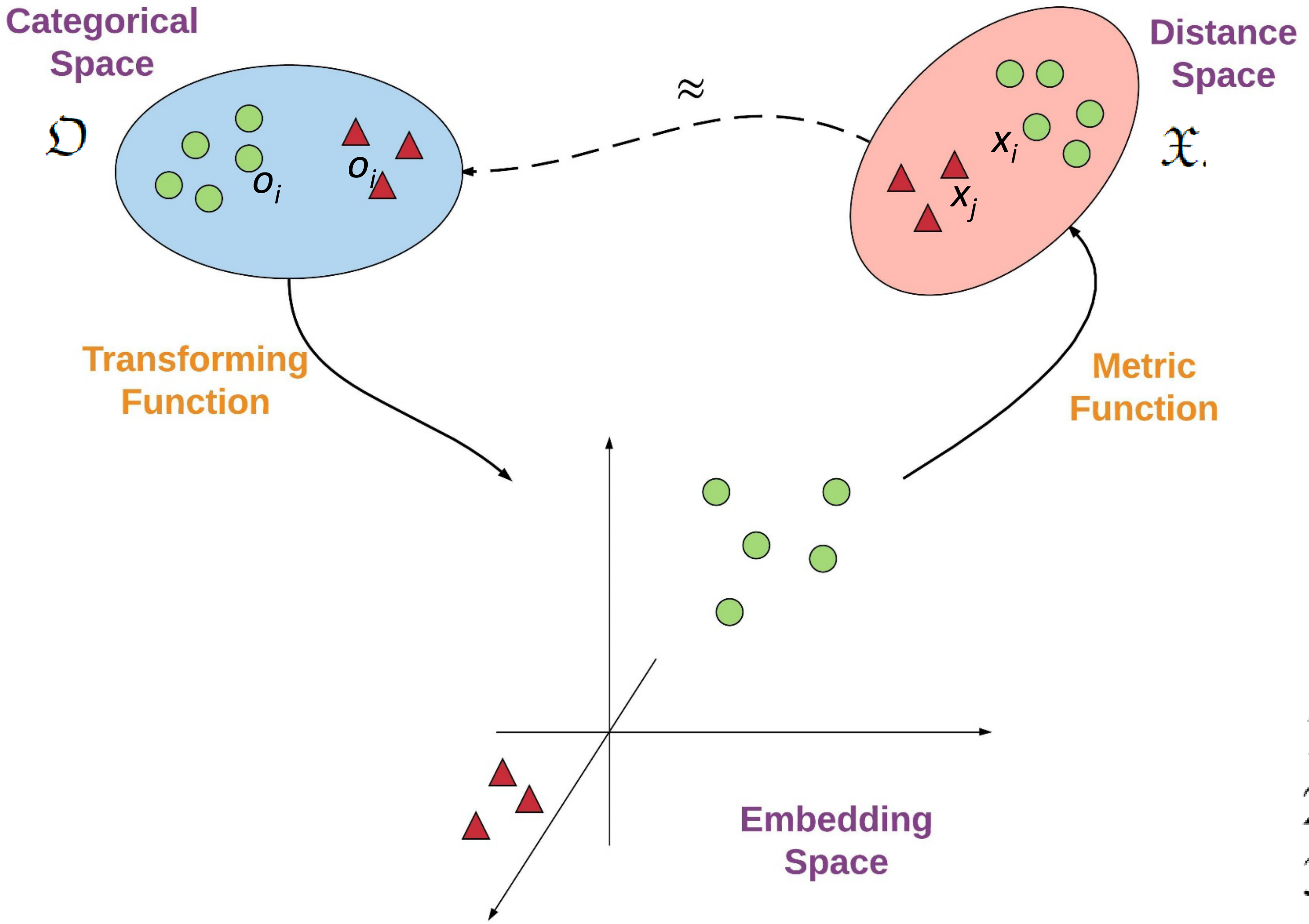
Hamming distance: Dis(H,I) = Dis(H,L) = 1

High (H) level commitment is closer to intermediate (I) instead of low (L) level.

Frequency-based distance: Dis(H, I) = 0

H commitment is different from I.

Problem Statement

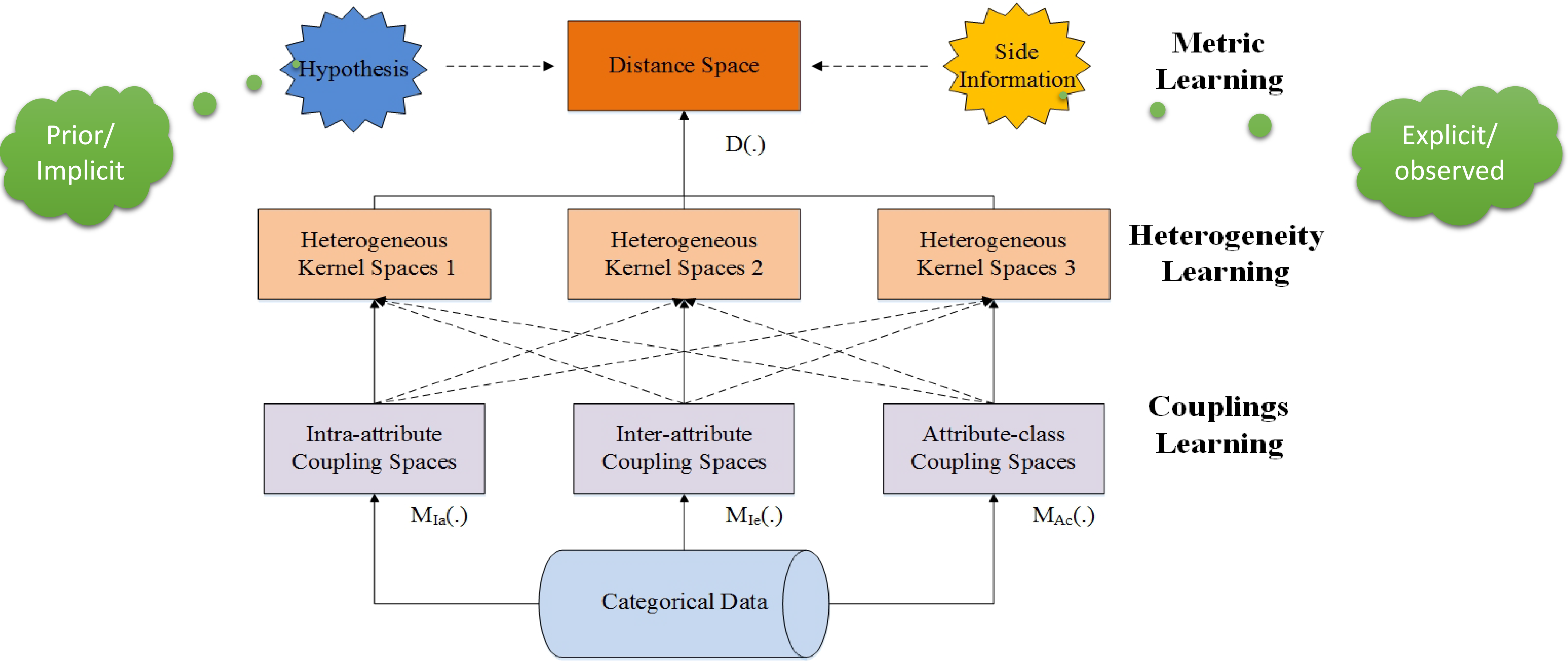


$$
\begin{aligned}
\underset{\mathbf{x}}{\text{minimize}} \quad & \widetilde{Div}(\mathfrak{O}||\mathfrak{X}) \\
\text{subject to} \quad & \mathsf{o} \sim \mathfrak{O} \\
& \mathbf{x} \sim \mathfrak{X} \\
& d(\mathsf{o}_i, \mathsf{o}_j) = \mathbf{x}_i \odot \mathbf{x}_j .
\end{aligned}
$$

Distance metric d(., .) satisfies:

1) $d(\mathsf{o}_i, \mathsf{o}_j) + d(\mathsf{o}_j, \mathsf{o}_k) \geq d(\mathsf{o}_i, \mathsf{o}_k)$,
2) $d(\mathsf{o}_i, \mathsf{o}_j) \geq 0$,
3) $d(\mathsf{o}_i, \mathsf{o}_j) = d(\mathsf{o}_j, \mathsf{o}_i)$.

HELIC Framework



HELIC: Heterogeneous Metric Learning with hIerarchical Couplings

Learning Value-to-Class Couplings

Learning Intra-attribute Couplings

$$m_{Ia}^{(j)}(\mathsf{v}_i^{(j)}) = \frac{|g^{(j)}(\mathsf{v}_i^{(j)})|}{n_o}.$$

Capture value frequency

Learning Inter-attribute Couplings

Capture value co-occurrence

$$m_{Ie}^{(j)}(\mathsf{v}_i^{(j)}) = \left[\ p(\mathsf{v}_i^{(j)}|\mathsf{v}_{*1}), \quad \cdots, \quad p(\mathsf{v}_i^{(j)}|\mathsf{v}_{*|V_*|})\ \right]^\top$$

Learning Attribute-class Couplings

Capture value distribution in each class

$$m_{Ac}^{(j)}(\mathsf{v}_i^{(j)}) = \left[\ p(\mathsf{v}_i^{(j)}|c_1) \quad \cdots \quad p(\mathsf{v}_i^{(j)}|c_{n_c})\ \right]^\top$$

Heterogeneity Learning

Construct Kernel Space:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{m}_1, \mathbf{m}_1) & k(\mathbf{m}_1, \mathbf{m}_2) & \cdots & k(\mathbf{m}_1, \mathbf{m}_{n_v^{(j)}}) \\ k(\mathbf{m}_2, \mathbf{m}_1) & k(\mathbf{m}_2, \mathbf{m}_2) & \cdots & k(\mathbf{m}_2, \mathbf{m}_{n_v^{(j)}}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_1) & k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_2) & \cdots & k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_{n_v^{(j)}}) \end{bmatrix}$$

Using various kernel functions for the value-to-class coupling spaces, a set of kernel matrices $\{\mathbf{K_1}, \cdots, \mathbf{K_{n_k}}\}$ can be obtained. Further, a set of transformation matrices $\{\mathbf{T_1}, \cdots, \mathbf{T_{n_k}}\}$ can be learned to guarantee that the space of the p-th transformed kernel $\mathbf{K}'_p$ only contains the p-th kernel sensitive information, where $\mathbf{K}'_p$ is defined as:

$$\mathbf{K}'_p = \mathbf{T}_p \cdot \mathbf{K}_p$$

Metric Learning

With a positive semi-definite matrix $\omega_p = \alpha_p \mathbf{T}_p^\top \mathbf{T}_p$, the metric d_{ij} is calculated as :

$$d_{ij} = \sum_{p=1}^{n_k} \mathbf{k}_{p,ij}^\top \boldsymbol{\omega}_p \mathbf{k}_{p,ij}$$

where $\mathbf{k}_{p,ij} = \mathbf{K}_{p,i\cdot} - \mathbf{K}_{p,j\cdot}$

The distance can be represented as

$$\boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{\omega}_1^{\text{diag}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\omega}_2^{\text{diag}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\omega}_{n_k}^{\text{diag}} \end{bmatrix}$$

$$d_{ij} = \sum_{p=1}^{n_k} \mathbf{k}_{p,ij}^\top \boldsymbol{\omega}_p \mathbf{k}_{p,ij}$$

$$\mathbf{k}_{ij} = \begin{bmatrix} \mathbf{k}_{1,ij}^\top & \mathbf{k}_{2,ij}^\top & \cdots & \mathbf{k}_{n_k,ij}^\top \end{bmatrix}^\top$$

Metric Learning

Objective function:

Selecting the kernels for their sensitive data distribution

$$
\begin{aligned}
\underset{\boldsymbol{\omega}, b}{\text{minimize}} \quad & \frac{1}{n_o^2} \sum_{i,j \in N_o} \xi_{ij} + \underline{\lambda \|\boldsymbol{\omega}\|}_1 \\
\text{subject to} \quad & \boldsymbol{\omega} \succcurlyeq 0, \\
& \boldsymbol{\omega}_{kl} = 0 \quad for \quad k \neq l, \\
& \underline{1 + r_{ij}(\mathbf{k}_{ij}^\top \boldsymbol{\omega} \mathbf{k}_{ij} - b) \leqslant \xi_{ij}} \\
& \xi_{ij} \geqslant 0, \forall i, j \in N_o.
\end{aligned}
$$

Force the distance between objects from different classes larger than a margin

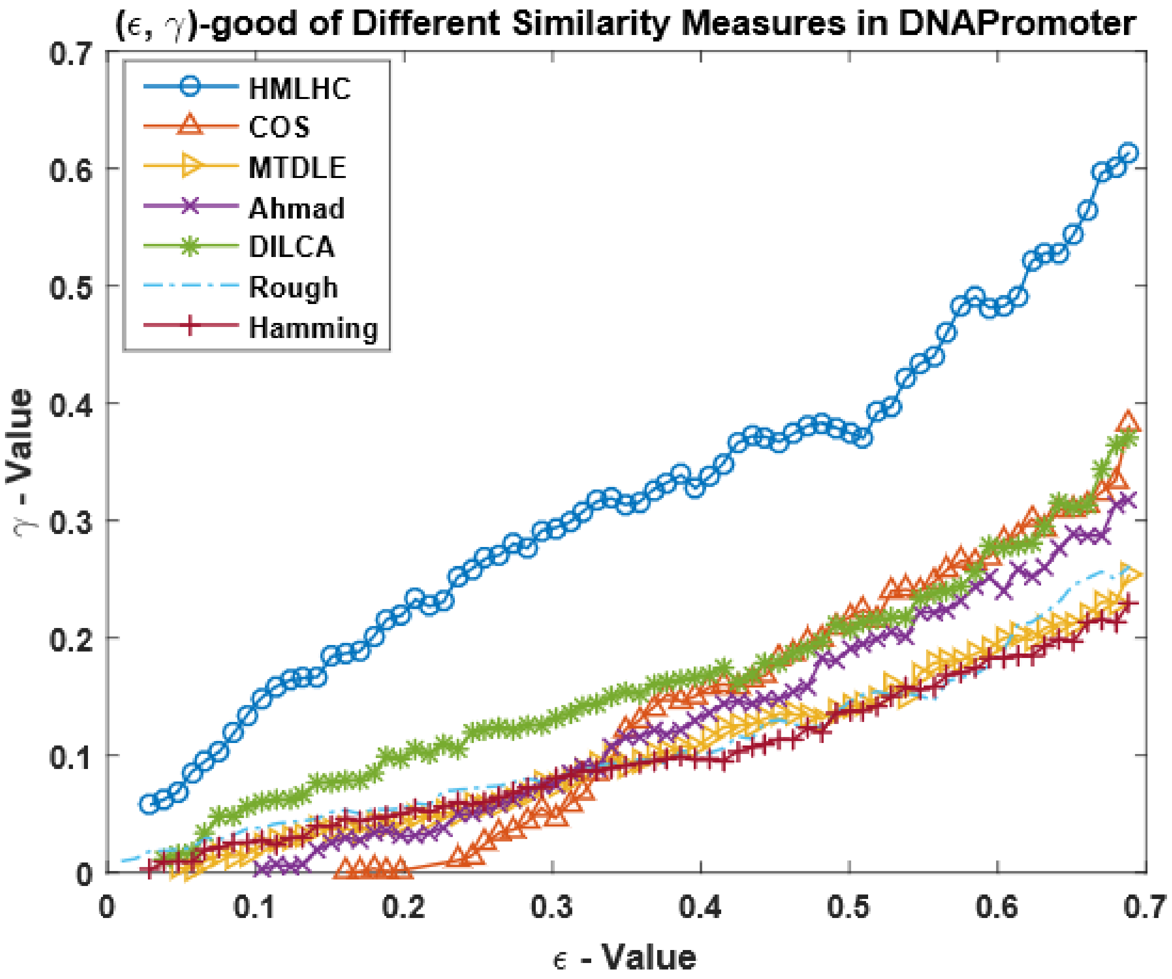
$$
r_{ij} = \begin{cases} 1, & c(\mathrm{o}_i) = c(\mathrm{o}_j) \\ -1, & c(\mathrm{o}_i) \neq c(\mathrm{o}_j) \end{cases}
$$

Representation Performance of HELIC

KNN Classification F-score (%) with Different Distance Measures

Data	HELIC	COS	MTDLE	Ahmad	DILCA	Rough	Hamming	Δ%
Zoo	**100***	**100***	**100***	**100***	**100***	97.75±11.11	**100***	0.00%
DNAPromoter	**92.90±5.85***	75.89±13.35	81.67±10.19	79.98±9.14	90.33±10.31	81.16±10.30	78.05±12.00	2.85%
Hayesroth	**90.85±5.07***	79.64±9.71	68.54±10.55	52.26±10.20	54.60±12.58	81.50±8.59	61.73±12.40	11.47%
Audiology	**75.44±7.60***	41.51±7.20	36.70±7.50	54.29±8.96	64.83±8.04	36.37±7.60	58.55±10.30	16.36%
Housevotes	**96.65 ± 3.40**	94.28 ± 4.95	91.09 ± 5.55	95.81 ± 4.15	94.90 ± 4.14	91.59 ± 5.14	93.77 ± 5.30	0.88%
Spect	**53.09 ±10.35***	51.31±9.16*	52.94±9.48*	52.70±9.69*	51.11±8.97*	51.18±7.90*	51.98±8.85*	0.28%
Mofn3710	**94.39 ±5.86***	79.35±9.07	68.74±10.58	79.35±9.07	71.21±8.42	77.70±11.44	74.82±8.08	18.95%
Monks3	**100***	34.85±0.00	99.88±0.52*	34.85±0.00	34.85±0.00	**100***	92.06±5.24	0.00%
ThreeOf9	**91.01 ±2.93***	32.00±0.00	75.88±8.41	32.00±0.00	32.00±0.00	78.84±5.09	78.84±5.09	15.44%
Balance	**58.91 ±1.31***	21.25±0.00	41.80±5.82	21.25±0.00	21.25±0.00	39.32±4.25	39.32±4.25	40.93%
Crx	**83.26±5.68***	78.58±4.74	77.54±5.68	82.79 ±3.86*	81.02±4.08	77.63±5.12	78.28±4.87	0.57%
Mammographic	**79.61 ±4.59***	70.22±7.12*	70.14±7.10*	70.20±7.02*	70.22±7.81*	69.79±7.11 *	69.95±7.29*	13.37%
Flare	**59.88 ± 3.36***	57.01 ± 4.38*	57.11 ± 3.09	54.41 ± 3.39	55.61 ± 3.13	55.88 ± 4.38	54.98 ± 4.00	4.85%
Titanic	**23.33 ± 2.48***	10.54 ± 1.76	10.06 ± 0.62	10.06 ± 0.99	10.54 ± 1.76	10.54 ± 1.76	10.54 ± 1.76	32.48 %
DNAnominal	**93.12 ± 1.05***	77.52 ± 1.21	52.22 ± 0.00	80.33 ± 1.48	91.65 ± 1.39	81.46 ± 1.75	69.11 ± 1.45	1.60 %
Splice	**93.69 ± 1.11***	77.25 ± 2.19	24.45 ± 0.00	79.85 ± 2.07	84.96 ± 2.21	81.05 ± 1.81	69.29 ± 2.24	10.28 %
Krvskp	**96.98 ± 1.06***	91.77 ± 1.66	90.04 ± 1.65	92.46 ± 1.74	91.39 ± 2.05	89.00 ± 1.43	91.48 ± 1.68	4.89%
Led24	**63.37 ± 1.94***	62.11 ± 1.85*	41.35 ± 2.74	61.81 ± 1.98*	62.58 ± 1.85*	47.89 ± 2.37	41.57 ± 2.19	1.26 %
Mushroom	**100 ± 0.00***	99.98 ± 0.06*	**100 ± 0.00***	**100 ± 0.00 ***	**100 ± 0.00***	**100 ± 0.00 ***	**100 ± 0.00***	0.00%
Krkopt	**53.62 ± 1.71***	52.66 ± 0.78*	NA	52.50 ± 0.96*	52.57 ± 1.02*	39.05 ± 0.70	10.42 ± 0.10	1.82%
Adult	**84.91 ± 0.86***	68.13 ± 1.12	NA	68.20 ± 1.07	68.16 ± 1.14	67.76 ± 1.04	68.01 ± 1.04	24.50%
Connect4	**56.33 ± 0.78***	48.23 ± 0.73	NA	46.95 ± 0.49	46.65 ± 0.55	53.22 ± 0.73	45.81 ± 0.72	5.84%
Census	**68.93 ± 0.55***	66.88 ± 0.40	NA	67.47 ± 0.43	66.66 ± 0.42	66.96 ± 0.55	67.16 ± 0.37	2.64%
Mean	**78.71***	63.95	65.27	63.89	65.09	68.51	65.47	14.89%

Representation Quality of HELIC



Classification Performance

KNN Classification F-score (%) with Couplings

Dataset	HELIC-KNN	HC-KNN	Δ%
Zoo	100	100	0%
DNAPromoter	92.90±5.85	94.93±7.00	0%
Hayesroth	90.85±5.07	85.89±6.39	5.77%
Audiology	75.44±7.60	54.94±11.85	37.31%
Housevotes	96.65 ± 3.40	95.43 ± 4.46	1.28%
Spect	53.09±10.35	51.40±9.51	3.28%
Mofn3710	94.39±5.86	94.92±3.36	0%
Monks3	100	100	0%
ThreeOf9	91.01±2.93	89.96±2.92	1.17%
Balance	58.91±1.31	59.64±1.46	0%
Crx	83.26±5.68	82.43±4.39	1.01%
Mammographic	79.61±4.59	70.31±7.00	13.23%
Flare	59.88 ± 3.36	55.40 ± 3.93	8.09%
Titanic	23.33 ± 2.48	12.15 ± 1.65	92.02%
DNAnominal	93.12 ± 1.05	91.83 ± 1.64	1.40%
Splice	93.69 ± 1.11	75.88 ± 2.03	23.47%
Krvskp	96.98 ± 1.06	92.49 ± 0.92	4.85%
Led24	63.37 ± 1.94	57.71 ± 2.46	9.81%
Mushroom	100 ± 0.00	100 ± 0.00	0.00%
Krkopt	53.62 ± 1.71	52.44 ± 1.58	2.25%
Adult	84.91 ± 0.86	84.32 ± 0.80	0.70%
Connect4	56.33 ± 0.78	43.07± 0.50	30.79%
Census	68.93 ± 0.55	64.23 ± 0.49	7.32%
Mean	78.71	74.32	5.91%

- HC: only learn the hierarchical couplings.
- HELIC: learn both hierarchical couplings and heterogeneity.

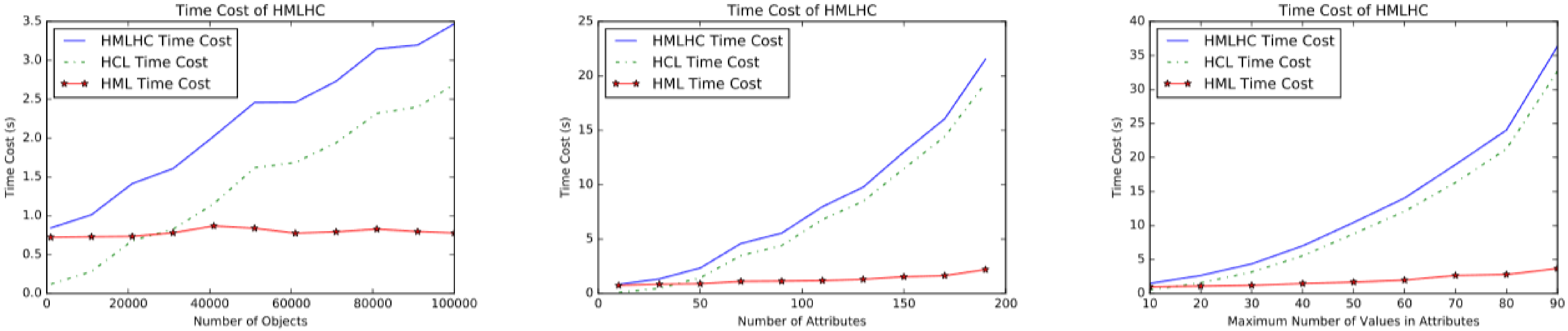
Flexibility of HELIC

LR, RF and SVM Classification F-score (%) with HELIC and MTDLE

Data	HELIC-LR	MTDLE-LR	Δ%	HELIC-RF	MTDLE-RF	Δ%	HELIC-SVM	MTDLE-SVM	Δ%
Zoo	100	92.50 ± 11.75	8.11%	100	99.64 ± 1.63	0.36%	100	100	0%
DNAPromoter	98.48 ± 3.70	89.84 ± 10.89	9.62%	93.88 ± 9.02	74.87 ± 11.89	25.39%	97.98 ± 4.15	89.88±10.35	9.01%
Hayesroth	83.56 ± 6.53	83.23 ± 8.16	0.40%	82.51±7.85	79.80± 10.66	3.40%	84.44 ± 8.62	81.64 ± 8.76	3.43%
Audiology	73.63 ± 6.33	49.88 ± 10.26	47.61%	73.04 ± 7.30	39.23 ± 13.19	86.18%	73.47 ± 6.07	62.15±10.70	18.21%
Spect	69.10±12.68	51.31 ± 8.79	34.67%	69.38±11.94	69.17 ±15.11	3.04%	69.65±12.22	69.33 ± 12.33	0.46%
Mofn3710	100	83.13 ± 16.47	20.29%	81.62±9.03	67.97± 9.94	20.08%	100	100	0%
Monks3	97.21 ± 1.79	100	0%	100	99.88 ± 0.52	0.12%	100	100	0%
ThreeOf9	80.54 ± 5.05	79.52 ± 5.20	1.29%	99.71±0.96	97.14 ± 2.60	2.65%	79.37±5.61	79.46 ± 5.48	0%
Balance	91.24 ± 7.00	63.94 ± 0.06	42.70%	58.52±1.86	58.17 ± 2.24	0.60%	97.45±2.49	98.09 ± 2.44	0%
Crx	85.76 ± 4.86	83.96 ± 4.82	2.14%	85.15±3.72	84.21 ± 4.00	1.12%	84.98±4.79	76.10 ± 5.99	11.67%
Mammographic	82.62 ± 5.13	82.36 ± 4.53	0.32%	82.75±5.36	80.61 ± 4.78	2.65%	82.59±4.32	80.91 ± 5.45	2.08%
Mean	87.96	78.51	12.04%	84.99	77.84	9.19%	88.61	85.91	3.14%

The HELIC framework can be incorporated into different classifiers

Scalability of HELIC



(a) Time Cost v.s. Number of Objects. (b) Time Cost v.s. Number of Attributes. (c) Time Cost v.s. Number of Attribute Values.

The Time Cost of HELIC w.r.t. Data Factors: Object Number n_o, Attribute Number n_a, and Maximum Number of Attribute Values n_{mv}. The solid line refers to the total time cost of HELIC. The dotted line refers to the time cost of the hierarchical coupling learning parts. The star line refers to the time cost of the heterogeneous metric learning parts.

Conclusions

- This work reports an effective heterogeneous metric for learning hierarchical couplings within and between attributes and between attributes and classes in categorical data.
- It analyzes the heterogeneity in the hierarchical interaction spaces and integrating heterogeneous couplings in complex categorical data.
- The proposed method can be applied to a variety of areas with categorical data. One thing in applications is to select appropriate kernels by considering specific data characteristics and domain knowledge of the problems.

Non-IID Representation Learning

Songlei Jian, Liang Hu, Longbing Cao and Kai Lu. Representation Learning with Multiple Lipschitz-constrained Alignments on Partially-labeled Cross-domain Data, AAAI2020

Songlei Jian, Longbing Cao, Guansong Pang, Kai Lu, Hang Gao. Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning. IJCAI2017

Songlei Jian, Liang Hu, Longbing Cao, and Kai Lu. Metric-based Auto-Instructor for Learning Mixed Data Representation. AAAI2018

Metric-based Auto-Instructor for Learning Mixed Data Representation

Songlei Jian, Liang Hu, Longbing Cao and Kai Lu. Metric-based Auto-Instructor for Learning Mixed Data Representation, AAAI2018

Source code is available at: https://github.com/jiansonglei/MAI

Background

- Categorical features
  - e.g., gender, education, brand
- Numerical features
  - e.g., age, length, price
- Mixed data contains both categorical features and numerical features
  - e.g., census data, product information

Representation of Categorical Features

- One-hot encoding:
- Distributional representation
  - Latent semantic analysis
  - Random projection
- Distributed representation
  - Embedding for categorical data
  - Word embedding

Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

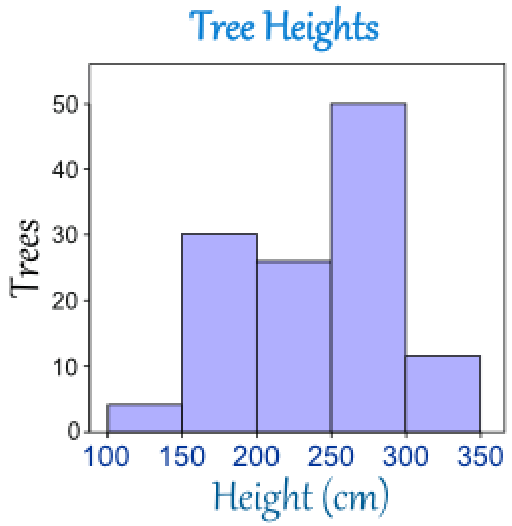
Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Representation of Numerical Features

- Raw representation
- Normalized representation
- Distributed representation
  - Dimension reduction
    - Principal component analysis (PCA)
    - Non-negative Matrix Factorization (NMF)
  - Autoencoder

Name	Formula
Standard score	$\frac{X-\mu}{\sigma}$
Student's t-statistic	$\frac{X-\overline{X}}{s}$
Studentized residual	$\frac{\hat{\epsilon}_i}{\hat{\sigma}_i}=\frac{X_i-\hat{\mu}_i}{\hat{\sigma}_i}$
Standardized moment	$\frac{\mu_k}{\sigma^k}$
Coefficient of variation	$\frac{\sigma}{\mu}$
Feature scaling	$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}$

Representation of Mixed Data



- Transform numerical data to categorical one
  - Discretization
- Transform categorical data to numerical data
  - Statistics: e.g., TF-IDF
- Concatenated representation: treat categorical and numerical features independently

weighting scheme	document term weight	query term weight
1	$f_{t,d} \cdot \log \frac{N}{n_t}$	$\left(0.5 + 0.5 \frac{f_{t,q}}{\max_t f_{t,q}}\right) \cdot \log \frac{N}{n_t}$
2	$1 + \log f_{t,d}$	$\log(1 + \frac{N}{n_t})$
3	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \cdot \log \frac{N}{n_t}$

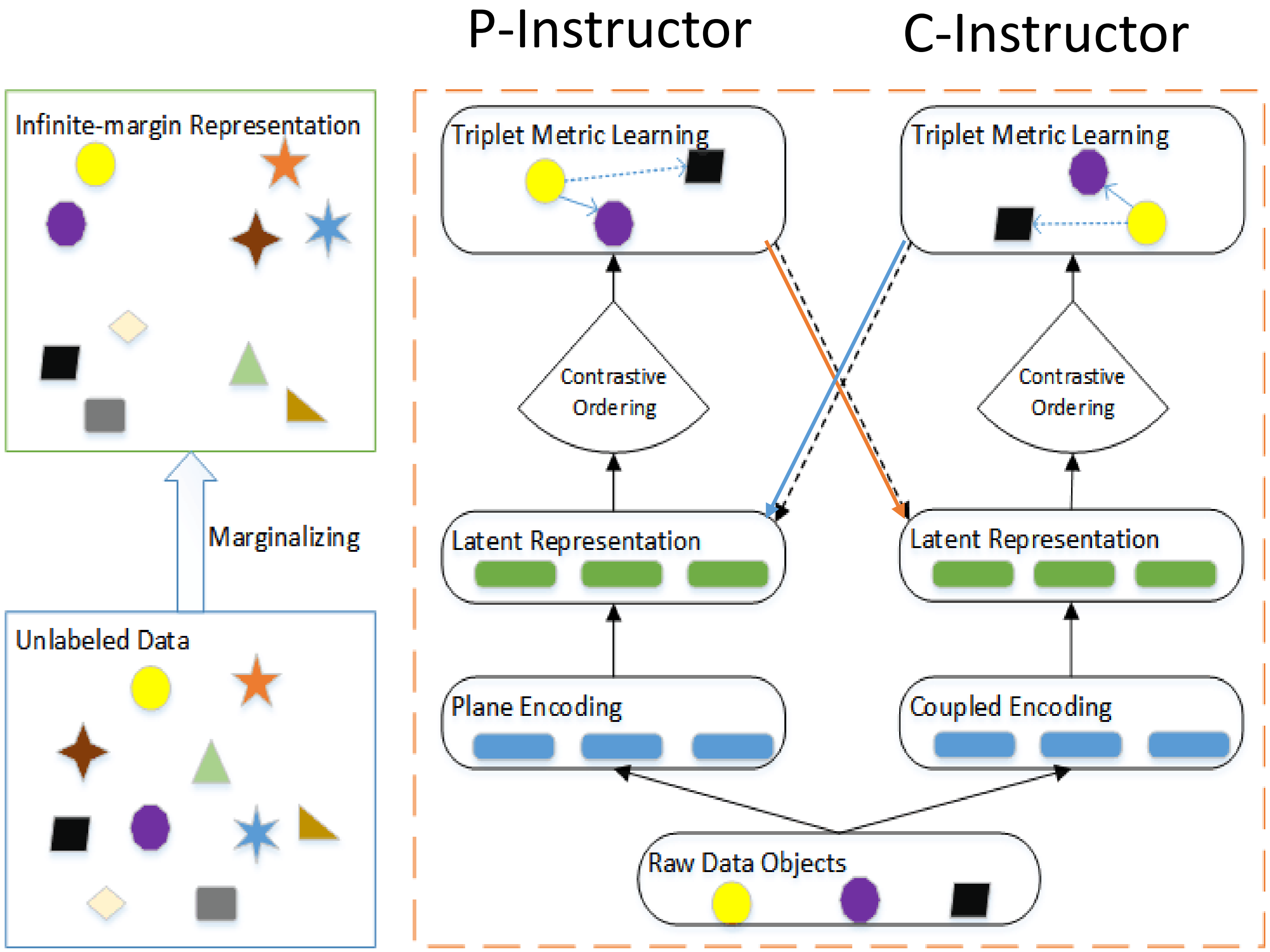
Name	Gender	Height
Alice	Female	1.75 m
Bob	Male	1.75 m

What Is A Good Representation for Mixed Data?

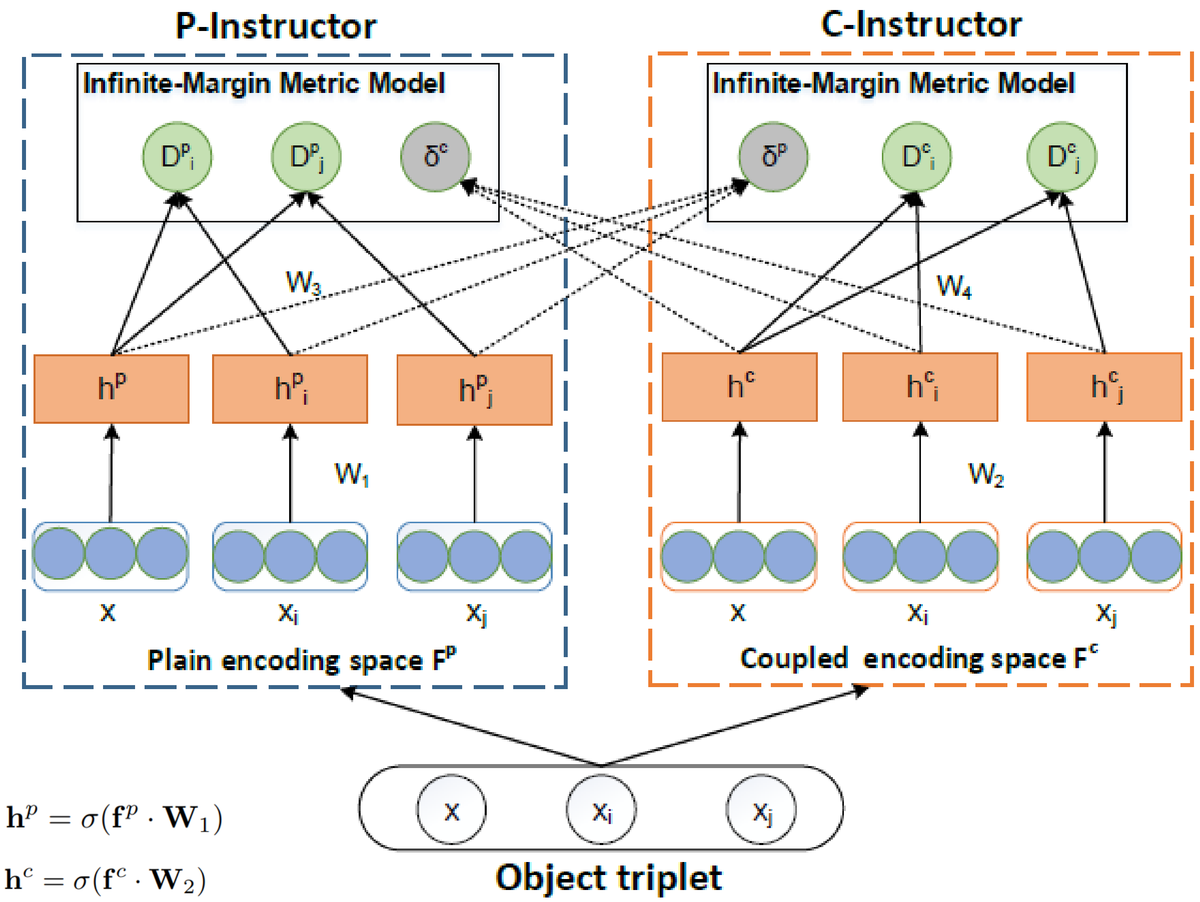
- At the feature level: capture the heterogeneous couplings (e.g., complex interactions, dependencies) between features
  - Couplings between categorical features
  - Couplings between numerical features
  - Couplings between categorical and numerical features
- At the object level: a good representation should express the discrimination and margins between objects to fertilize learning tasks.

MAI Architecture

- Consists of two instructors in two encoding spaces
  - P-Instructor in plain encoding space
  - C-Instructor in coupled encoding space



Coupled Metric Learning Process



$$\mathbf{h}^p = \sigma(\mathbf{f}^p \cdot \mathbf{W}_1)$$

$$\mathbf{h}^c = \sigma(\mathbf{f}^c \cdot \mathbf{W}_2)$$

$$D^p(\mathbf{h}^p, \mathbf{h}^p_i) = (\mathbf{h}^p - \mathbf{h}^p_i)\mathbf{W}_3(\mathbf{h}^p - \mathbf{h}^p_i)^\top$$

$$D^c(\mathbf{h}^c, \mathbf{h}^c_i) = (\mathbf{h}^c - \mathbf{h}^c_i)\mathbf{W}_4(\mathbf{h}^c - \mathbf{h}^c_i)^\top$$

$$\delta_{\mathbf{h}}(\mathbf{h}_i, \mathbf{h}_j) = \begin{cases} 1, \text{ if } d(\mathbf{h}, \mathbf{h}_i) > d(\mathbf{h}, \mathbf{h}_j) \\ 0, \text{ otherwise.} \end{cases}$$

- Plain features: Concatenation of one-hot representation of categorical data and numerical data
- Coupled features: product kernel of numerical variable and categorical value

$$p(a_i^x, v_j) = \frac{1}{N}\sum_{k=1}^{N}\{L_\lambda(v_j^k, v_j)W(\frac{a_i^k - a_i^x}{h_i})\}$$

$$\begin{cases} L_{\Theta^p} = -\sum_{\langle x, x_i, x_j\rangle} \log P_{\Theta^p}(D_i^p > D_j^p | \delta^c_{\mathbf{h}^c}) \\ L_{\Theta^c} = -\sum_{\langle x, x_i, x_j\rangle} \log P_{\Theta^c}(D_i^c > D_j^c | \delta^p_{\mathbf{h}^p}) \end{cases}$$

Experiments

- Application: clustering
  - Partition-based: k-means
  - Density-based: DBSCAN
- Evaluation metrics:
  - AMI
  - Calinski-Harabasz index

Table 1: Statistics of UCI datasets

Datasets	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{F}^c\rvert$	$\lvert\mathcal{F}^n\rvert$	$\lvert Class\rvert$
Echo	132	2	8	3
Hepatitis	155	13	6	2
MPG	398	2	5	6
Heart	270	8	5	2
ACA	690	8	6	2
CRX	690	9	6	2
CMC	1473	7	2	3
Income	32561	8	6	2

Table 2: K-means clustering performance w.r.t. AMI $\pm$ standard deviation. The top two performers for each are boldfaced.

Datasets	Plain encoding	Coupled encoding	CoupledMC	Autoencoder	MAI-F	MAI-D
Echo	0.1789±0.1033	0.1749±0.0444	0.1237±0.1147	0.2493±0.0207	**0.3246±0.0000**	**0.3304±0.0000**
Hepatitis	0.1453±0.0703	0.1761±0.0292	0.1532±0.0342	0.1689±0.0163	**0.1848±0.0000**	**0.1905±0.0000**
MPG	0.1490±0.0106	0.1477±0.0184	0.1373±0.0347	0.1536±0.0086	**0.1831±0.0232**	**0.1770±0.0000**
Heart	**0.3130±0.0688**	0.1439±0.0642	0.1037±0.1215	**0.3302±0.0042**	0.2632±0.0000	0.2774±0.0000
ACA	0.3204±0.1518	0.3433±0.1726	0.3182±0.0627	0.3477±0.0844	**0.4258±0.0000**	**0.4258±0.0000**
CRX	0.2322±0.1191	0.0836±0.1109	0.2714±0.1361	0.1445±0.1477	**0.4267±0.0000**	**0.4267±0.0000**
CMC	0.0293±0.0052	0.0269±0.0013	**0.0333±0.0070**	0.0292±0.0037	**0.0327±0.0077**	0.0303±0.0081
Income	0.1139±0.0361	**0.1414**±0.0291	0.1258±0.0658	0.1314±0.0000	**0.1325±0.0000**	**0.1325±0.0000**
Average	0.1853±0.0707	0.1547±0.0588	0.1583±0.0722	0.1944±0.0353	**0.2467±0.0064**	**0.2488±0.0010**

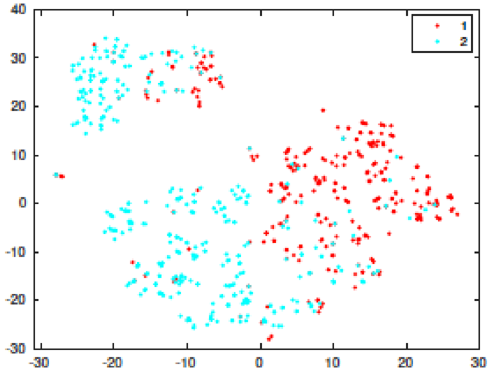
Table 3: DBSCAN clustering performance w.r.t. AMI/Clusters.

Datasets	PF($\|C\|$)	CF($\|C\|$)	CMC($\|C\|$)	AE($\|C\|$)	MAI-F($\|C\|$)
Echo	0.123(5)	0.011(3)	0.067(2)	0.188(7)	**0.392**(3)
Hepatitis	0.019(4)	0.044(2)	0.037(5)	0.016(2)	**0.075**(3)
MPG	0.031(20)	0.037(16)	0.049(13)	0.149(2)	**0.237**(3)
Heart	0.024(4)	0.001(2)	0.003(2)	0.003(2)	**0.130**(3)
ACA	0.003(4)	0.021(7)	0.031(2)	0.087(20)	**0.227**(6)
CRX	0.003(4)	0.018(6)	0.061(2)	0.102(16)	**0.242**(5)
CMC	0.002(21)	0.009(2)	0.115(5)	0.003(13)	**0.043**(2)
Income	**0.157**(493)	0.052(6)	0.052(6)	0.108(291)	0.1304(15)
Average	0.0451	0.0242	0.0519	0.0818	0.1845

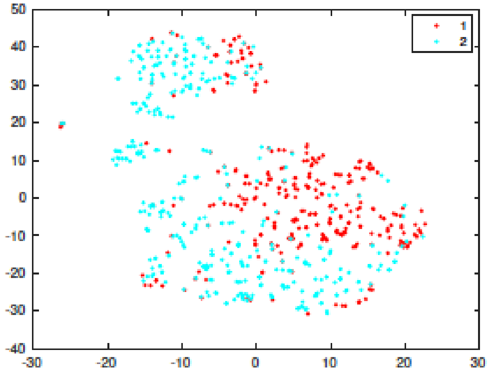
Table 4: Calinski-Harabasz index on representation w.r.t. the Euclidean distance for ground-truth labels

Datasets	PF	CF	CMC	AE	MAI-F
Echo	14.60	7.14	5.12	21.99	**56.81**
Hepatitis	11.76	8.65	15.91	16.05	**44.15**
MPG	19.18	7.34	7.53	41.88	**45.91**
Heart	32.35	16.83	5.64	56.49	**91.85**
ACA	72.90	31.69	16.92	124.37	**288.31**
CRX	67.78	65.94	20.77	106.97	**226.55**
CMC	16.82	12.46	17.21	22.44	**35.35**
Income	1419.90	2029.04	1729.04	3009.80	**5045.45**

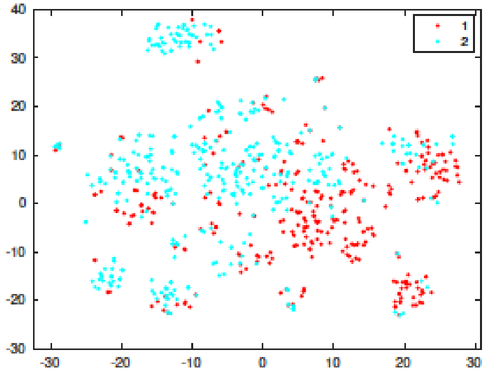
Visualization



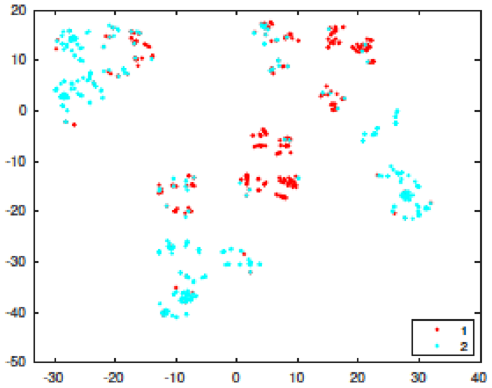
(a) Plain encoding



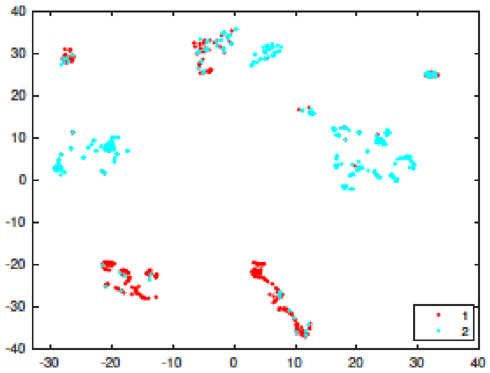
(b) Coupled encoding



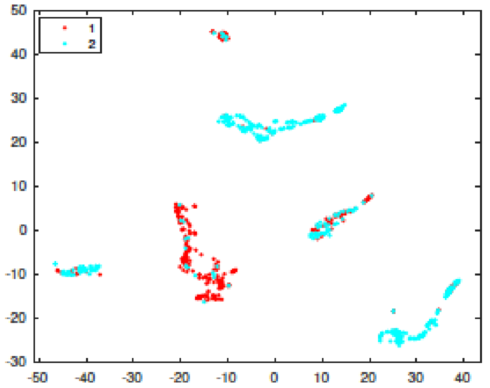
(c) CoupledMC



(d) Autoencoder



(e) MAI-F



(f) MAI-D

Conclusion

- A comprehensive representation for mixed data simultaneously learns the couplings at feature level and the discrimination between objects at the object level.

- A metric-based auto-instructor (MAI) model with two collaborative instructors learns more discriminative representation between objects by learning the margin enhanced distance metric.

- MAI is a general representation learning framework not limited to mixed data, which has the potential to be applied to multimodal learning and domain adaption.

Coupling Learning of complex interactions and relations

Songlei Jian, Liang Hu, Longbing Cao, Kai Lu, Hang Gao. Evolutionarily Learning Multi-aspect Interactions and Influences from Network Structure and Node Content, AAAI2019

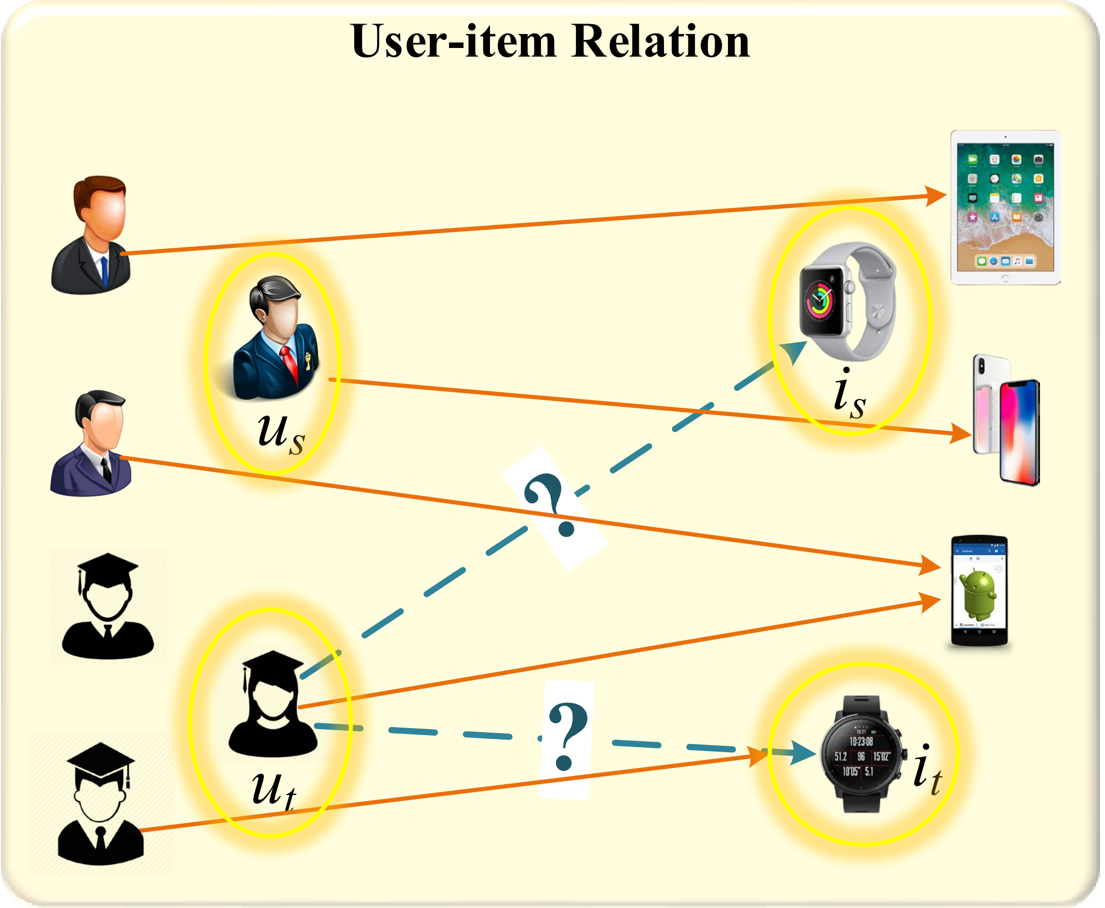
Liang Hu, Songlei Jian, Longbing Cao, Zhiping Gu, Qingkui Chen, Artak Amirbekyan. HERS: Modeling Influential Contexts with Heterogeneous Relations for Sparse and Cold-start Recommendation, AAAI2019.

Learning Heterogeneous Couplings – Multi-relation Learning

Hu, L., Jian, S., Cao, L., Gu, Z., Chen, Q., Amirbekyan, A. HERS: Modeling Influential Contexts with Heterogeneous Relations for Sparse and Cold-start Recommendation. In AAAI-19

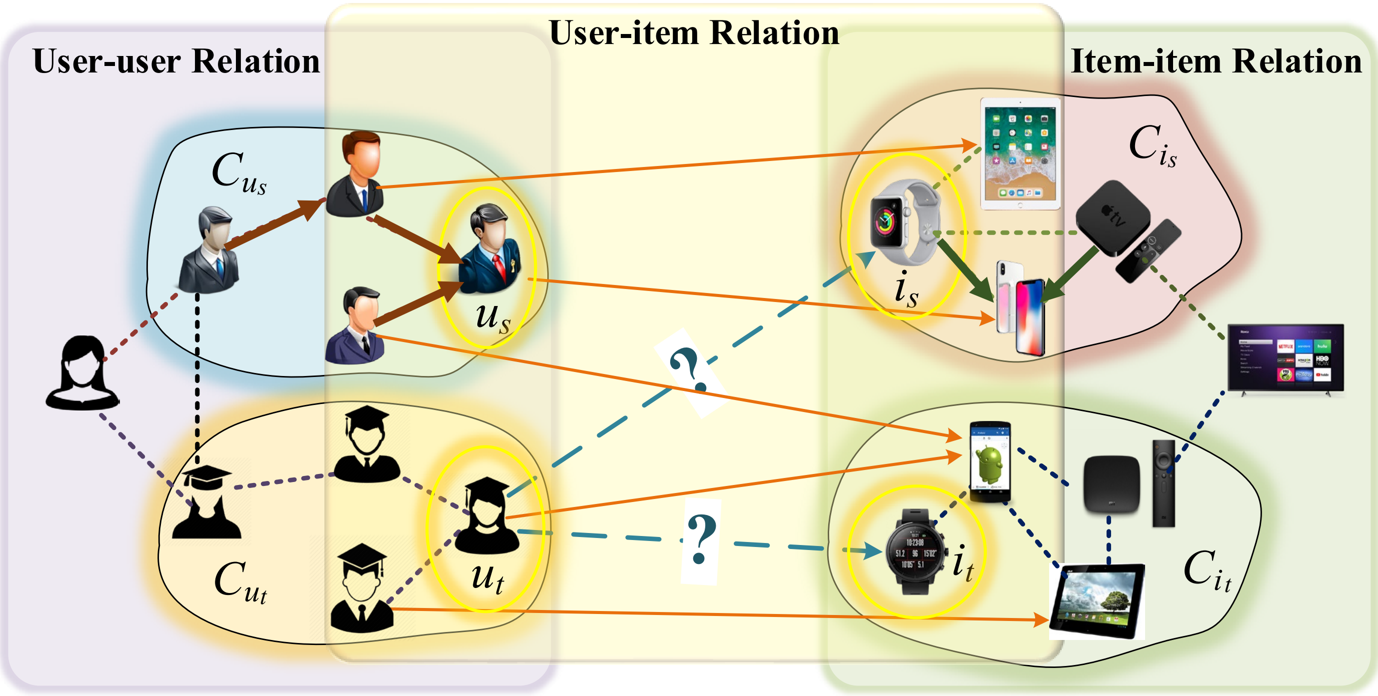
Heterogeneous couplings

- The basic problem in RS is to study the ***user-item*** relation.
- Besides ***user-item*** relation, ***user-user*** relation (e.g. social network) and ***item-item*** relation (e.g. compatibility)
- In fact, ***user-user*** relation and ***item-item*** relation have direct influence on user selection, so they should be considered when modeling RS.



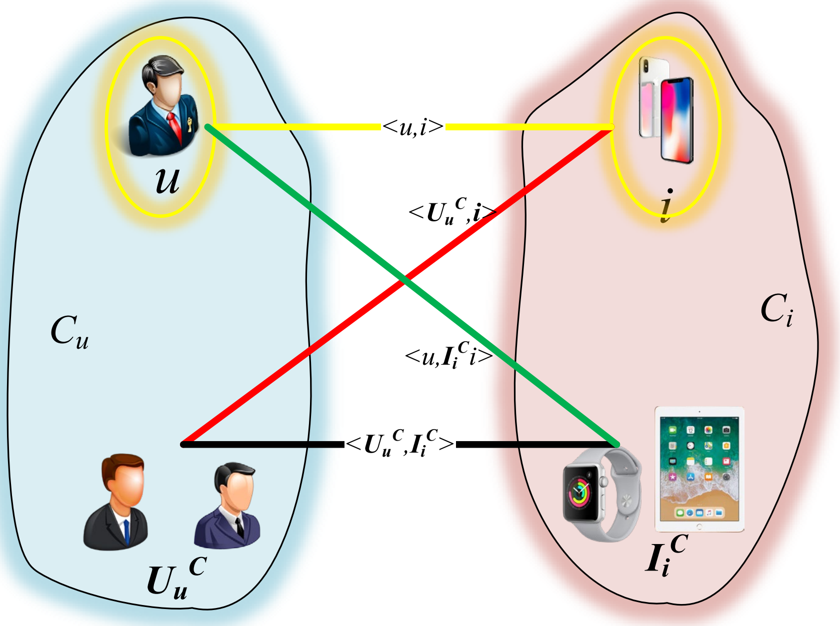
Influence contexts for making decision

- A user u is influenced by friends and friends' friends. C_u signifies the user influential context.
- User selection on an item i is also influenced by relevant items which form item influential context C_i.
- Influential contexts of users and items indicate how a user's choice on items is made, thus making recommendation more accurate and interpretable.



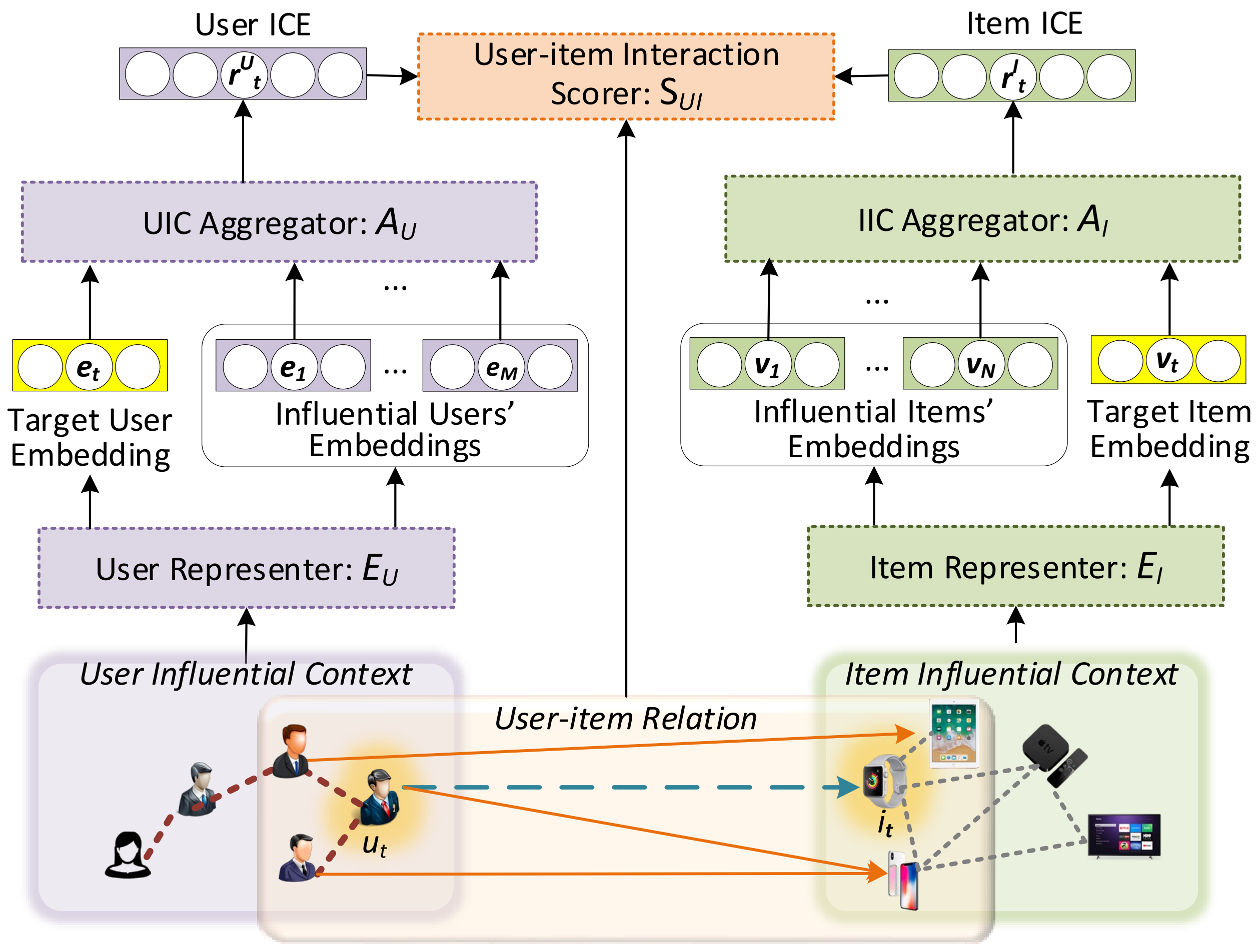
Influential context interaction decomposition

- Coupling Modeling
  - Heterogeneous couplings
  - Influential-context couplings



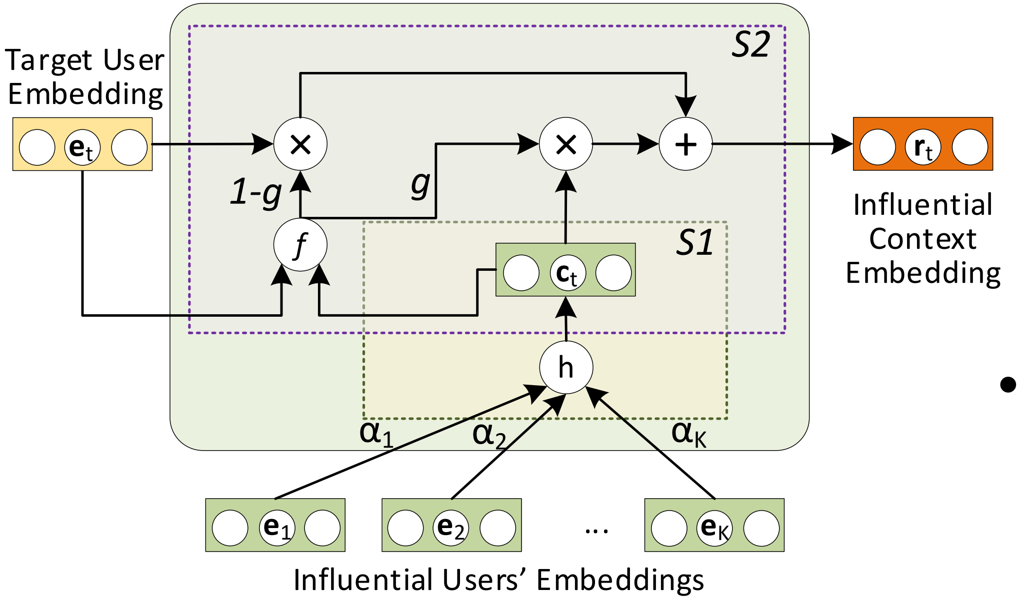
- $s_{\langle C_u, C_i \rangle} = \lambda_1 s_{\langle u,i \rangle} + \lambda_2 s_{\langle u, I_i^c \rangle} + \lambda_3 s_{\langle U_u^c, i \rangle} + \lambda_4 s_{\langle U_u^c, I_i^c \rangle}$
- $s_{\langle C_u, C_i \rangle}$: overall interaction score
- $s_{\langle u,i \rangle}$: scores u's preference on preference on item i
- $s_{\langle u, I_i^c \rangle}$: scores u's preference on influential items I_i^c
- $s_{\langle U_u^c, i \rangle}$: scores relevant users' preference on item i
- $s_{\langle U_u^c, I_i^c \rangle}$: scores the subsidiary preference between influential users U_u^c and influential items I_i^c

Architecture of modeling HERS



- User Representer E_U: it maps target user u_t and its influential users in UIC to the corresponding user embeddings, i.e., $E_U(\mathcal{U}_{u_t}) \mapsto \mathcal{E}_{u_t}$ where $\mathcal{E}_{u_t} = \{\mathbf{e}_t, \mathbf{e}_1, \cdots \mathbf{e}_M\}$.
- Item Representer E_I: it maps target item i_t and its influential items in IIC to the corresponding item embeddings, i.e., $E_I(\mathcal{I}_{i_t}) \mapsto \mathcal{E}_{i_t}$ where $\mathcal{E}_{i_t} = \{\mathbf{v}_t, \mathbf{v}_1, \cdots \mathbf{v}_N\}$.
- UIC Aggregator A_U: it learns a representation $\mathbf{r}^U_t$ for the influential context $\mathcal{C}_{u_t}$, namely influential context embedding (ICE). Formally, we have $A_U(\mathcal{C}_{u_t}, \mathcal{E}_{u_t}) \mapsto \mathbf{r}^U_t$.
- IIC Aggregator A_I: it learns i_t's ICE by aggregating the influential context $\mathcal{C}_{i_t}$, that is, $A_I(\mathcal{C}_{i_t}, \mathcal{E}_{i_t}) \mapsto \mathbf{r}^I_t$.
- User-item Interaction Scorer S_{UI}: it learns to score the interaction strength between the target user-item pair $\langle u_t, i_t \rangle$ in terms of the user ICE $\mathbf{r}^U_t$ and the item ICE $\mathbf{r}^I_t$, namely $S_{UI}(\mathbf{r}^U_t, \mathbf{r}^I_t, y_{u_t,i_t}) \mapsto s_{\langle \mathcal{C}_u, \mathcal{C}_i \rangle}$ (cf. Eq. $\boxed{1}$).

Influential-Context Aggregation Unit (ICAU)



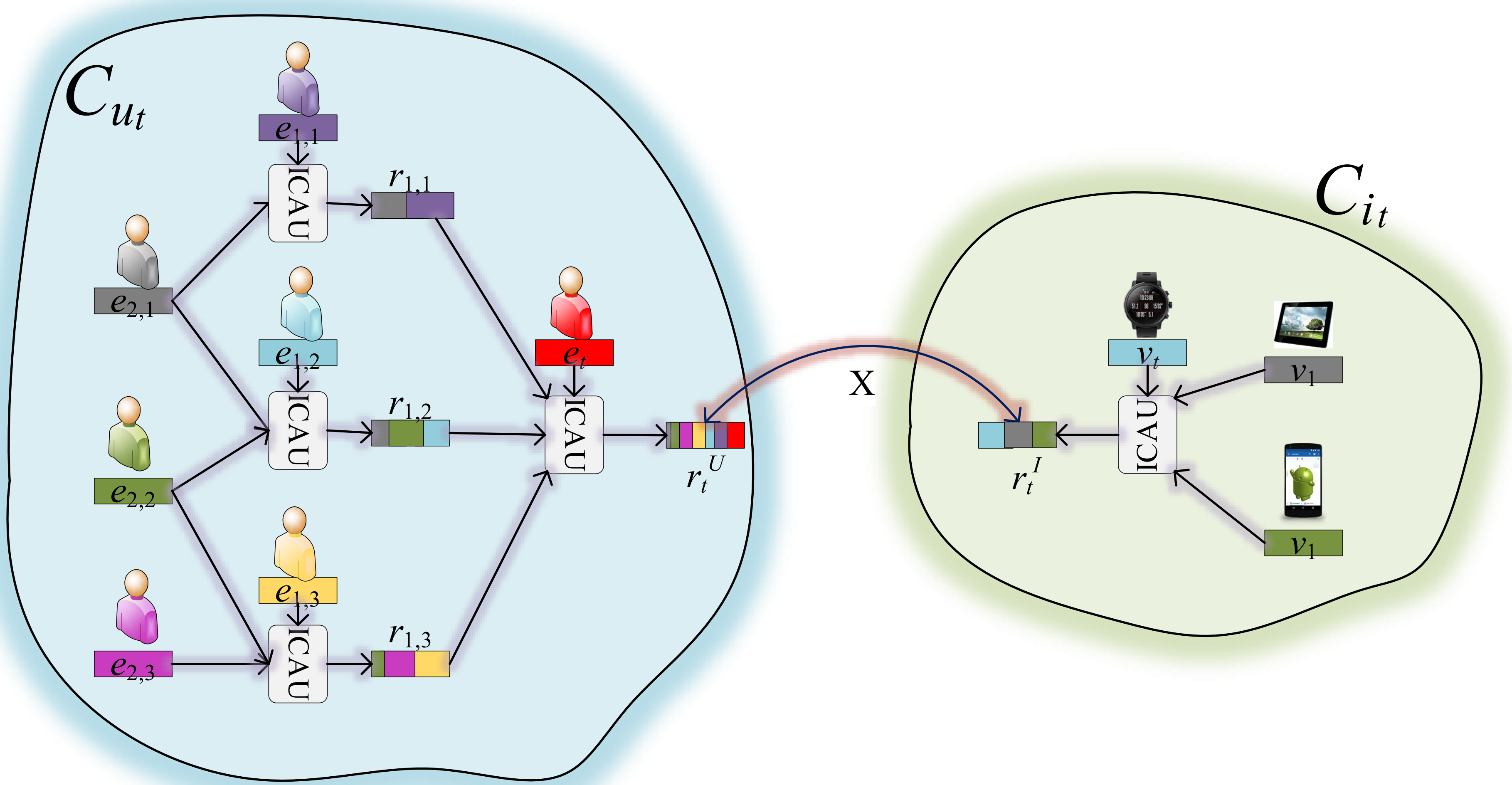
- S1: This stage outputs the subsidiary influence embedding $\boldsymbol{c}_t$ through an aggregation function $h(\cdot)$ over the influential users' embeddings $\boldsymbol{e}_k$:

$$\{\alpha_1, \cdots, \alpha_K\} = a(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K)$$
$$\boldsymbol{c}_t = h(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K | \alpha_1, \cdots, \alpha_K)$$

- S2: This stage generates the ICE by aggregating the subsidiary influence context embedding $\boldsymbol{c}_t$ and the target embedding $\boldsymbol{e}_t$ through a gate function $f(\cdot)$:

$$g = f(\boldsymbol{c}_t, \boldsymbol{e}_t)$$
$$\boldsymbol{r}_t = g\boldsymbol{c}_t + (1 - g)\boldsymbol{e}_t$$

ICE is a representation for influential coupling



Statistics of datasets: Delicious and Lastfm

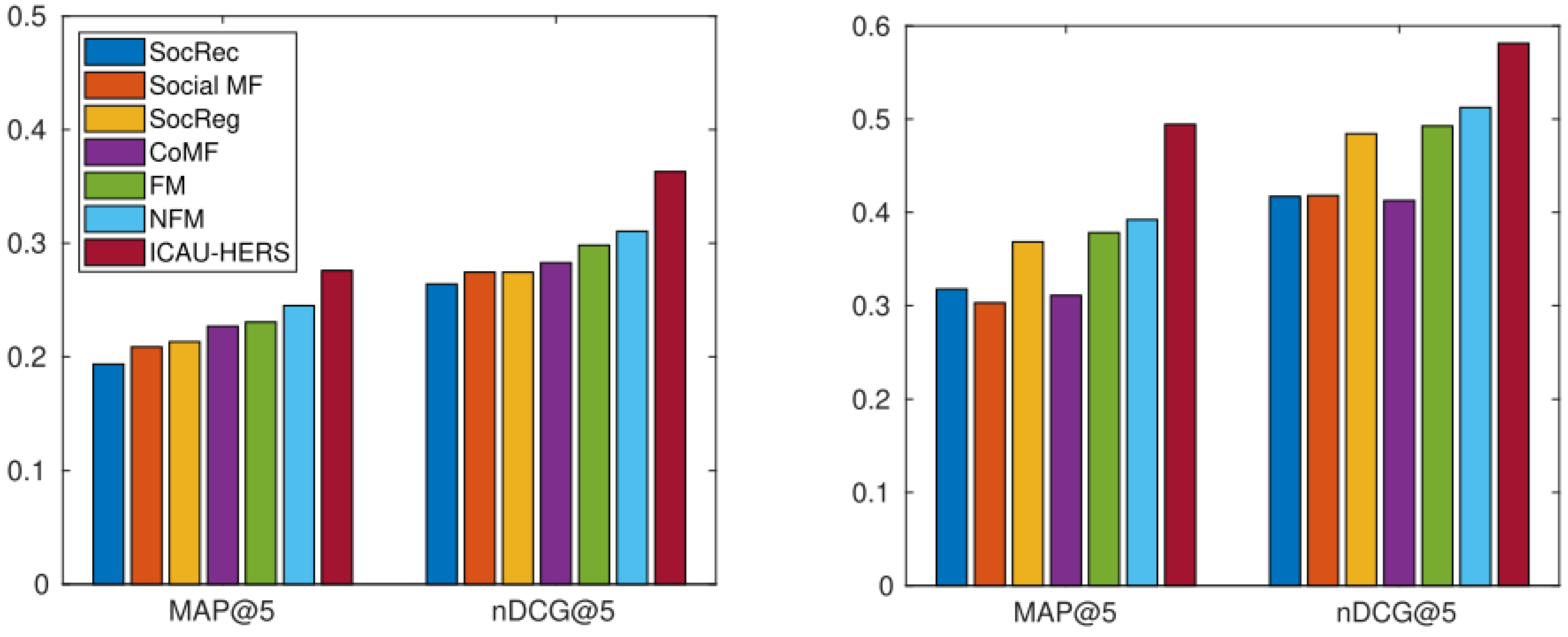
- Two datasets, ***Delicious*** and ***Lastfm*** provided by RecSys Challenge 2011

	Property	User-user	Item-item	User-Item
Delicous	#Entity	1,892	17,632	1,892+17,632
	#Link	25,434	199,827	104,799
	#Link/#Entity	13.44	22.66	5.37
	Sparsity	0.0071	0.0006	0.0031
Lastfm	#Entity	1,867	69,226	1,867+69,226
	#Link	15,328	682,314	92,834
	#Link/#Entity	8.24	15.75	3.03
	Sparsity	0.0044	0.0001	0.0007

Recommendation accuracy

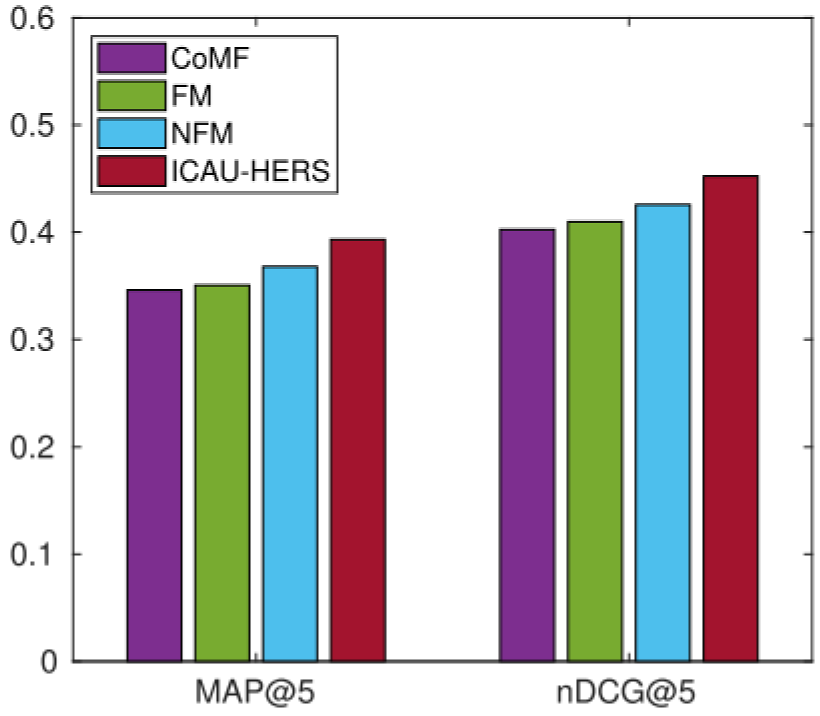
	Delicious				Lastfm			
	MAP@5	MAP@20	nDCG@5	nDCG@20	MAP@5	MAP@20	nDCG@5	nDCG@20
BPR-MF	0.4157	0.3225	0.4318	0.3744	0.5154	0.4586	0.6252	0.6334
SoRec	0.4174	0.3390	0.4476	0.3965	0.5350	0.4775	0.6412	0.6457
Social MF	0.4181	0.3409	0.4520	0.4017	0.5489	0.4907	0.6544	0.6575
SoReg	0.4239	0.3444	0.4577	0.4056	0.5495	0.4878	0.6548	0.6541
CMF	0.4375	0.3507	0.4739	0.4158	0.5530	0.4928	0.6549	0.6749
FM	0.4246	0.3363	0.4522	0.3896	0.5366	0.4837	0.6453	0.6723
NFM	0.4565	0.3754	0.4924	0.4347	0.5462	0.4885	0.6516	0.6702
ICAU-HERS	**0.5477**	**0.4200**	**0.6064**	**0.5273**	**0.5865**	**0.5302**	**0.6913**	**0.7021**

Item recommendation for cold-start users

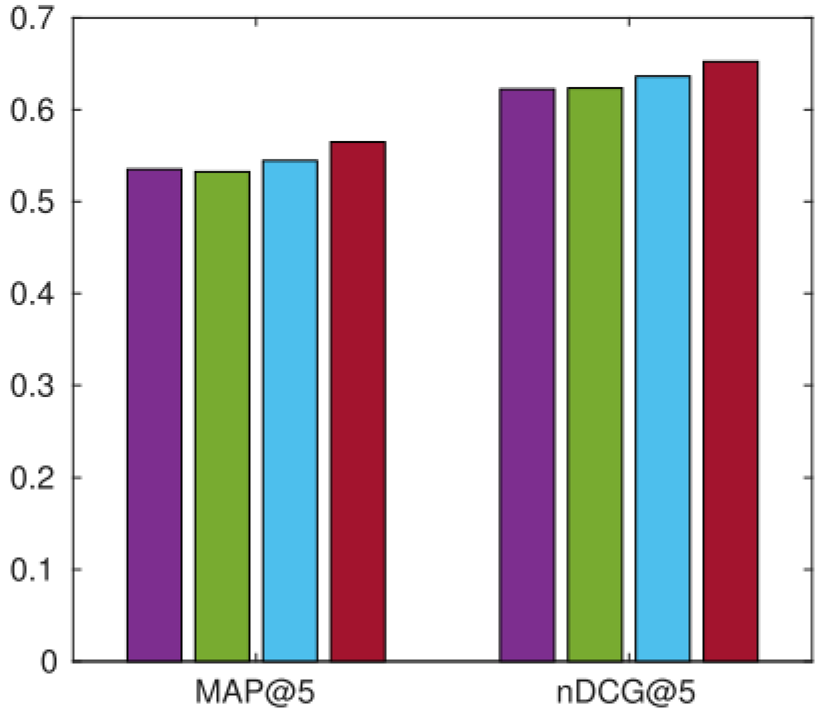


(a) Delicious (b) Lastfm

User recommendation for cold-start items

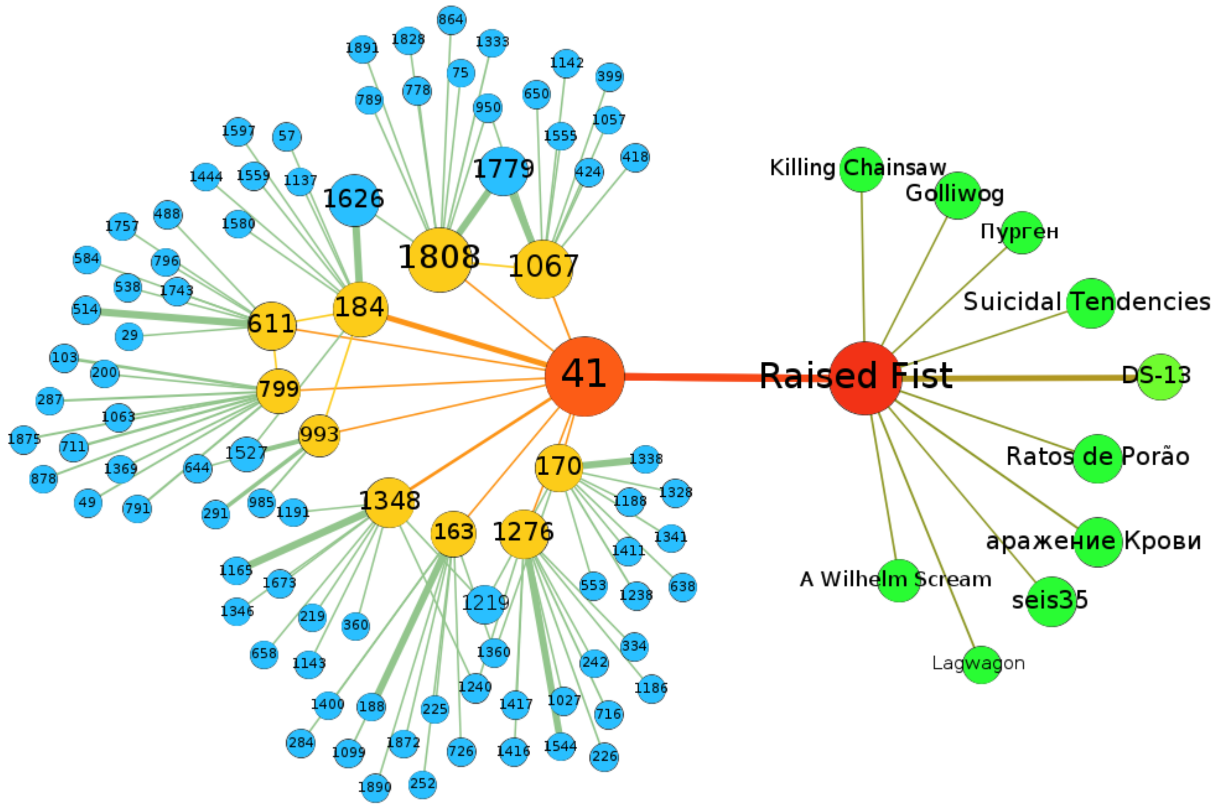


(a) Delicious



(b) Lastfm

Visualization and Interpretation



- The artists in the item network are labeled by their names.
- The anonymous users in the user network are labeled with their IDs.
- The thickness of edges specifies the significance of influence.

Pattern Relation Analysis/ Combined Pattern Mining

Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery, 3(2): 140-155, 2013

Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, IEEE Trans. SMC Part B, 41(3): 699 – 712, 2011

Longbing Cao. Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008

Combined Pattern Pairs

- Pair patterns

$$\mathcal{P} ::= \mathcal{G}(P_1, P_2)$$

$$\mathcal{P}: \begin{cases} X_1 \rightarrow T_1 \\ X_2 \rightarrow T_2 \end{cases}$$

$$\mathcal{E}: \begin{cases} X_\mathrm{p} \rightarrow T_1 \\ X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T_2 \end{cases}$$

$$I_{\mathrm{pair}}(\mathcal{P}) = \begin{cases} |Conf(P_1) - Conf(P_2)|, & \text{if } T_1 = T_2; \\ \sqrt{Conf(P_1)\,Conf(P_2)}, & \text{if } T_1 \text{ and } T_2 \text{ are contrary;} \\ 0, & \text{otherwise;} \end{cases}$$

$$I_{\mathrm{pair}}(\mathcal{P}) = Lift_V(R_1)\, Lift_V(R_2)\, dist(T_1, T_2)$$

$$\begin{aligned} Cont_\mathrm{e}(P) &= \frac{Lift(X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T)}{Lift(X_\mathrm{p} \rightarrow T)} \\ &= \frac{Conf(X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T)}{Conf(X_\mathrm{p} \rightarrow T)} \end{aligned}$$

$$I_{\mathrm{rule}}(X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T) = \frac{Cont_\mathrm{e}(X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T)}{Lift(X_\mathrm{e} \rightarrow T)}$$

$$Cps(X_\mathrm{e} \rightarrow T|X_\mathrm{p}) = Prob(X_\mathrm{e} \rightarrow T|X_\mathrm{p}) - Prob(X_\mathrm{e}|X_\mathrm{p}) \times Prob(T|X_\mathrm{p})$$

$$= \frac{Prob(X_\mathrm{p} \wedge X_\mathrm{e} \rightarrow T)}{Prob(X_\mathrm{p})} - \frac{Prob(X_\mathrm{p} \wedge X_\mathrm{e})}{Prob(X_\mathrm{p})} \times \frac{Prob(X_\mathrm{p} \rightarrow T)}{Prob(X_\mathrm{p})}$$

Longbing Cao. Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008.

Combined Pattern Clusters

- Cluster patterns

$$\mathcal{P} ::= \mathcal{G}(P_1, \ldots, P_n)(n > 2)$$

$$\mathcal{C}\colon \begin{cases} X_1 \rightarrow T_1 \\ \cdots \\ X_k \rightarrow T_k \end{cases}$$

$$\mathcal{S}\colon \begin{cases} X_{\mathrm{p}} \rightarrow T_1 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \rightarrow T_2 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \rightarrow T_3 \\ \cdots \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \wedge \cdots \wedge X_{\mathrm{e,k-1}} \rightarrow T_k \end{cases}$$

$$I_{\mathrm{cluster}}(\mathcal{C}) = \max_{P_i, P_j \in \mathcal{C}, i \neq j} I_{\mathrm{pair}}(P_i, P_j)$$

Combined Pattern Clusters

An Example of Combined Pattern Clusters

Clusters	Rules	X_{p}	X_{e}		T	Cnt	$Conf$	I_{r}	I_{c}	$Lift$	$Cont_{\mathrm{p}}$	$Cont_{\mathrm{e}}$	$Lift$ of	$Lift$ of
		demographics	arrangements	repayments			(%)						$X_{\mathrm{p}} \rightarrow T$	$X_{\mathrm{e}} \rightarrow T$
$\mathcal{P}_1$	P_5	marital:sin &gender:F &benefit:N	irregular	cash or post	A	400	83.0	1.12	0.67	1.80	1.01	2.00	0.90	1.79
	P_6		withhold	cash or post	A	520	78.4	1.00		1.70	0.89	1.89	0.90	1.90
	P_7		withhold & irregular	cash or post & withhold	B	119	80.4	1.21		2.28	1.33	2.06	1.10	1.71
	P_8		withhold	cash or post & withhold	B	643	61.2	1.07		1.73	1.19	1.57	1.10	1.46
	P_9		withhold & vol. deduct	withhold & direct debit	B	237	60.6	0.97		1.72	1.07	1.55	1.10	1.60
	P_{10}		cash	agent	C	33	60.0	1.12		3.23	1.18	3.07	1.05	2.74
$\mathcal{P}_2$	P_{11}	age:65+	withhold	cash or post	A	1980	93.3	0.86	0.59	2.02	1.06	1.63	1.24	1.90
	P_{12}		irregular	cash or post	A	462	88.7	0.87		1.92	1.08	1.55	1.24	1.79
	P_{13}		withhold & irregular	cash or post	A	132	85.7	0.96		1.86	1.18	1.50	1.24	1.57
	P_{14}		withhold & irregular	withhold	C	50	63.3	2.91		3.40	2.47	4.01	0.85	1.38

Pattern Relation Analysis

- Shoujin Wang, Longbing Cao. Inferring Implicit Rules by Learning Explicit and Hidden Item Dependency. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(3): 935-946, 2020.
- Jingyu Shao, Junfu Yin, Wei Liu,, Longbing Cao. Mining actionable combined patterns of high utility and frequency. DSAA 2015: 1-10
- Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery, 3(2): 140-155, 2013
- Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, IEEE Trans. SMC Part B, 41(3): 699 – 712, 2011
- Yanchang Zhao, Huaifeng Zhang, Longbing CaoChengqi Zhang. Combined Pattern Mining: from Learned Rules to Actionable Knowledge, LNCS 5360/2008, 393-403, 2008
- Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang. Combined Association Rule Mining, PAKDD2008
- Longbing Cao. Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008

Non-IID Statistical Learning

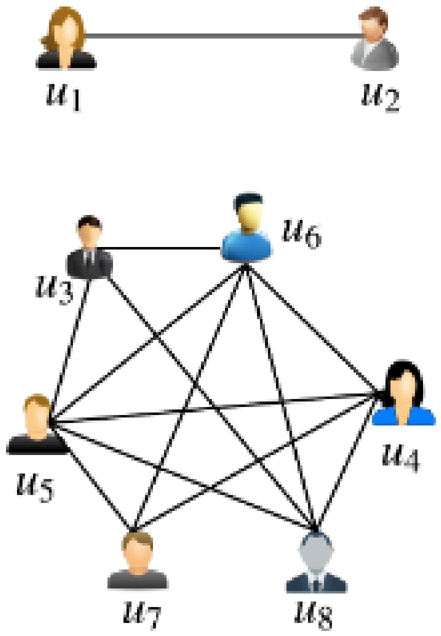
PAKDD2019 Tutorial on Large-scale statistical learning

https://datasciences.org/large-scale-statistical-learning/

Large-scale, Sparse, Multi-source Data: Non-IIDness

	The Godfather	The Dark Knight	Goodfellas	Toy Story 3	Alien
u_1	5	3	5	4	?
u_2	5	?	5	?	?
u_3	1	3	?	?	?
u_4	1	?	?	?	?
u_5	1	3	?	4	?
u_6	1	3	?	4	?
u_7	?	3	?	5	?
u_8	?	?	?	?	?

(a) Rating table



(b) User friendship

	Age	Location	Occupation	Education
u_1	28	NY	Developer	Bac
u_2	27	NY	Nurse	Bac
u_3	42	HI	Prof.	PhD
u_4	40	HI	Prof.	PhD
u_5	43	HI	Prof.	PhD
u_6	41	HI	Prof.	PhD
u_7	42	HI	Prof.	PhD
u_8	45	HI	Prof.	PhD

(c) User metadata

Bayesian Probabilistic Models

In Equation:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta}$$

In Plain English:

Posterior = $\frac{Likelihood * Prior}{Evidence}$

Bayesian Probabilistic Models

- $X = \{x_1, x_2, \ldots, x_n\}$ represents the data and θ represents the model parameters.
- It is assumed that $\{x_i\}$ are independent and identically distributed (i.i.d) conditioning on the prior ϑ.

$$P(X|\theta) = \prod_{i=1}^{n} P(x_i|\theta).$$

- The data in X is exchangeable.

Hierarchical Priors

- One may construct a complex prior distribution using a hierarchy of simple distributions as

$$P(\theta) = \int \ldots \int P(\theta|\alpha_t)P(\alpha_t|\alpha_{t-1})\ldots P(\alpha_1)d\alpha_1 \ldots d\alpha_t$$

- For example: One can construct a hierarchy of Gamma distribution.

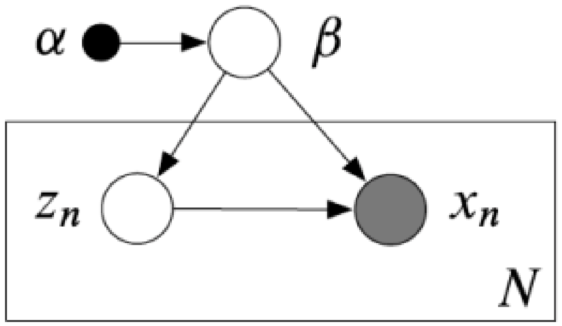
E.g., Gamma-Gamma-Gamma-Poisson distribution Compound models

Large-scale Bayesian Inference

- Sampling methods:
  - Markov Chain Monte Carlo (MCMC):
    - Metropolis-Hastings Sampling.
    - Gibbs Sampling
    - …
- Optimization methods
  - Variational Inference (VI)
  - Stochastic Variational Inference (SVI)

Stochastic Variational Inference (SVI)

- Model



$$z_n = z_{n,1:J}$$

$$p(x, z, \beta \mid \alpha) = p(\beta \mid \alpha) \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$

- Our goal: approximate the posterior

$$p(\beta, z \mid x)$$

- Locally independence

$$p(x_n, z_n \mid x_{-n}, z_{-n}, \beta, \alpha) = p(x_n, z_n \mid \beta, \alpha).$$

https://www.cs.ubc.ca/labs/lci/mlrg/slides/SVI.pdf

Stochastic Variational Inference (SVI)

- Conjugacy relation between the global variable and local variable

$$p(x_n, z_n|\beta) = h(x_n, z_n)\exp\{\beta^\top t(x_n, z_n) - a_\ell(\beta)\}.$$

- Prior of global variable is also exponential

$$p(\beta) = h(\beta)\exp\{\alpha^\top t(\beta) - a_g(\alpha)\}$$

- Posterior

$$p(z, \beta|x) = \frac{p(x, z, \beta)}{\int p(x, z, \beta) dz d\beta}.$$

https://www.cs.ubc.ca/labs/lci/mlrg/slides/SVI.pdf

Stochastic Variational Inference (SVI)

- ELBO

$$
\begin{aligned}
\log p(x) &= \log \int p(x,z,\beta)\,dz\,d\beta \\
&= \log \int p(x,z,\beta)\frac{q(z,\beta)}{q(z,\beta)}\,dz\,d\beta \\
&= \log \left(\mathbb{E}_q \left[\frac{p(x,z,\beta)}{q(z,\beta)} \right] \right) \\
&\geq \mathbb{E}_q[\log p(x,z,\beta)] - \mathbb{E}_q[\log q(z,\beta)] \\
&\triangleq \mathcal{L}(q).
\end{aligned}
$$

https://www.cs.ubc.ca/labs/lci/mlrg/slides/SVI.pdf

Statistical Learning of Large-scale, Sparse and Multi-source Data

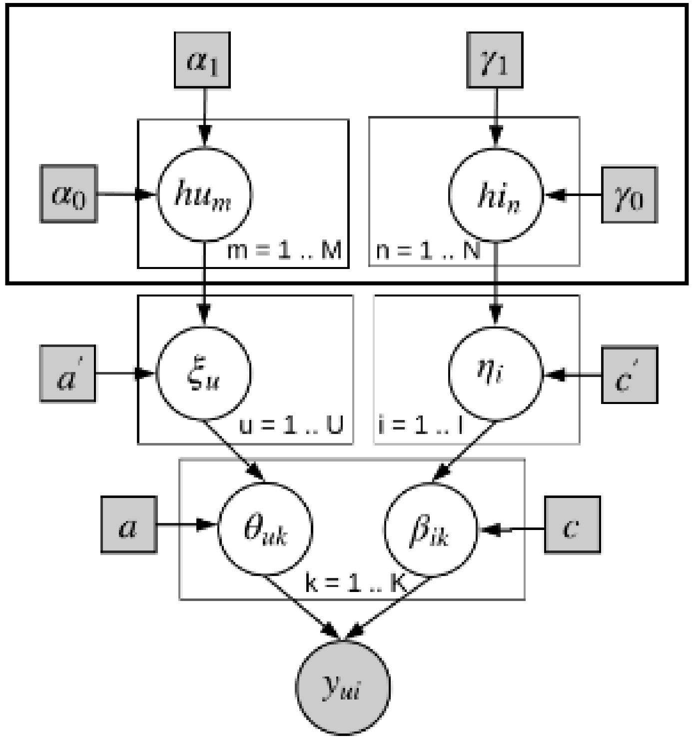
Trong Dinh Thac Do and Longbing Cao. Metadata-dependent Infinite Poisson Factorization for Efficiently Modelling Sparse and Large Matrices in Recommendation, IJCAI2018

Motivations

- User/item Sparsity:
  - PF is inefficient when working with a column or row with very few observations (corresponding to a sparse item or user) due to poor priors in the Gamma distribution.
- Dynamics/infinity:
  - Solve the challenge in automatically choosing the number of latent components.

Metadata-integrated Poisson Factorization (MPF)

Enrich prior using user and item metadata



(a) MPF

Metadata-integrated Poisson Factorization (MPF)

(1) For the m^{th} user attribute in the metadata, sample the weight:

$$hu_m \sim Gamma(\alpha_0, \alpha_1) \quad (1)$$

(2) For the n^{th} item attribute, sample the weight:

$$hi_n \sim Gamma(\gamma_0, \gamma_1) \quad (2)$$

(3) For each user u, sample latent behavior:

$$\xi_u \sim Gamma(a', \prod_{m=1}^{M} hu_m^{fu_{u,m}}) \quad (3)$$

(4) For each item i, sample latent attractiveness:

$$\eta_i \sim Gamma(c', \prod_{n=1}^{N} hi_n^{fi_{i,n}}) \quad (4)$$

(5) For each component k in the PF factorization:

(a) Sample user's latent preference:

$$\theta_{uk} \sim Gamma(a, \xi_u) \quad (5)$$

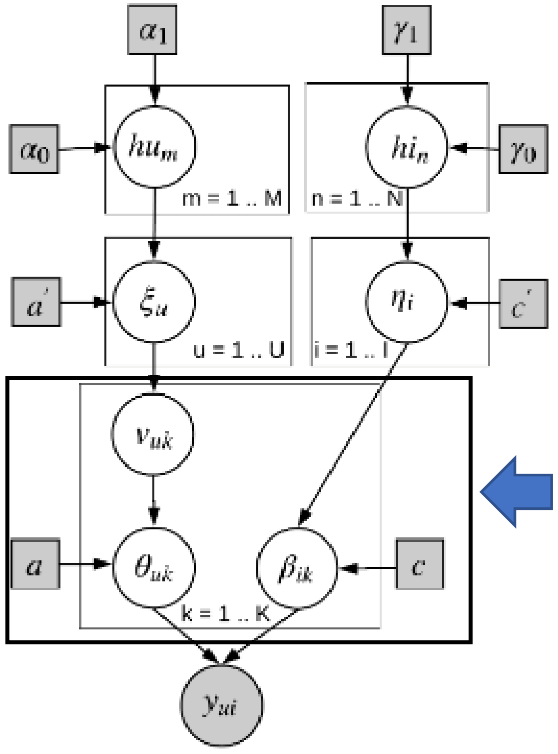
(b) Sample item's latent feature:

$$\beta_{ik} \sim Gamma(c, \eta_i) \quad (6)$$

(6) Sample rating:

$$y_{ui} \sim Poisson\left(\sum_k \theta_{uk}\beta_{ik}\right) \quad (7)$$

Metadata-integrated Infinite Poisson Factorization (MIPF)



(b) MIPF

Using Bayesian Nonparametric techniques to automatically determines the number of latent components

Metadata-integrated Infinite Poisson Factorization (MIPF)

(1) For the m^{th} user attribute, sample the weight:

$$hu_m \sim Gamma(\alpha_0, \alpha_1) \tag{8}$$

(2) For the n^{th} item attribute, sample the weight:

$$hi_n \sim Gamma(\gamma_0, \gamma_1) \tag{9}$$

(3) For each user $u(= 1, \ldots, M)$:

(a) Draw the user's latent behavior:

$$\xi_u \sim Gamma(a', \prod_{m=1}^{M} hu_m^{fu_{u,m}}) \tag{10}$$

(b) For $k(= 1..\infty)$, draw stick-breaking proportion:

$$v_{uk} \sim Beta(1, a') \tag{11}$$

(c) For $k(= 1..\infty)$, set the user's latent preference:

$$\theta_{uk} = \xi_u . v_{uk} \prod_{l=1}^{k-1} (1 - v_{ul}) \tag{12}$$

(4) For each item $i(= 1...N)$:

(a) Draw the item's latent attractiveness:

$$\eta_i \sim Gamma(c', \prod_{n=1}^{N} hi_n^{fi_{i,n}}) \tag{13}$$

(b) For $k = (1...\infty)$, set the item's latent feature:

$$\beta_{ik} \sim Gamma(c, \eta_i) \tag{14}$$

(5) For $u(= 1...M)$ and $i(= 1...N)$, draw

$$y_{ui} \sim Poisson\left(\sum_{k=1}^{\infty} \theta_{uk}\beta_{ik}\right) \tag{15}$$

Inference

- Variational Inference for MPF:
  - The mean-field family assumes each distribution is independent of the others.

$$q(hu, hi, \theta, \beta, \xi, \eta, z) = \prod_m q(hu_m|\zeta_m) \prod_n q(hi_n|\rho_n)$$
$$\prod_{u,k} q(\theta_{uk}|\nu_{uk}) \prod_{i,k} q(\beta_{ik}|\mu_{ik}) \prod_u q(\xi_u|\kappa_u) \quad (17)$$
$$\prod_i q(\eta_i|\tau_i) \prod_{u,i,k} q(z_{ui,k}|\phi_{ui,k})$$

We use the class of conditionally conjugate priors for hu_m, hi_n, θ_{uk}, β_{ik}, ξ_u, η_i and $z_{ui,k}$ to update the variational parameters $\{\zeta, \rho, \nu, \mu, \kappa, \tau, \phi\}$. For the Gamma distribution, we update both hyper-parameters: *shape* and *rate*.

IID assumption:
- **Independent**

Non-IID reality:
- **What if variables are non-IID?**

Inference

- Variational Inference for MiPF:
  - The mean-field family assumes each distribution is independent of the others.

$$q(hu, hi, v, \beta, \xi, \eta, z) = \prod_m q(hu_m|\zeta_m) \prod_n q(hi_n|\rho_n)$$

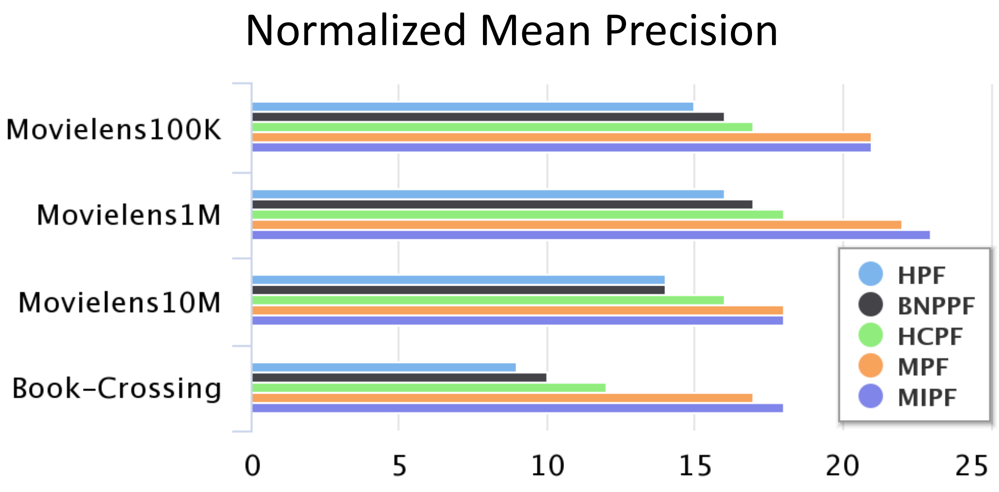
$$\prod_{k=1}^{\infty} \prod_u q(v_{uk}|\sigma_{uk}) \prod_{k=1}^{\infty} \prod_i q(\beta_{ik}|\mu_{ik}) \prod_u q(\xi_u|\kappa_u)$$

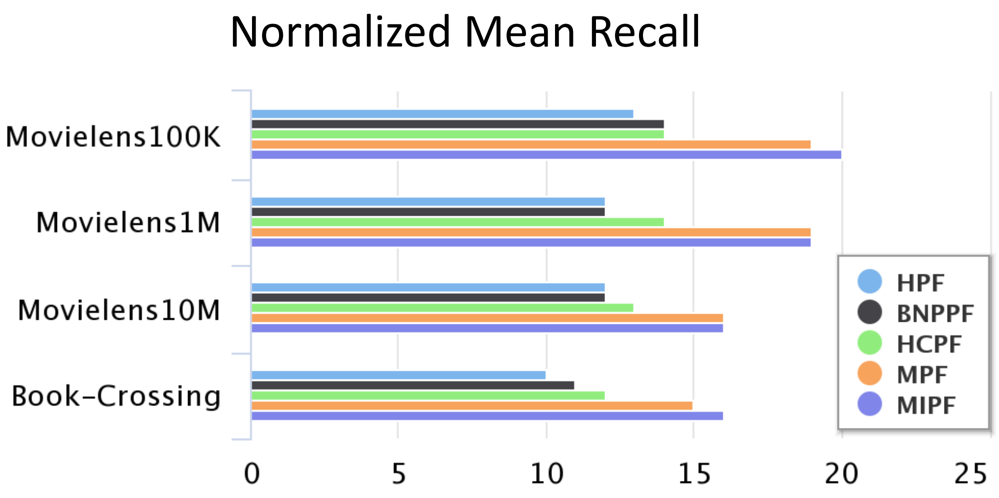
$$\prod_i q(\eta_i|\tau_i) \prod_{k=1}^{\infty} \prod_{u,i} q(z_{ui,k}|\phi_{ui,k})$$

IID assumption:
- **Independent**

Non-IID reality:
- **What if variables are non-IID?**

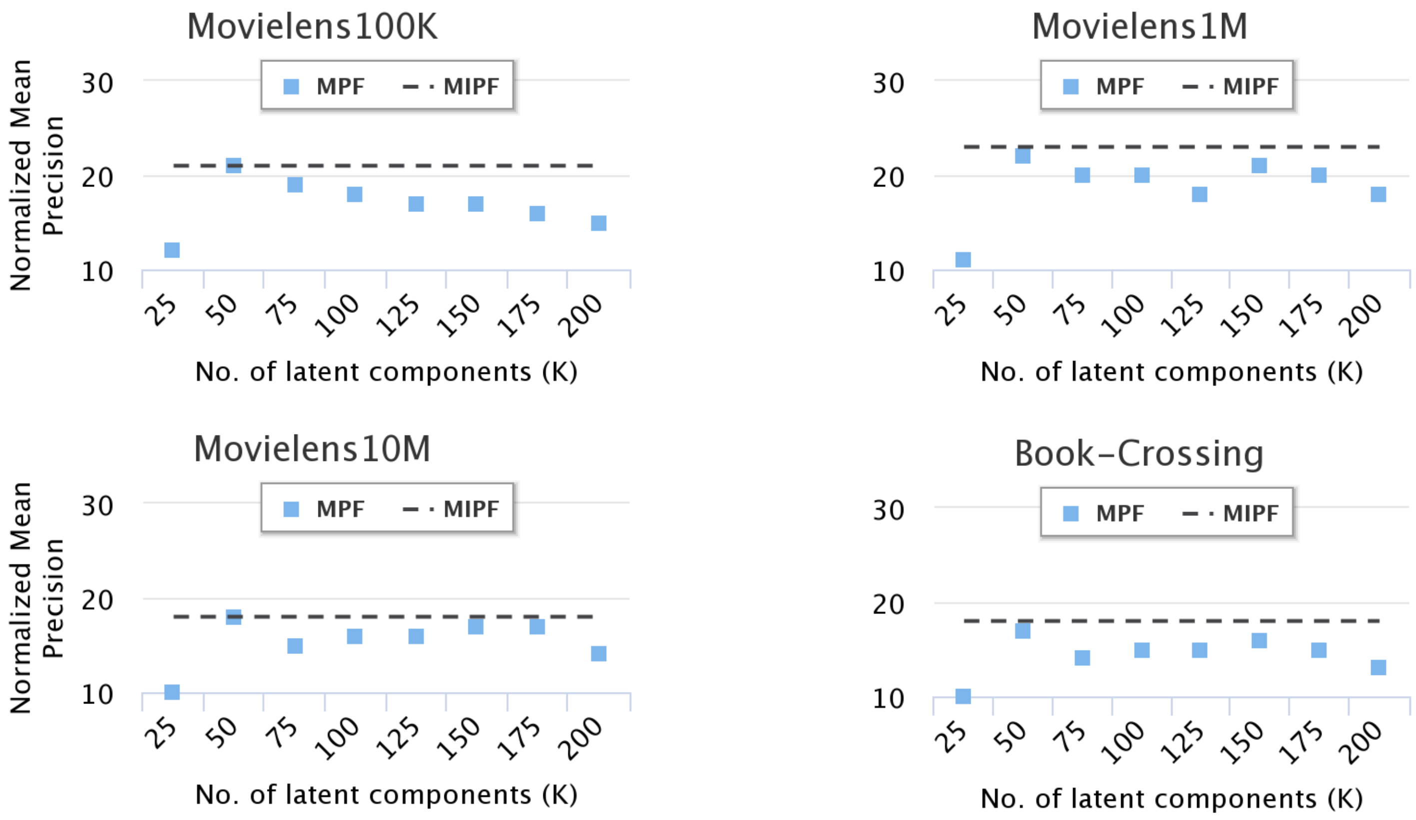
How Do MPF/MIPF Significantly Outperform Other PF Models?





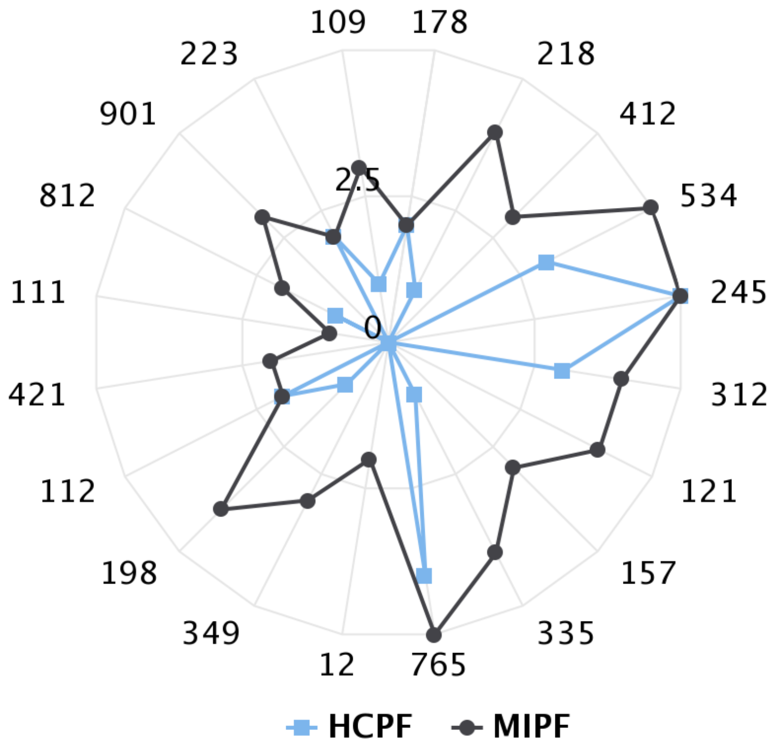
Top-20 Recommendation Compared with baselines

How Does MIPF Effectively Estimate the Number of Unbounded Latent Components?



Performance of top-30 recommendations made by finite model MPF and infinite model MIPF.

How Do MPF/MIPF Deal with Sparse Items/users?



Example of MIPF in handling sparse items in comparison with HCPF.

Contributions

- MPF/MIPF improve precision when working with large and sparse data by integrating user/item metadata.

- MIPF efficiently estimates the number of latent components.

- The variational inference for MPF and MIPF applies to massive data.

Non-IID Recommender Systems

Longbing Cao. Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting. Engineering, 2: 212-224, 2016.

https://datasciences.org/recommender-systems/

Framework of Non-IID Recommender Systems

Longbing Cao. Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting. Engineering, 2: 212-224, 2016.

Longbing Cao, Philip Yu. Non-IID Recommendation Theories and Systems. IEEE Intelligent Systems, 31(2), 81-84, 2016.

Challenges

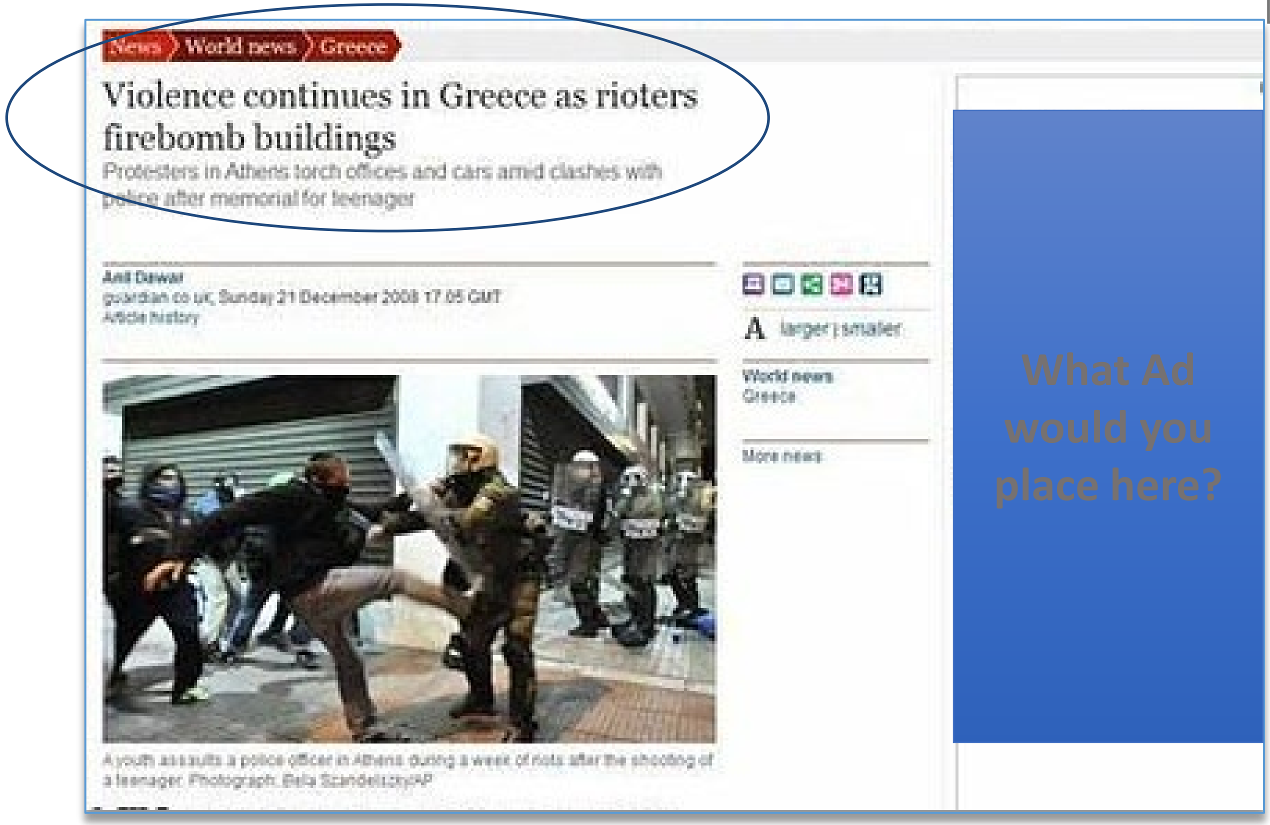
Amazon

Recommendation problems:

- Duplicated
- Irrelevant
- Missing
- Falsified
- ...

Big Data Challenges Existing Theories and Systems

Irrelevant and Damaging to Brand



Why the Prediction Doesn’t Work?

- There may be many reasons,
  - Content understanding
  - Understand the semantic hidden in contents
  - Analyze the relevance between news and ads from every possible aspect
  - Treat each piece of news differently
  - …

- A fundamental assumption - IIDness
  - Weaken or overlook the data complexities
    - Relationships between objects, syntactically, semantically,
    - Heterogeneity between objects, sources, …

A Systematic View of Recommendation

NS	SS	AS	CS	Subcategory
NC	SC	AC	CC	Category
NP	SP	AP	CP	Price
Name	Sex	Age	City	

(D). Implicit user-item interactions

Subcategory	C1.6	C2.2	C2.3
Category	C1	C2	C2
Price	100	800	1200
	i1	i2	i3

(C). Item properties

Name	Sex	Age	City	
John	M	45	Sydney	u1
Cindy	F	42	Sydney	u2
Julie	F	20	Sydney	u3

(B). User demographics

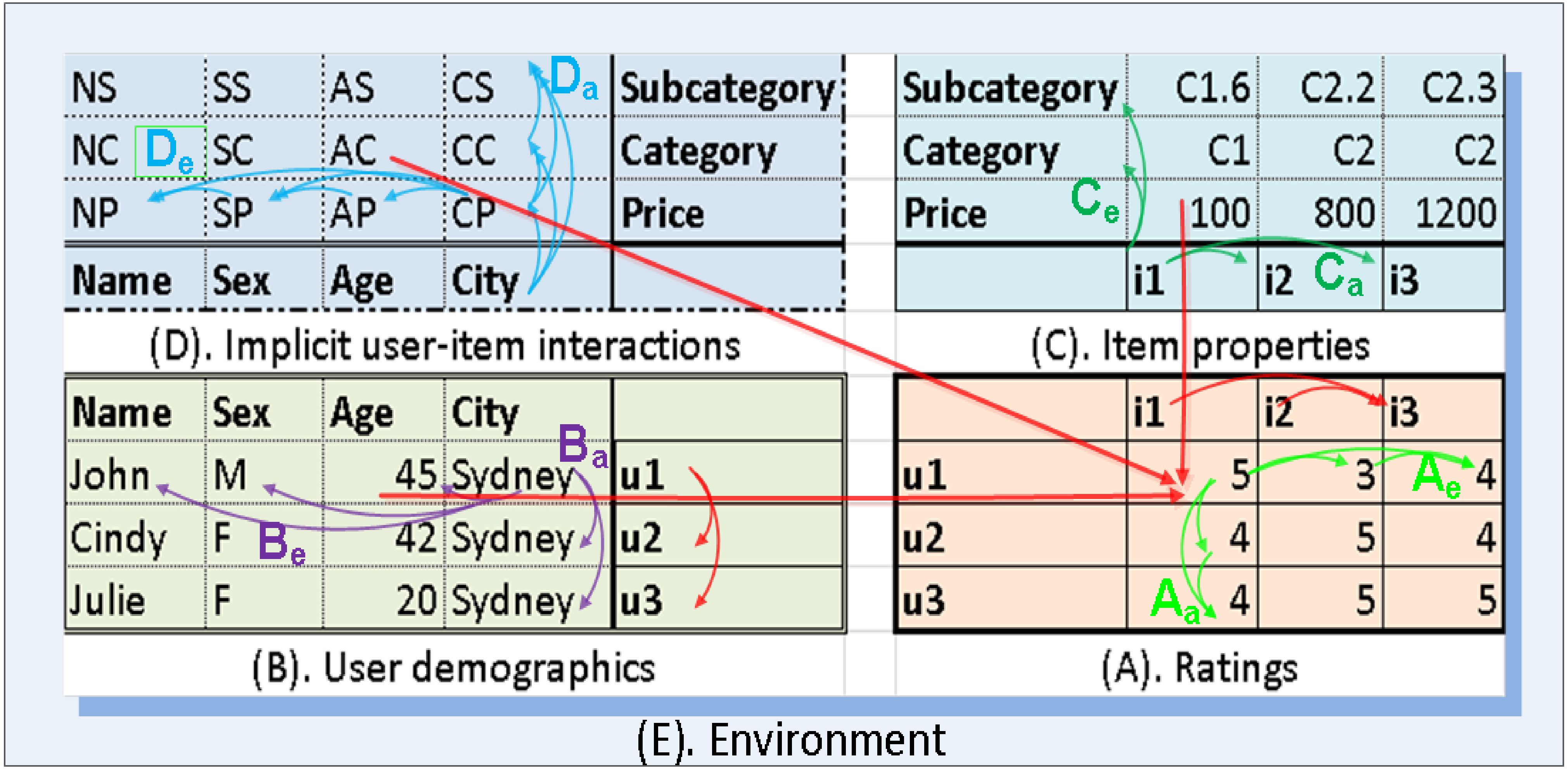
	i1	i2	i3
u1	5	3	4
u2	4	5	4
u3	4	5	5

(A). Ratings

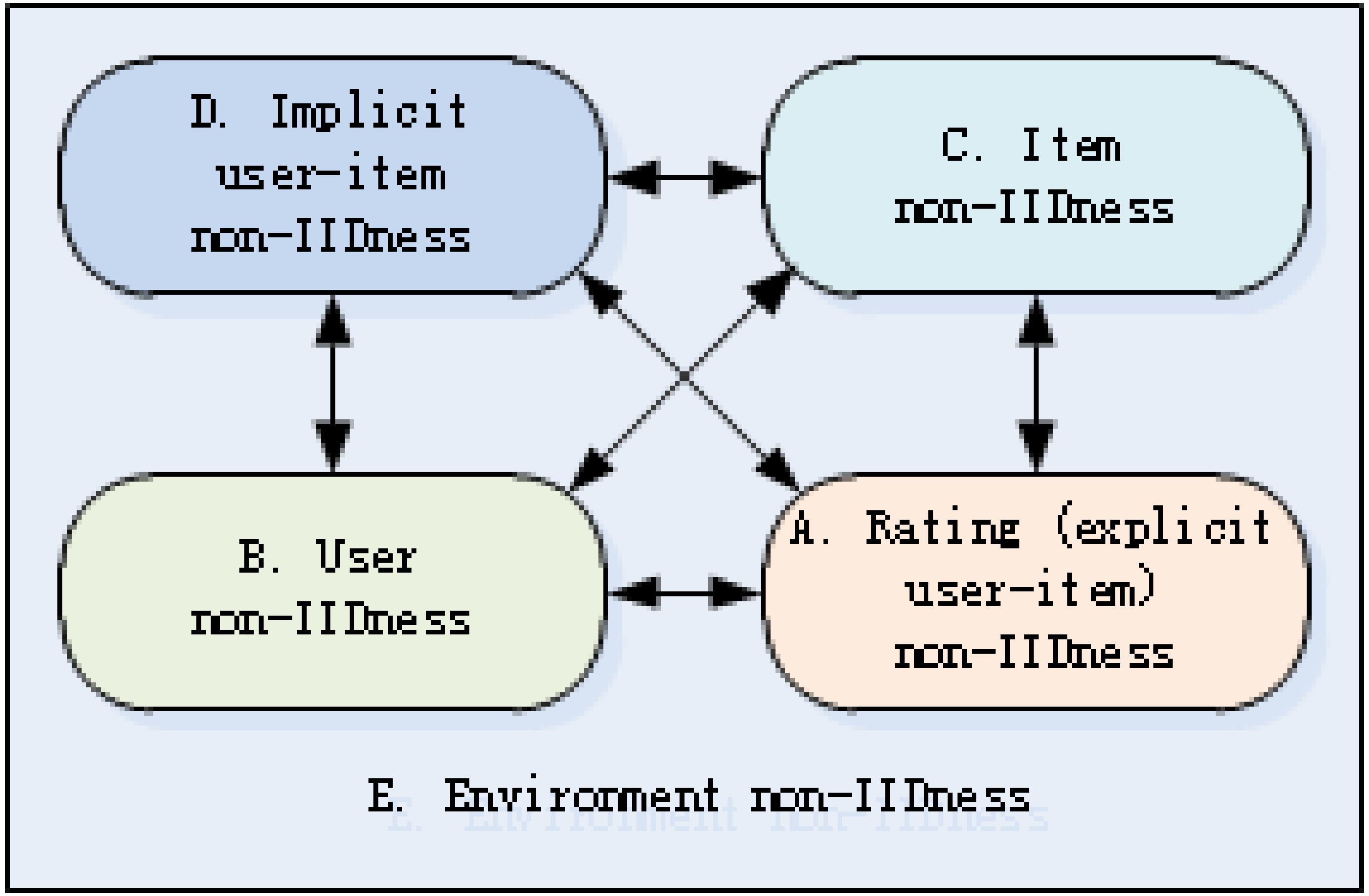
(E). Environment

Longbing Cao. *Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting*. Engineering, 2: 212-224, 2016.

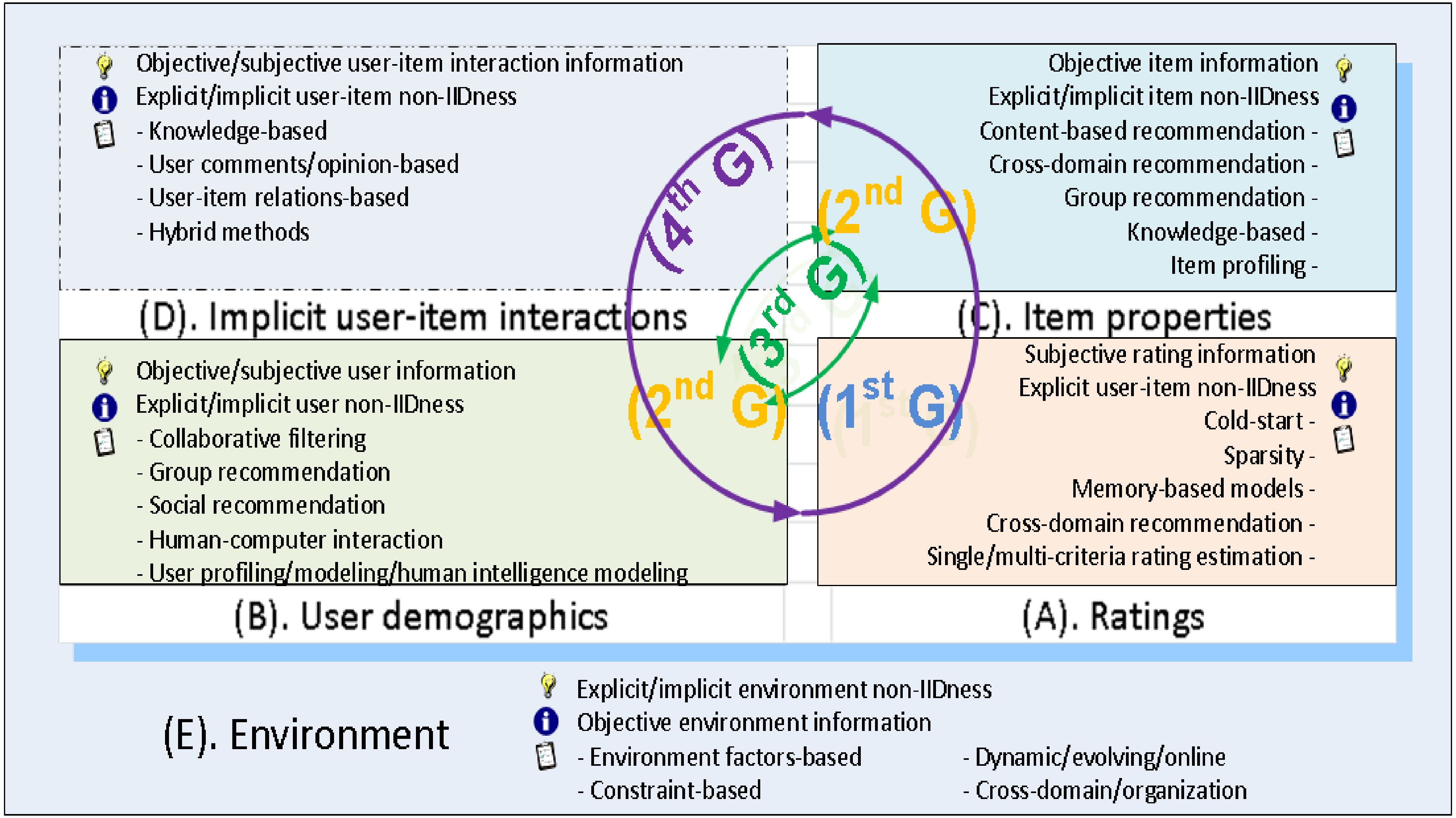
Non-IIDness in Recommendation



Non-IIDness in Recommendation



Four-stage Recommendation Research



Non-IIDness in Modern Recommendation

- Heterogeneity (Non-identical distribution)
  - Due to the **heterogeneity** of users, items and domains, it is improper to model the features of all users or items using identical distributions
  - Heteroskedastic modeling for recommendation in long tail
  - Modeling non-identical user feature distribution, non-identical item feature distribution and non-identical choice distribution
  - Cross-domain data (non-identical domain distribution due to heterogeneity)

Liang Hu, Wei Cao, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Bayesian Heteroskedastic Choice Modeling on Non-identically Distributed Linkages, ICDM 2014

Hu, L., Cao, L., Cao, J., Gu, Z., Xu, G., and Wang, J. Improving the Quality of Recommendations for Users and Items in the Tail of Distribution. ACM Trans. Inf. Syst., 2017

Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Can Zhu: Personalized recommendation via cross-domain triadic factorization. WWW 2013

Liang Hu, Longbing, Jian Cao, Zhiping Gu, Guandong Xu, & Dingyu Yang: Learning Informative Priors from Heterogeneous Domains to Improve Recommendation in Cold-Start User Domains. ACM Trans. Inf. Syst., (2016)

Liang Hu, Jian Cao, Guandong Xu, Jie Wang, Zhiping Gu, Longbing Cao, Cross-Domain Collaborative Filtering via Bilinear Multilevel Analysis, IJCAI 2013

Modeling Non-IID Recommender Systems

- Couplings (Non-independency)
  - Recommender systems were born with non-independency, they always try to find the coupling relationships among users, items, domains and other information
  - Social Influence (coupling related users' feedback)

    Hu, L., Cao, L., Cao, J., Gu, Z., Xu, G., and Wang, J. Improving the Quality of Recommendations for Users and Items in the Tail of Distribution. ACM Trans. Inf. Syst., 2017
  - Group-based Recommendation (joint decision)

    Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Wei Cao, Deep Modeling of Group Preferences for Group-based Recommendation, AAAI 2014
  - Session-based Recommendation (context dependent)

    Hu, L., Cao, L., Wang, S., Xu, G., Cao, J. and Gu, Z. 2017. Diversifying personalized recommendation with user-session context. (IJCAI'17)
  - Cross-domain recommendation (multi-domain dependency)

    Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Can Zhu: Personalized recommendation via cross-domain triadic factorization. WWW 2013

    Liang Hu, Longbing, Jian Cao, Zhiping Gu, Guandong Xu, & Dingyu Yang: Learning Informative Priors from Heterogeneous Domains to Improve Recommendation in Cold-Start User Domains. ACM Trans. Inf. Syst., (2016

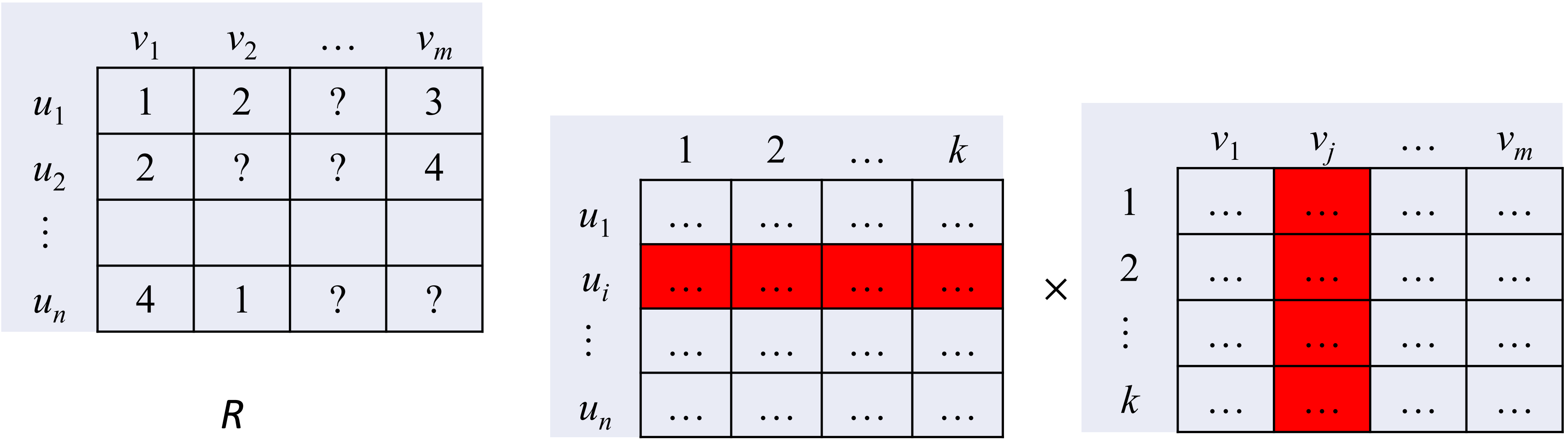
Coupled Matrix Factorization within Non-IID Context

Fangfang Li, Guandong Xu, Longbing Cao. Coupled Matrix Factorization within Non-IID Context, PAKDD2015, 707-719.

One Basic Approach: MF (Matrix Factorization)

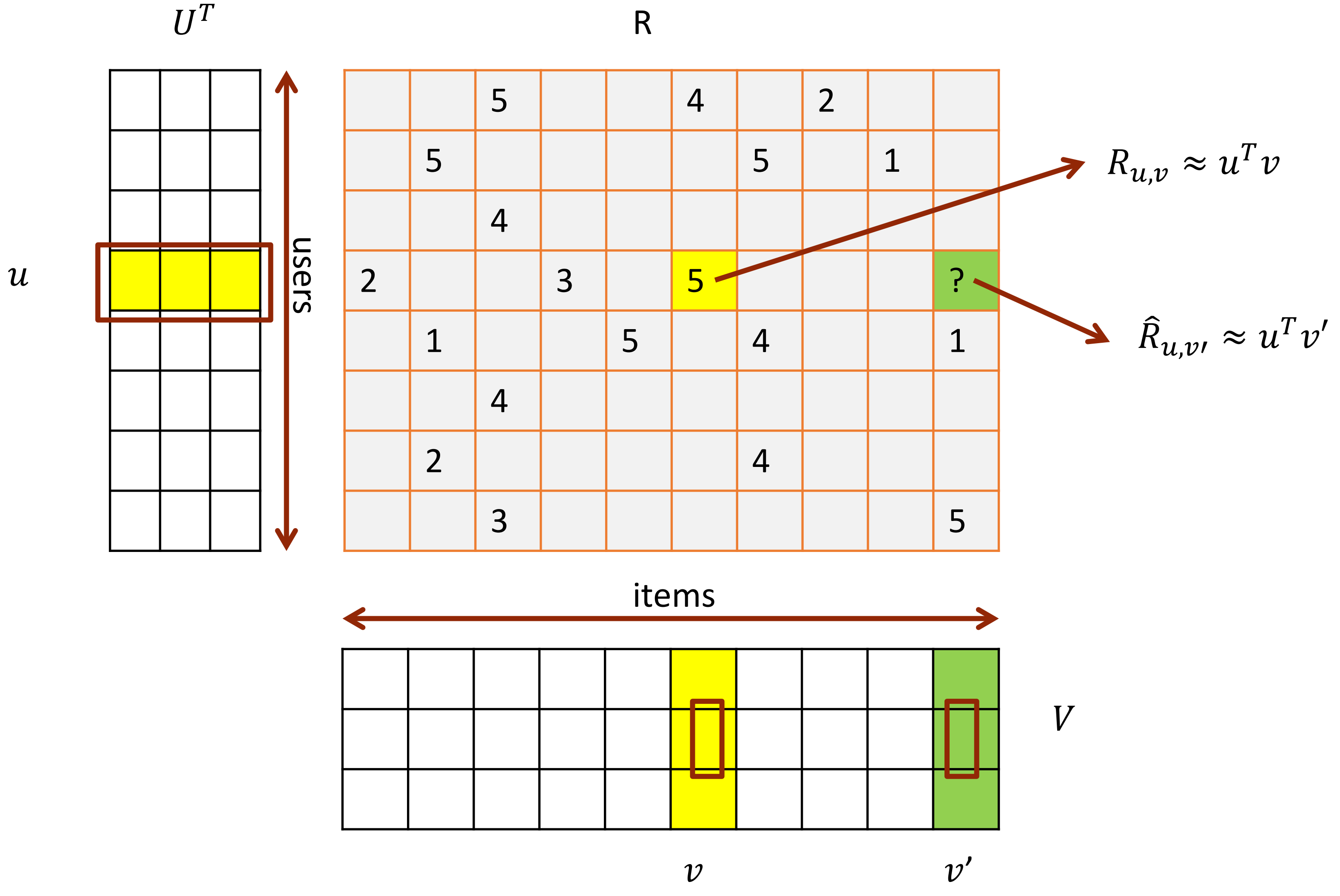
- Idea: project users and items into a joint k-dimensional space.
  - Represent user u_i, and item v_j using P_i and Q_j as their latent profile respectively
  - Rating R_{ij} is predicted as:

$$\boxed{\begin{aligned} R \approx \hat{R} &= P^T Q \\ \hat{R}_{ij} &= {P^T}_i \cdot Q_j \end{aligned}}$$



R P^T Q

Matrix Factorization

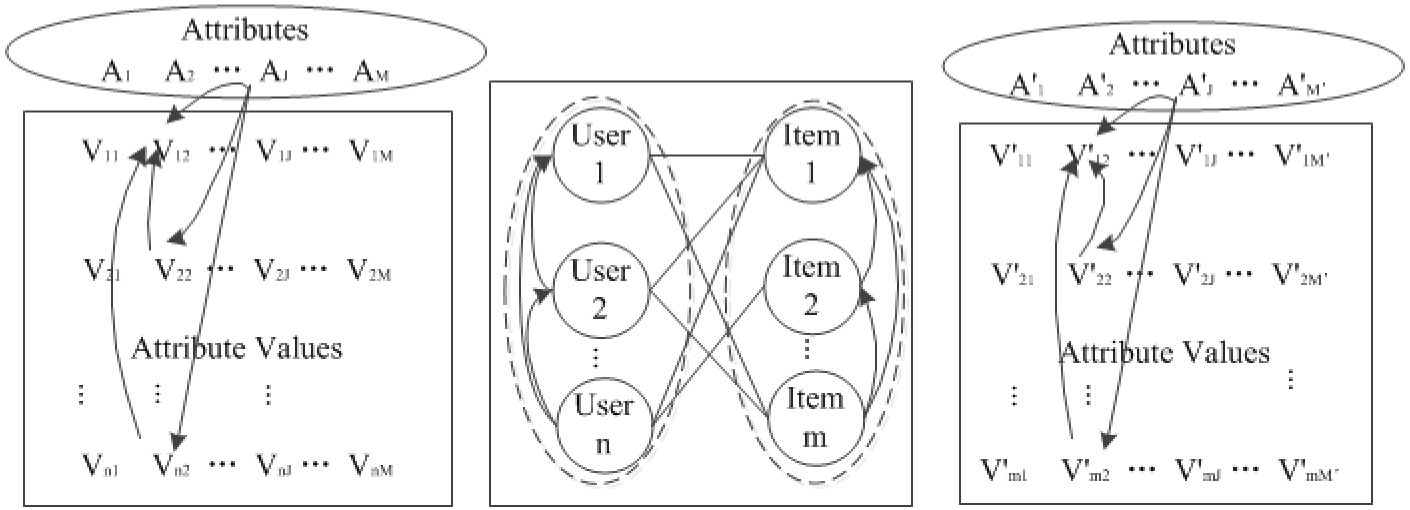


Problems and Solution

- MF problems:
  - MF solve the rating estimation as a mathematical problem
  - Same rating table for different businesses would lead to same rating estimation
  - User/item non-IIDness are not involved
- Solution:
  - Combine CF and content-based method together.
  - Deeper analysis by considering the non-IID characteristics for items and users.

User/Item Coupling Analysis

- Deep couplings within users and items contribute to the rating behavior.
  - Attribute values are coupled and not independent,
  - Attributes are also coupled and influence each other.



Non-IID Users

- For two users described by the attribute space, the Coupled User Similarity (CUS) is defined to measure the similarity between users.

Definition 1. *Formally, given user attribute space* $S_U = < U, A, V, f >$, *the* ***Coupled User Similarity (CUS)*** *between two users* u_i *and* u_j *is defined as follows.*

$$CUS(u_i, u_j) = \sum_{k=1}^{J} \delta_k^{Ia}(V_{ik}, V_{jk})) * \delta_k^{Ie}(V_{ik}, V_{jk})) \qquad (1)$$

where V_{ik} *and* V_{jk} *are the values of attribute* k *for users* u_i *and* u_j, *respectively; and* δ_k^{Ia} *is the intra-coupling within attribute* A_k, δ_k^{Ie} *is the inter-coupling between different attributes.*

Non-IID Items

- For two items described by the attribute space, the Coupled Item Similarity (CIS) is defined to measure the similarity between items.

Definition 2. *Formally, given item attribute space* $S_O = < O, A', V', f' >$, *the* ***Coupled Item Similarity (CIS)*** *between two items* o_i *and* o_j *is defined as follows.*

$$CIS(o_i, o_j) = \sum_{k=1}^{J'} \delta_k^{Ia}(V'_{ik}, V'_{jk})) * \delta_k^{Ie}(V'_{ik}, V'_{jk})) \quad (2)$$

where V'_{ik} *and* V'_{jk} *are the values of attribute* j *for items* o_i *and* o_j, *respectively; and* δ_k^{Ia} *is the intra-coupling within attribute* A_k, δ_k^{Ie} *is the inter-coupling between different attributes.*

Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao, Chi-Hung Chi: *Coupled Attribute Similarity Learning on Categorical Data*. IEEE Trans. Neural Netw. Learning Syst. 26(4): 781-797 (2015)

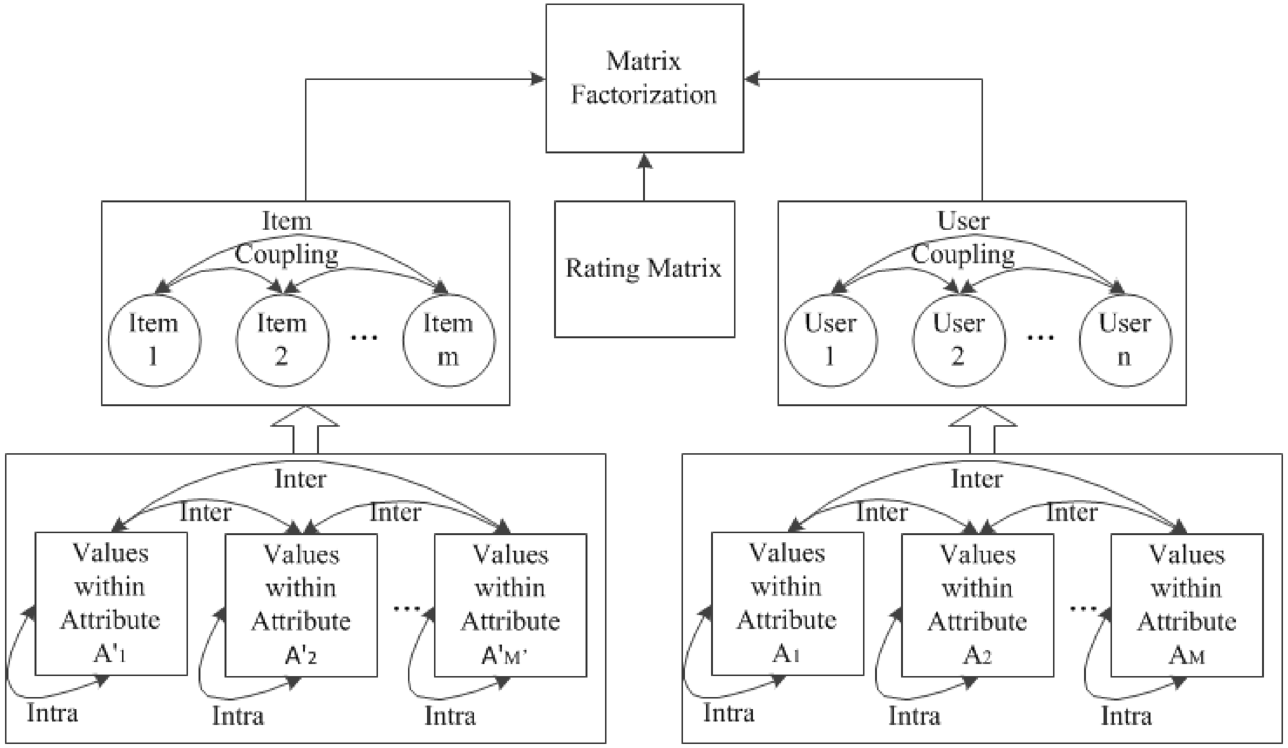
Matrix Factorization

- Traditionally, the rating matrix can be modeled by MF as:
  - The prediction task of matrix is transformed to compute user's factor matrix P and item's factor matrix Q.
  - Once P and Q are calculated, R can be easily reconstructed to predict the rating given by one user to an item.

$$\hat{R} = r_m + PQ^T$$

Coupled MF (CMF)

- CMF considers three sorts of information
  - Traditional rating matrix
  - Non-IID User coupling based on users' attributes
  - Non-IID Item coupling based on items' attributes



CMF Model

- Objective Function

$$L = \frac{1}{2}\sum_{(u,o_i)\in K}\left(R_{u,o_i} - \hat{R}_{u,o_i}\right)^2 + \frac{\lambda}{2}\left(\|Q_i\|^2 + \|P_u\|^2\right) + \frac{\alpha}{2}\sum_{all(u)}$$

$$\left\| P_u - \sum_{v\in\mathbb{N}(u)} CUS(u,v)P_v \right\|^2 + \frac{\beta}{2}\sum_{all(o_i)}\left\| Q_i - \sum_{o_j\in\mathbb{N}(o_i)} CIS(o_i,o_j)Q_j \right\|^2$$

- Optimization

$$\frac{\partial L}{\partial P_u} = \sum_{o_i} I_{u,o_i}(r_m + P_u {Q_i}^T - R_{u,o_i})Q_i + \lambda P_u + \alpha(P_u -$$

$$\sum_{v\in\mathbb{N}(u)} CUS(u,v)P_v) - \alpha \sum_{v:u\in\mathbb{N}(v)} CUS(u,v)(P_v - \sum_{w\in\mathbb{N}(v)} CUS(v,w)P_w)$$

$$\frac{\partial L}{\partial Q_i} = \sum_{u} I_{u,o_i}(r_m + P_u {Q_i}^T - R_{u,o_i})P_u + \lambda Q_i + \beta(Q_i - \sum_{o_j\in\mathbb{N}(o_i)}$$

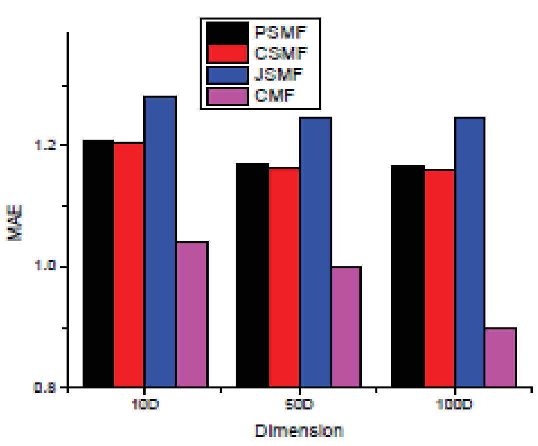
$$CIS(o_i,o_j)Q_j) - \beta \sum_{o_j:o_i\in\mathbb{N}(o_j)} CIS(o_j,o_i)(Q_j - \sum_{o_k\in\mathbb{N}(o_j)} CIS(o_j,o_k)Q_k)$$

Compared to MF and CF

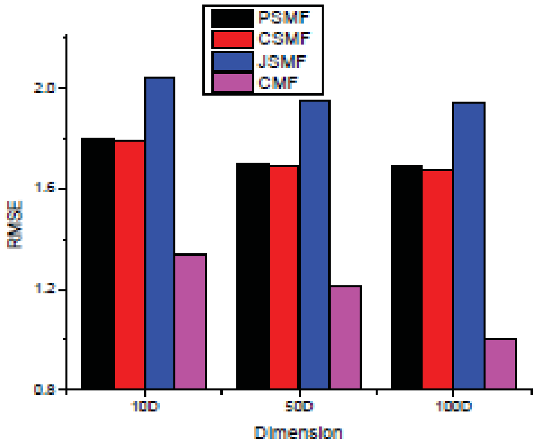
Data Set	Dim	Metrics	PMF (Improve)	ISMF (Improve)	RSVD (Improve)	CMF
Movielens	100D	MAE	1.1787(28.09%)	1.1125 (21.47%)	1.1076 (20.98%)	**0.8978**
		RMSE	1.7111 (71.07%)	1.5918 (59.14%)	1.5834 (58.30%)	**1.0004**
	50D	MAE	1.1852 (18.43%)	1.1188 (11.79%)	1.1088 (10.79%)	**1.0009**
		RMSE	1.8051 (58.98%)	1.6103 (39.50%)	1.5835 (36.82%)	**1.2153**
	10D	MAE	1.2129 (17.19%)	1.1651 (12.41%)	1.1098 (6.88%)	**1.0410**
		RMSE	1.8022 (46.25%)	1.7294 (38.97%)	1.5863 (24.66%)	**1.3397**
Bookcrossing	100D	MAE	1.5127 (3.65%)	1.5102 (3.40%)	1.5131 (3.69%)	**1.4762**
		RMSE	3.7455 (0.76%)	3.7397 (0.18%)	3.7646 (2.67%)	**3.7379**
	50D	MAE	1.5128 (3.67%)	1.5100 (3.39%)	1.5131 (3.70%)	**1.4761**
		RMSE	3.7452 (0.74%)	3.7415 (0.37%)	3.7648 (2.70%)	**3.7378**
	10D	MAE	1.5135 (3.73%)	1.5107 (3.45%)	1.5134 (3.72%)	**1.4762**
		RMSE	3.7483 (1.20%)	3.7440 (0.77%)	3.7659 (2.96%)	**3.7363**

Data Set	Metrics	UBCF (Improve)	IBCF (Improve)	CMF
Movielens	MAE	0.9027 (0.49%)	0.9220 (2.42%)	**0.8978**
	RMSE	1.0022 (0.18%)	1.1958 (19.54%)	**1.0004**
Bookcrossing	MAE	1.8064 (33.02%)	1.7865 (31.03%)	**1.4762**
	RMSE	3.9847 (24.68%)	3.9283 (19.04%)	**3.7379**

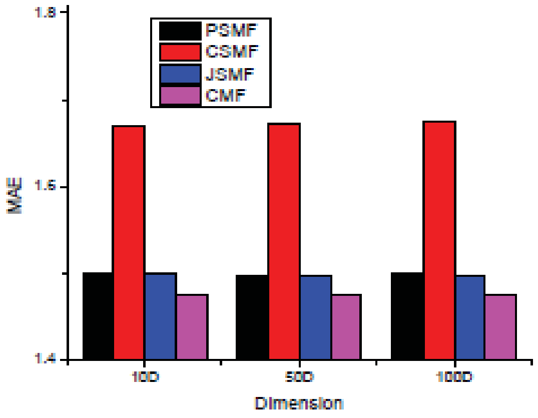
Compared to Hybrid Methods



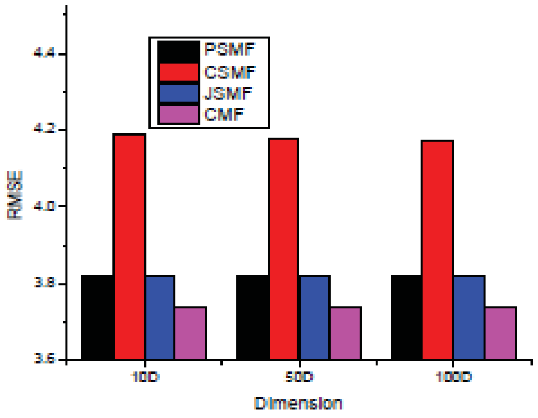
(a) MAE on Movielens



(b) RMSE on Movielens



(c) MAE on Bookcrossing



(d) RMSE on Bookcrossing

Summary of CMF

- Contributions
  - Applied a NonIID-based method to capture the couplings between users and items, based on their objective attribute information;
  - Integrated user coupling, item coupling and users’ subjective rating preferences into matrix factorization learning model;
  - Evaluated the effectiveness of Coupled MF model.

More Recent Work on non-IID recommender systems

- *Trong Dinh Thac Do and Longbing Cao. Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence, NIPS2018*
- *CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering, IJCAI2018*
- *Interpretable Recommendation via Attraction Modeling: Learning Multilevel Attractiveness over Multimodal Movie Contents, IJCAI2018*
- *Attention-based Transactional Context Embedding for Next-Item Recommendation. AAAI2018*

Dynamic, Continuous (Next-item), Personalized Recommendations within Session & Context

- Personalized recommendations
- With user/product sessions as context
- Behavior-based recommendations
- Continuous (next-product/moment/ interest/etc.) recommendations

Table 3: Accuracy comparisons on Tafang

Model	REC@10	REC@50	MRR
PBRS	0.0307	0.0307	0.0133
FPMC	0.0191	0.0263	0.0190
PRME	0.0212	0.0305	0.0102
GRU4Rec	0.0628	0.0907	0.0271
ATEM	**0.1089**	**0.2016**	**0.0347**
TEM	0.0789	0.1716	0.0231

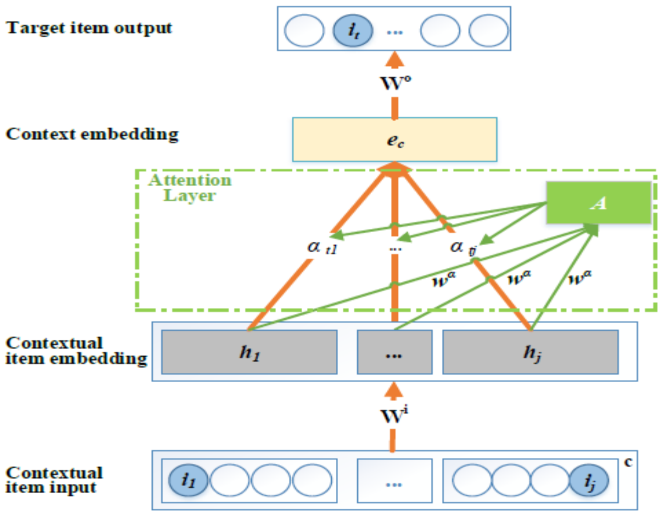


Figure 1: The ATEM architecture, which first learns item embeddings and then integrates them into the context embedding for target item prediction, where '*A*' represents the attention model.

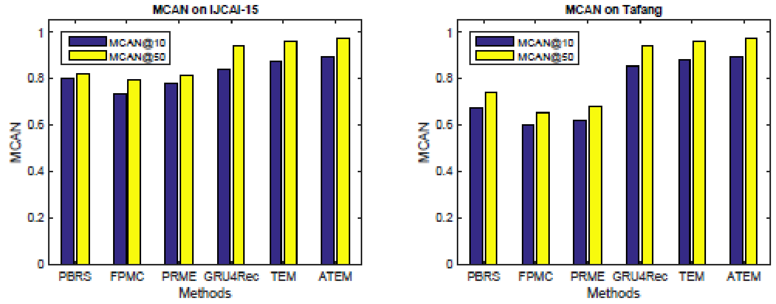
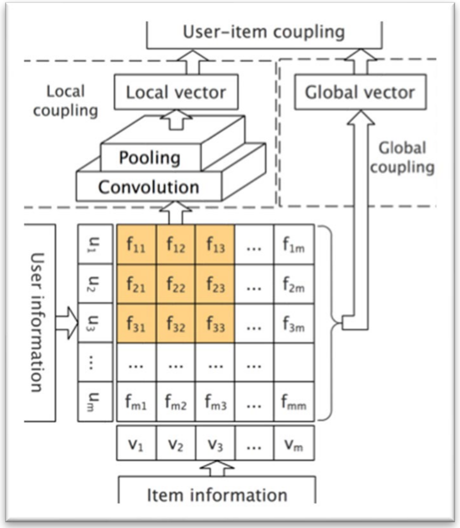


Figure 3: ATEM achieves higher novelty than the other approaches.

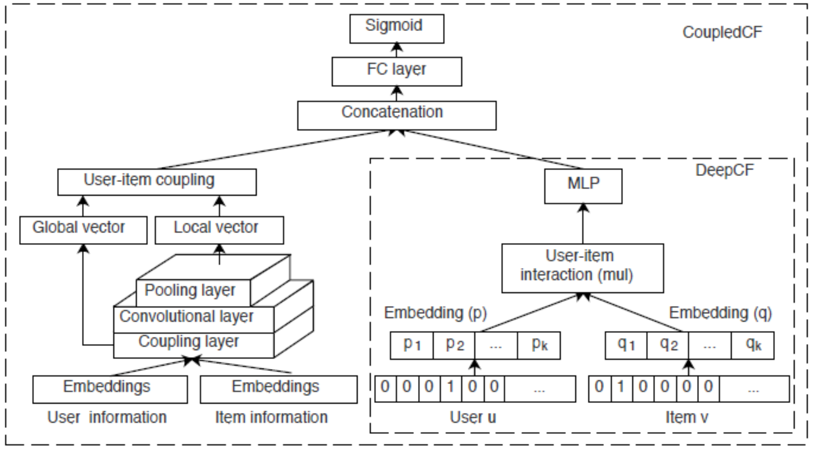
- *Attention-based Transactional Context Embedding for Next-Item Recommendation. AAAI2018*
- *Diversifying Personalized Recommendation with User-session Context. IJCAI2017*

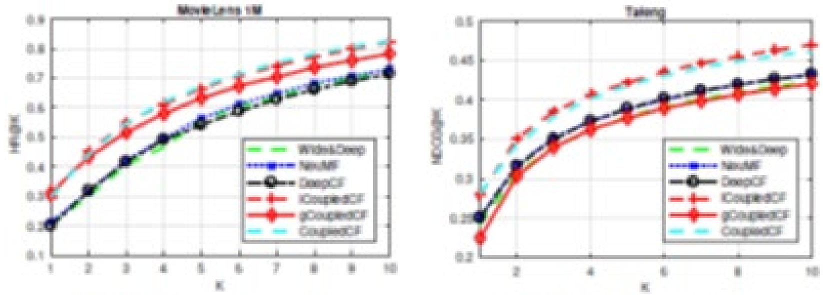
Deep Representation with Explicit and Implicit Feature Couplings

- Learn explicit user-product couplings by metadata-enabled CNN
- Build a deep collaborative filter model to learn the latent user-product relations
- Integrate both local and global user-product interactions components



- User's dense vector U
- Item's dense vector V
- User-item coupling F



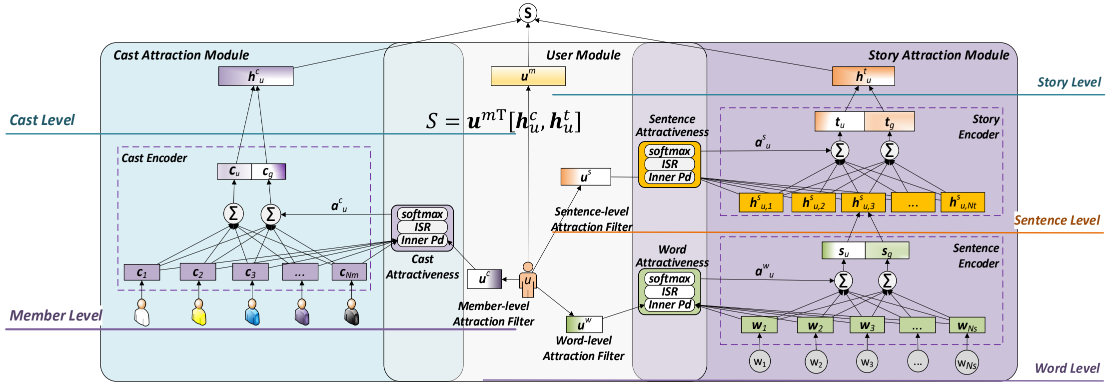


(a) HR@K on MovieLens (b) NDCG@K on Tafeng

- *CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering, IJCAI2018*

Attraction Modeling:
Learning Multilevel Attractiveness over Multimodal Content

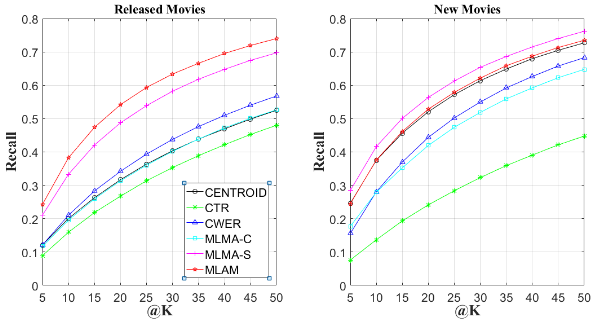
- One **multilevel neural model on the movie story** to capture
  - Word-level attraction: e.g., some characters, some place
  - Sentence-level attraction: e.g., some interesting plot
  - Story-level attraction: e.g., like the movie to what extent
- Another **multilevel neural model on the cast** to capture
  - Member-level attraction: e.g., a fan of some actor
  - Cast-level attraction: e.g., attracted by the movie to what extent



$$a_u^{c_i} = softmax\left(isr(\boldsymbol{u}^{c\mathrm{T}}\boldsymbol{c}_i)\right) \quad \boldsymbol{c}_u = \sum a_u^{c_i}\boldsymbol{c}_i \quad a_u^{w_i} = softmax\left(isr(\boldsymbol{u}^{w\mathrm{T}}\boldsymbol{w}_i)\right) \quad \boldsymbol{s}_u = \sum a_u^{w_i}\boldsymbol{w}_i$$

$$a_u^{s_i} = softmax\left(isr(\boldsymbol{u}^{s\mathrm{T}}\boldsymbol{h}_i^s)\right) \quad \boldsymbol{t}_u = \sum a_u^{s_i}\boldsymbol{h}_i^s$$

$$L_{m_{u,i} \succeq m_{u,j}} = \max(0, margin + S_{m_{u,j}} - S_{m_{u,i}})$$



User 156	Sentence level attractiveness	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title. The plot revolves around a high school election and satirizes both suburban high school life and politics. The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction. The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.
	Word level attractiveness	Election is a 1999 American **comedy-drama** film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title.
	Cast member attractiveness	**Alexander Payne**, Reese Witherspoon, Matthew Broderick, Jim Taylor
User 2163	Sentence level attractiveness	Election is a 1999 American comedy-drama film directed and written by Alexander Payne and adapted by him and Jim Taylor from Tom Perrotta's 1998 novel of the same title. The plot revolves around a high school election and satirizes both suburban high school life and politics. The film stars Matthew Broderick as Jim McAllister, a popular high school social studies teacher in suburban Omaha, Nebraska, and Reese Witherspoon as Tracy Flick, around the time of the school's student body election. When Tracy qualifies to run for class president, McAllister believes she does not deserve the title and tries his best to stop her from winning. Election opened to acclaim from critics, who praised its writing and direction. **The film received an Academy Award nomination for Best Adapted Screenplay, a Golden Globe nomination for Witherspoon in the Best Actress category, and the Independent Spirit Award for Best Film in 1999.**
	Word level attractiveness	The film received an Academy **Award** nomination for **Best** Adapted Screenplay, a Golden Globe nomination for Witherspoon in the **Best** Actress category, and the Independent Spirit **Award** for **Best** Film in 1999
	Cast member attractiveness	Alexander Payne, **Reese Witherspoon**, Matthew Broderick, Jim Taylor

Interpretable Recommendation via Attraction Modeling: Learning Multilevel Attractiveness over Multimodal Movie Contents, IJCAI2018

Statistical attractiveness on movie ***Election (1999)*** w.r.t. sentences, words in the most attractive sentences and cast members. The larger size and deeper color of font denote the larger attractiveness weight is assigned.

Non-IID Behavior Analytics

More at KDD2018 Tutorial on Behavior Analytics

https://datasciences.org/behavior-informatics/

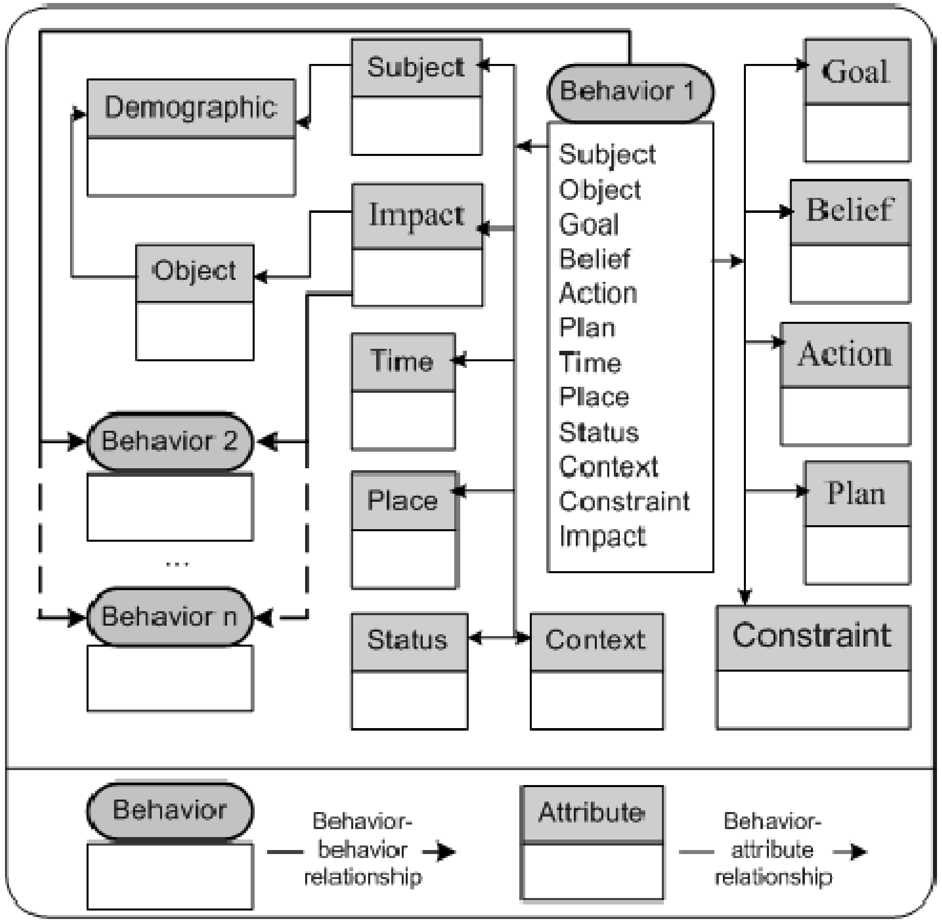
Behavior Model

Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, Information Science, 180(17); 3067-3085, 2010.

Examples of Coupled Objects and Behaviors

An Abstract Behavior Model: behavior computing

- An abstract behavior model
  - Demographics and circumstances of behavioral subjects and objects
  - Associates of a behavior may form into certain behavior sequences or network;
  - Social behavioral network consists of sequences of behaviors that are organized in terms of certain social relationships or norms.
  - Impact, costs, risk and trust of behavior/behavior network



Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, Information Science, 180(17); 3067-3085, 2010.

Behavior Vector & Couplings

- Behavior instance: behavior vector

$$\vec{\gamma} = \{s, o, e, g, b, a, l, f, c, t, w, u, m\}$$

  - basic properties
  - social and organizational factors
- Vector-based behavior sequences
- Vector-oriented behavior representation

$$\vec{\Gamma} = \{\vec{\gamma_1}, \vec{\gamma_2}, ..., \vec{\gamma_n}\}$$

- Behavior Coupling Relationships
  - ✓ Logic/semantic behavior couplings
  - ✓ Statistical/Probabilistic behavior couplings

Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, Information Science, 180(17); 3067-3085, 2010.

Group/Coupled Behavior Analysis

Yin Song, Longbing Cao, et al. Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation, KDD 2012, 976-984.

Yin Song and Longbing Cao. Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets, IJCNN 2012, 1-8.

Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Applications, IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).

Pool Manipulation

TABLE 1
An example of buy and sell orders

Investor	Time	Direction	Price	Volume
(1)	09:59:52	Sell	12.0	155
(2)	10:00:35	Buy	11.8	2000
(3)	10:00:56	Buy	11.8	150
(2)	10:01:23	Sell	11.9	200
(1)	10:01:38	Buy	11.8	200
(4)	10:01:47	Buy	11.9	200
(5)	10:02:02	Buy	11.9	250
(2)	10:02:04	Sell	11.9	500

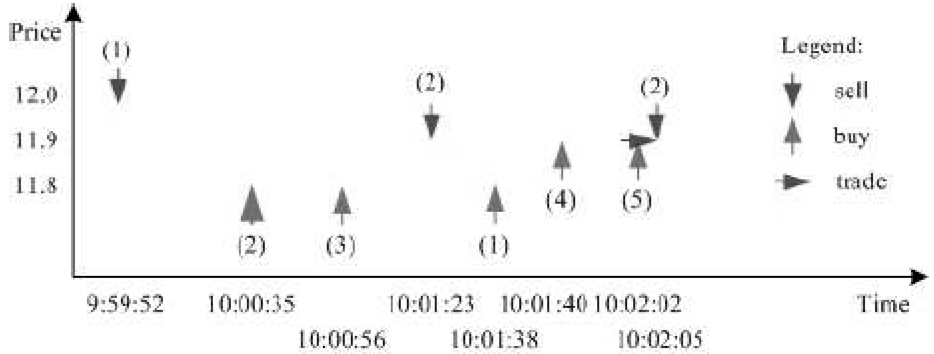


Fig. 1. Coupled Trading Behaviors

Behavior Formal Descriptor

We tackle the coupled behaviors from either one or different actors, denoted as intra-coupling and inter-coupling, respectively.

Behavior Feature Matrix

intra-coupling

$$FM(\mathbb{B}) = \begin{pmatrix} \mathscr{O}_{11} & \mathscr{O}_{12} & \cdots & \mathscr{O}_{1J_{max}} \\ \mathscr{O}_{21} & \mathscr{O}_{22} & \cdots & \mathscr{O}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathscr{O}_{I1} & \mathscr{O}_{I2} & \cdots & \mathscr{O}_{IJ_{max}} \end{pmatrix}$$

inter-coupling

An actor $\mathscr{A}_i$ undertakes J_i operations $\{\mathscr{O}_{i1}, \mathscr{O}_{i2}, \ldots, \mathscr{O}_{iJ_i}\}$

I actors: $\{\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_I\}$

Intra-Coupling

- The intra-coupling reveals the complex couplings within an actor's distinct behaviors.

Definition 2 (Intra-Coupled Behaviors): Actor $\mathscr{A}_i$'s behaviors $\mathbb{B}_{ij}$ $(1 \leq j \leq J_{max})$ are intra-coupled in terms of coupling function $\theta_j(\mathbb{B})$,

$$\mathbb{B}_{i\cdot}^{\theta} ::= \mathbb{B}_{i\cdot}(\mathscr{A}, \mathscr{O}, \theta) | \sum_{j=1}^{J_{max}} \theta_j(\mathbb{B}) \odot \mathbb{B}_{ij}, \qquad \text{(IV.2)}$$

where $\sum_{j=1}^{J_{max}} \odot$ means the subsequent behavior of $\mathbb{B}_i$ is $\mathbb{B}_{ij}$ intra-coupled with $\theta_j(\mathbb{B})$, and s

$$FM(\mathbb{B}) = \begin{pmatrix} \mathbb{B}_{11} & \mathbb{B}_{12} & \dots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \dots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \dots & \mathbb{B}_{IJ_{max}} \end{pmatrix}$$

For instance, in the stock market, the investor will place a sell order at some time after buying his or her desired instrument due to a great rise in the trading price. This is, to some extent, one way to express how these two behaviors are intra-coupled with each other.

Inter-Coupling

- **The inter-coupling embodies the way multiple behaviors of different actors interact.**

Definition 3 (Inter-Coupled Behaviors): Actor $\mathscr{A}_i$'s behaviors $\mathbb{B}_{ij}$ $(1 \leq i \leq I)$ are inter-coupled with each other in terms of coupling function $\eta_i(\mathbb{B})$,

$$\mathbb{B}^{\eta}_{\cdot j} ::= \mathbb{B}_{\cdot j}(\mathscr{A}, \mathscr{O}, \eta) | \sum_{i=1}^{I} \eta_i(\mathbb{B}) \odot \mathbb{B}_{ij}, \qquad \text{(IV.3)}$$

where $\sum_i^I \odot$ means the subsequent behavior of $\mathbb{B}_i$ is $\mathbb{B}_{ij}$ inter-coupled with $\eta_i(\mathbb{B})$, and so on.

$$FM(\mathbb{B}) = \left(\begin{array}{c|ccc} \mathbb{B}_{11} & \mathbb{B}_{12} & \cdots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \cdots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \cdots & \mathbb{B}_{IJ_{max}} \end{array}\right)$$

For instance, a trading happens successfully only when an investor sells the instrument at the same price as the other investor buys this instrument. This is another example of how to trigger the interactions between inter-coupled behaviors.

Coupling

- **In practice, behaviors may interact with one another in both ways of intra-coupling and inter-coupling.**

Definition 4 (Coupled Behaviors): Coupled behaviors $\mathbb{B}_c$ refer to behaviors $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$ that are coupled in terms of relationships $h(\theta(\mathbb{B}), \eta(\mathbb{B}))$, where $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$

$$\mathbb{B}_c = (\mathbb{B}^{\theta}_{i_1 j_1})^{\eta} * (\mathbb{B}^{\theta}_{i_2 j_2})^{\eta} ::= \mathbb{B}_{ij}(\mathscr{A}, \mathscr{O}, \mathscr{C}) | \sum_{i_1, i_2 = 1}^{I} \sum_{j_1, j_2 = 1}^{J_{max}} h(\theta_{j_1 j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B})) \odot (\mathbb{B}_{i_1 j_1} \mathbb{B}_{i_2 j_2}), \tag{IV.4}$$

where $h(\theta_{j_1, j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B}))$ is the coupling function denoting the corresponding relationships between $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$, $\sum_{i_1, i_2 = 1}^{I} \sum_{j_1, j_2 = 1}^{J_{max}} \odot$ means the subsequent behaviors of $\mathbb{B}$ are $\mathbb{B}_{i_1 j_1}$ coupled with $h(\theta_{j_1}(\mathbb{B}), \eta_{i_1}(\mathbb{B}))$, $\mathbb{B}_{i_2 j_2}$ with $h(\theta_{j_2}(\mathbb{B}), \eta_{i_2}(\mathbb{B}))$, and so on.

For instance, we consider both the successful trading between investor A_1 (buy) and investor A_2 (sell), and then the selling behavior conducted by A_1 after he or she has bought the instrument at a relative low price.

Coupled Behavior Analysis (CBA)

Theorem 1. *(Coupled Behavior Analysis (CBA)) The analysis of coupled behaviors* (CBA Problem *for short) is to build the objective function $g(\cdot)$ under the condition that behaviors are coupled with each other by coupling function $f(\cdot)$, and satisfy the following conditions.*

$$f(\cdot) ::= f(\theta(\cdot), \eta(\cdot)), \tag{9}$$

$$g(\cdot)|(f(\cdot) \geq f_0) \geq g_0 \tag{10}$$

TABLE 1
An example of buy and sell orders

Investor	Time	Direction	Price	Volume
(1)	09:59:52	Sell	12.0	155
(2)	10:00:35	Buy	11.8	2000
(3)	10:00:56	Buy	11.8	150
(2)	10:01:23	Sell	11.9	200
(1)	10:01:38	Buy	11.8	200
(4)	10:01:47	Buy	11.9	200
(5)	10:02:02	Buy	11.9	250
(2)	10:02:04	Sell	11.9	500

CHMM-based Coupled Sequence Modeling

- Coupled behavior sequences
  - Multiple sequences

$$\Phi_1 = \{\phi_{11}, \ldots, \phi_{1T}\}$$
$$\Phi_2 = \{\phi_{21}, \ldots, \phi_{2F}\}$$
$$\Phi_C = \{\phi_{C1}, \ldots, \phi_{CG}\}$$

  - Coupling relationship

$$R_{ij}(\Phi_i, \Phi_j)$$
$$R_{ij} \subset R, R_{ij}(\Phi_i, \Phi_j) = \varnothing$$

  - Behavior properties

$$\phi_{ik}(p_{ik,1}, \ldots, p_{ik,L})$$

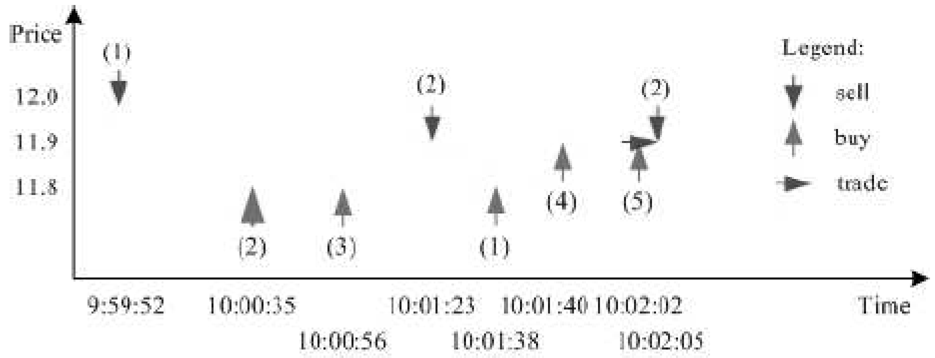
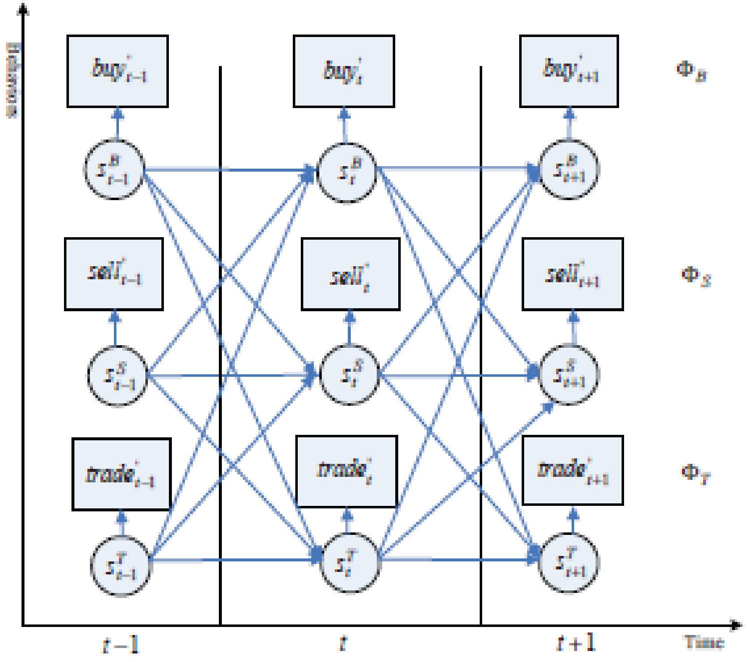


Fig. 1. Coupled Trading Behaviors

CBA – CHMM



(b) The Structure of the CHMM

$$CBA\ problem \rightarrow CHMM\ model \quad (15)$$

$$\Phi(\mathbb{B}_c)|category \rightarrow X \quad (16)$$

$$M(\Phi(\mathbb{B}_c))|\phi_{ik}([p_{ij}]_1, \ldots, [p_{ij}]_K) \rightarrow Y \quad (17)$$

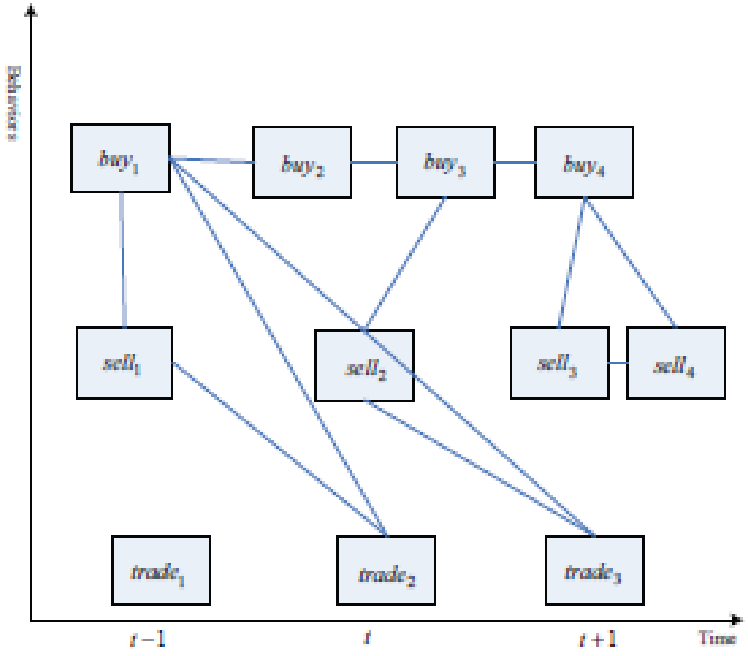
$$f(\theta(\cdot), \eta(\cdot)) \rightarrow Z \quad (18)$$

$$Initial\ distribution\ of\ \Phi(\mathbb{B}_c)|category \rightarrow \pi \quad (19)$$

- Wei Cao, Liang Hu, Longbing Cao. Deep Modeling Complex Couplings within Financial Markets, AAAI2015, 2518-2524.
- Wei Cao, Longbing Cao, Yin Song. Coupled Market Behavior Based Financial Crisis Detection, IJCNN2013
- Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Applications, IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).
- Longbing Cao, Yuming Ou, Philip S YU, Gang Wei. Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors, KDD2010, 85-94

Graph-based Coupled Behavior Presentation

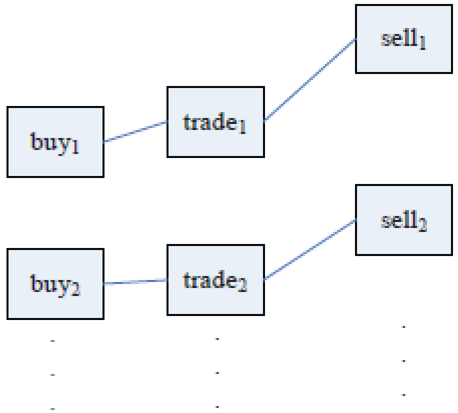
- Coupled hidden Markov Model (CHMM)
- Relational probability tree (RPT)
- Relational Bayesian Classifier (RBC)



(c) The Structure of Graph-based Coupled Behavior Model

- Yin Song, Longbing Cao, et al. [Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation](), KDD 2012, 976-984.
- Yin Song and Longbing Cao. [Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets](), IJCNN 2012, 1-8

CBA - Conditional Probability Distribution



(a) An Example of the Subgraphs for Each Target Behavior

	$X^{(t)}$	RF_1	RF_2	$\cdots$	RF_n
$trade_1$	x_1	rf_{11}	rf_{21}	$\cdots$	rf_{n1}
$trade_2$	x_2	rf_{12}	rf_{22}	$\cdots$	rf_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(b) An Example of the Relational Features for Each Target Behavior

$$CBA\ problem \rightarrow SRL\ Modeling \qquad (5)$$

$$f(\theta(\cdot),\eta(\cdot)) \rightarrow the\ CPD\ p(X^{(t)}|RF_1,\cdots,RF_n) \qquad (6)$$

$$p(X^{(t)}|RF_1,RF_2,\cdots,RF_n)$$

$$CL(\mathbf{b^k}) = \prod_{\mathbf{b_i^{(t)}}\in\mathbf{b^k}} p(X^{(t)} = x_{b_i^{(t)}}|rf_{1i},rf_{2i},\cdots,rf_{ni};M)$$

- Yin Song, Longbing Cao, et al. Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation, KDD 2012, 976-984.

Empirical Results

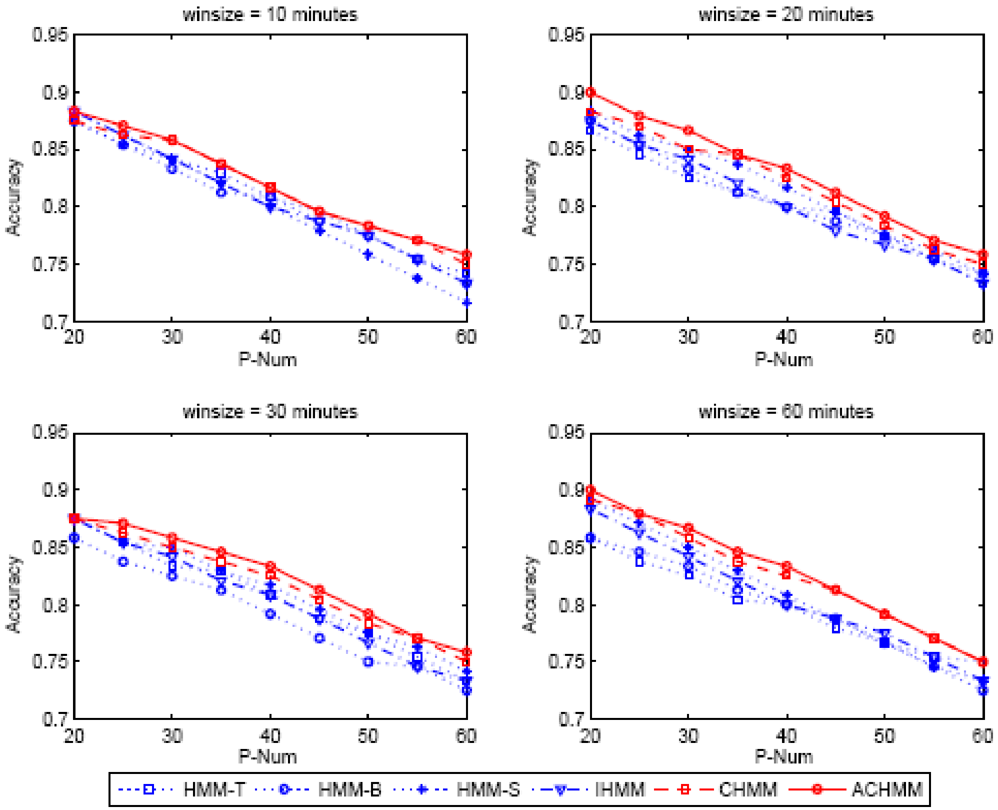


Figure 4: Accuracy of Six Models

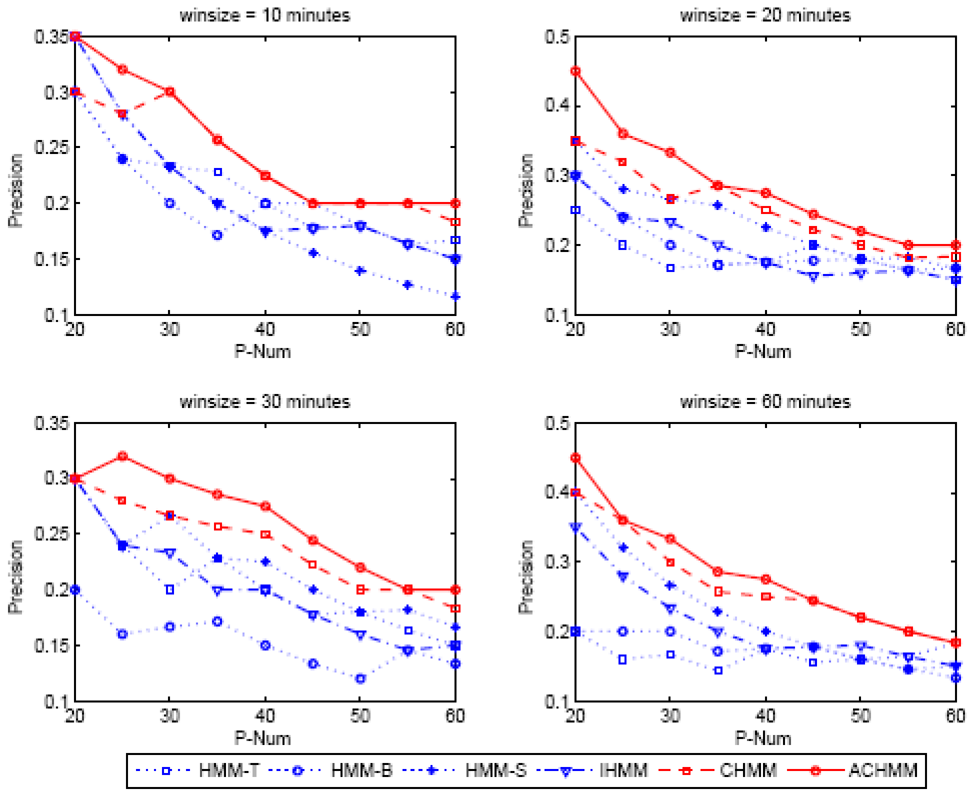


Figure 5: Precision of Six Models

Next-best Action Recommendation with multi-party interactions

Longbing Cao, Chengzhang Zhu. Personalized next-best action recommendation with multi-party interaction learning for automated decision-making, PLoS ONE, 17(1): e0263010, 2022

The NBA problem

- NBA-based personalized decision-making process

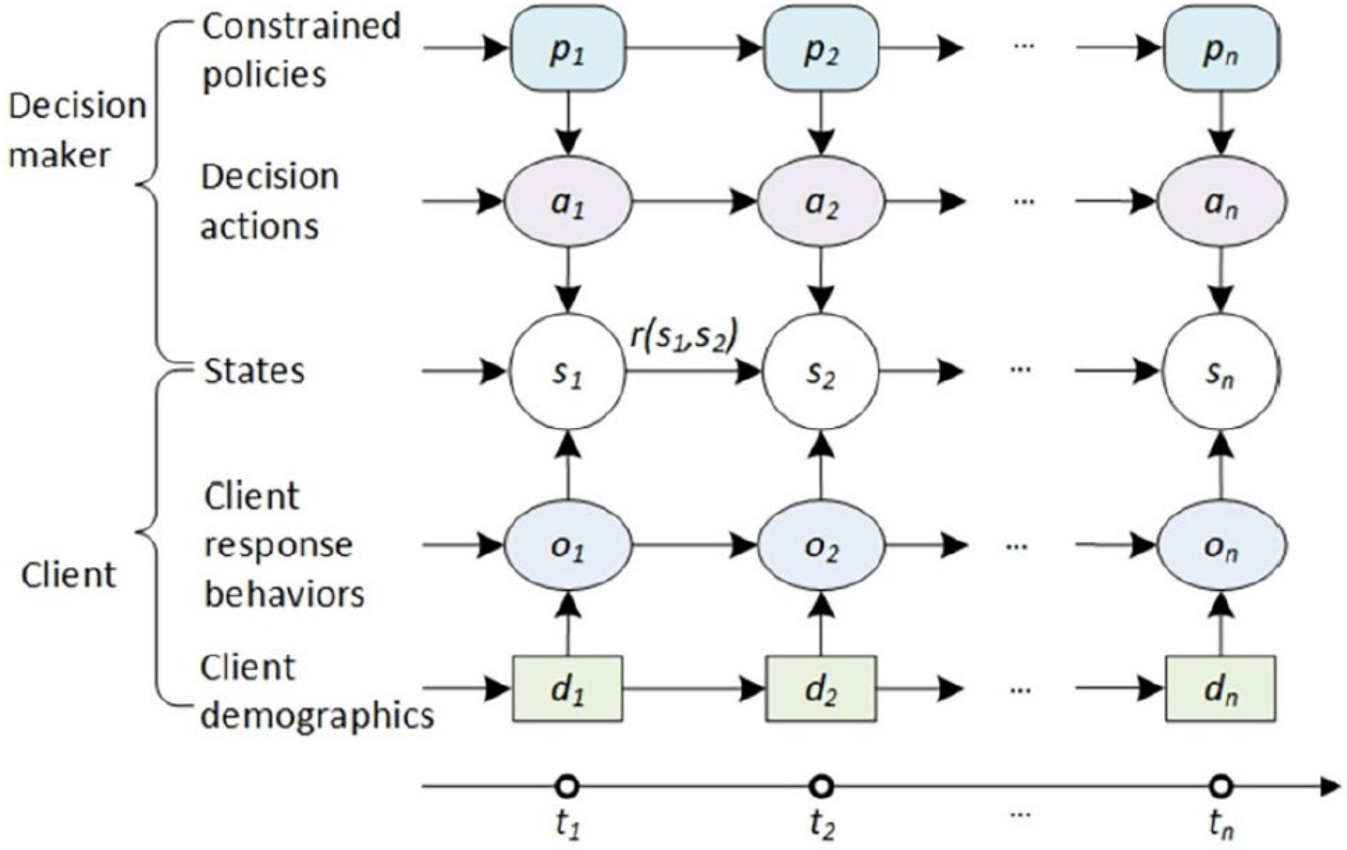


Fig 1. Next-best action-based personalized decision-making in constrained, tailored, sequential and interactive dynamic processes with state-action-response-coupled sequences.

The NBA problem

- NBA objective function

$$\underset{\{a_t^j | j=1,\cdots,k\}}{\text{minimize}} \quad Div(\hat{\mathcal{R}} || \mathcal{R}) - \sum_{j=1}^{k} r_{\boldsymbol{\theta}}(C_t, a_t^j)$$

$$\text{subject to} \quad a_t^j \in A^*,$$

where $Div(\cdot||\cdot)$ is the divergence between the estimated reward space $\hat{\mathcal{R}}$ and the actual reward space $\mathcal{R}$, and $\boldsymbol{\theta}$ refers to the parameters in the action-value function $r_{\boldsymbol{\theta}}(\cdot, \cdot)$.

action-value function $r_{\theta}(\cdot, \cdot) : \mathcal{C} \times \mathcal{A} \rightarrow \hat{\mathcal{R}}$

k next-best action set

The NBA problem

- Learn multi-party past-to-present interactions and decision-making

NBA action-value function

$$r_\theta(\cdot,\cdot) : \mathcal{C} \times \mathcal{A} \rightarrow \hat{\mathcal{R}}$$

client descriptions C_t
decision-making actions A_{t-1}
and estimated rewards

$$C_t = < D_t, A_{t-1}, O_t >$$

RL action-value function

$$r_\theta(\cdot,\cdot) : \mathcal{O} \times \mathcal{A} \rightarrow \mathcal{R}$$

decision actions a_t

client responses $O_{t,t}$

The NBA problem

- Personalized NBA set

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{j=1}^{n_c} \sum_{i=1}^{t^{(j)}} l(r_{\boldsymbol{\theta}}(C_i^{(j)}, a_i^{(j)}), r_{<C_i^{(j)}, a_i^{(j)}>})$$

$l(\cdot, \cdot)$ a loss function that measures the difference between the real and estimated rewards

$C_i^{(j)}$ description of the *j*-th client at time step *i*

$a_i^{(j)}$ historical decision action on the *j*-th client at time step *i*

$t^{(j)}$ maximal length of historical sequence of the *j*-th client

The NBA problem

- Personalized Next-k Best Action/NBA

$$\begin{aligned} \underset{\{a_t^j | j=1,\cdots,k\}}{\text{maximize}} \quad & \sum_{j=1}^{k} r_{\boldsymbol{\theta}}(C_t, a_t^j) \\ \text{subject to} \quad & a_t^j \in A_t^*. \end{aligned}$$

$$\hat{A}_t^* = \{a_t^j | j = 1, \cdots, k\}$$

A_t^* candidate action set

PNBA learning framework

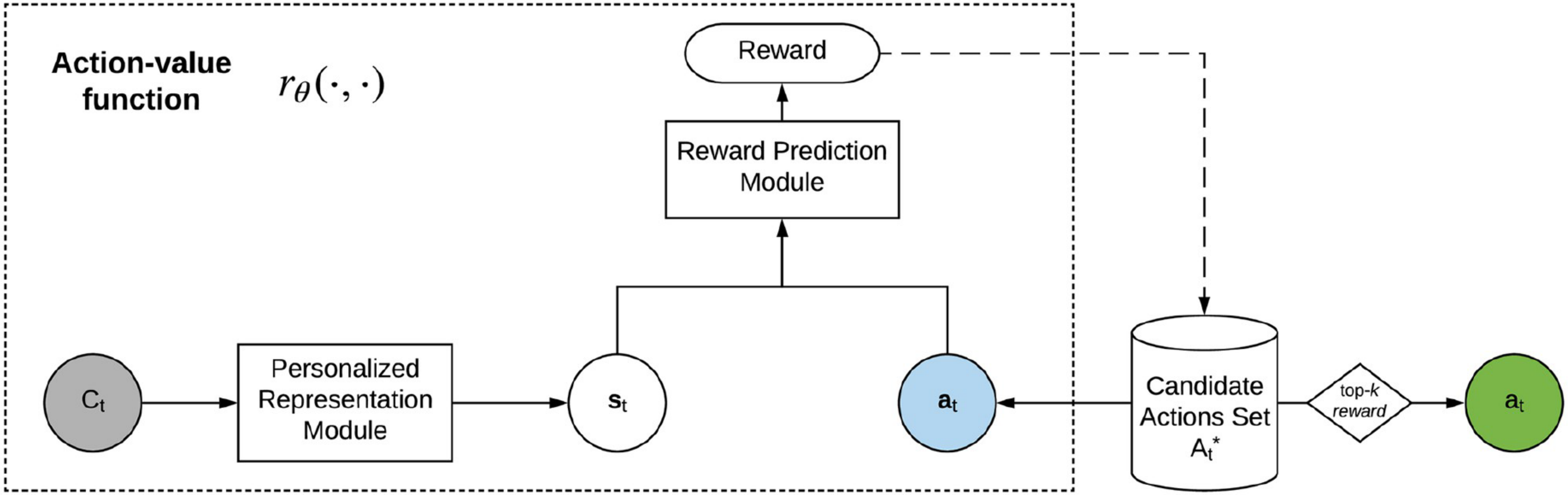


Fig 2. The framework for modeling the next-best action-oriented personalized decision-making.

Learn personalized client representation

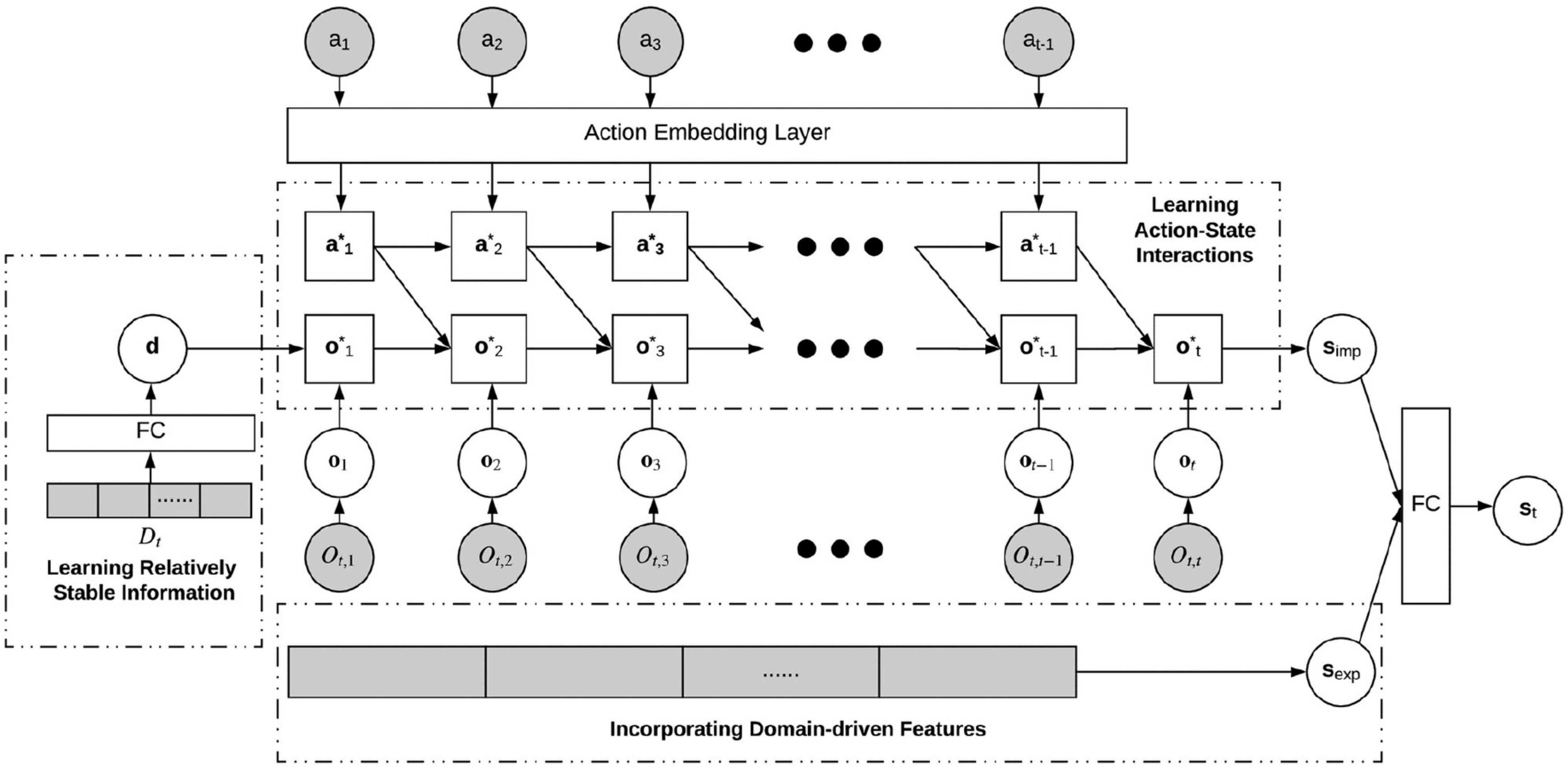
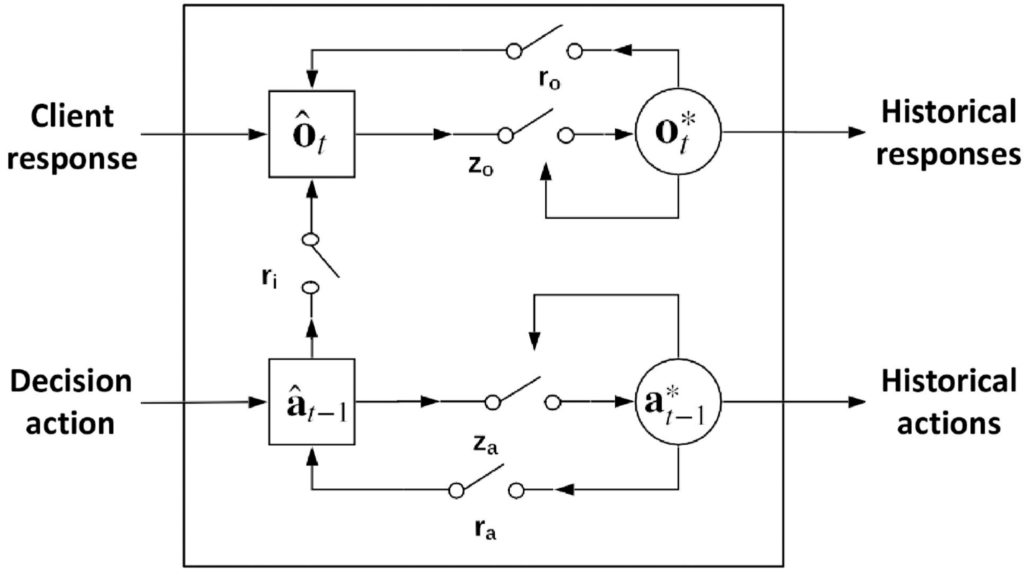


Fig 3. A reinforced coupled recurrent network to learn personalized client representation.

Learn state-action-response couplings



$$\mathbf{z}_a = \sigma(\mathbf{W}_{z_a}\mathbf{a}_{t-1} + \mathbf{U}_{z_a}\mathbf{a}^*_{t-2})$$

$$\mathbf{r}_a = \sigma(\mathbf{W}_{r_a}\mathbf{a}_{t-1} + \mathbf{U}_{r_a}\mathbf{a}^*_{t-2})$$

$$\hat{\mathbf{a}}_{t-1} = tanh(\mathbf{W}_a\mathbf{a}_{t-1} + \mathbf{U}_a(\mathbf{r}_a \circ \mathbf{a}^*_{t-2}))$$

$$\mathbf{a}^*_{t-1} = (\mathbf{1}_a - \mathbf{z}_a) \circ \mathbf{a}^*_{t-2} + \mathbf{z}_a \circ \hat{\mathbf{a}}_{t-1}$$

$$\mathbf{z}_o = \sigma(\mathbf{W}_{z_o}\mathbf{o}_t + \mathbf{U}_{z_o}\mathbf{o}^*_{t-1})$$

$$\mathbf{r}_o = \sigma(\mathbf{W}_{r_o}\mathbf{o}_t + \mathbf{U}_{r_o}\mathbf{o}^*_{t-1})$$

$$\hat{\mathbf{o}}_t = tanh(\mathbf{W}_o\mathbf{o}_t + \mathbf{U}_o(\mathbf{r}_o \circ \mathbf{o}^*_{t-1}) + \mathbf{I}_o(\mathbf{r}_i \circ \hat{\mathbf{a}}_{t-1}))$$

$$\mathbf{o}^*_t = (\mathbf{1}_o - \mathbf{z}_o) \circ \mathbf{o}^*_{t-1} + \mathbf{z}_o \circ \hat{\mathbf{o}}_t$$

$$\mathbf{r}_i = \sigma(\mathbf{W}_i\mathbf{a}_{t-1} + \mathbf{U}_i\mathbf{o}^*_{t-1})$$

$\boldsymbol{r_{o/a}}$: historical responses and actions on their current states
z: current response and action states on history
$\boldsymbol{r_i}$: interaction between decision action and client response

Fig 4. A coupled recurrent unit (CRU) for modeling state-action-response-coupled long-term dependencies.

Learn client representations

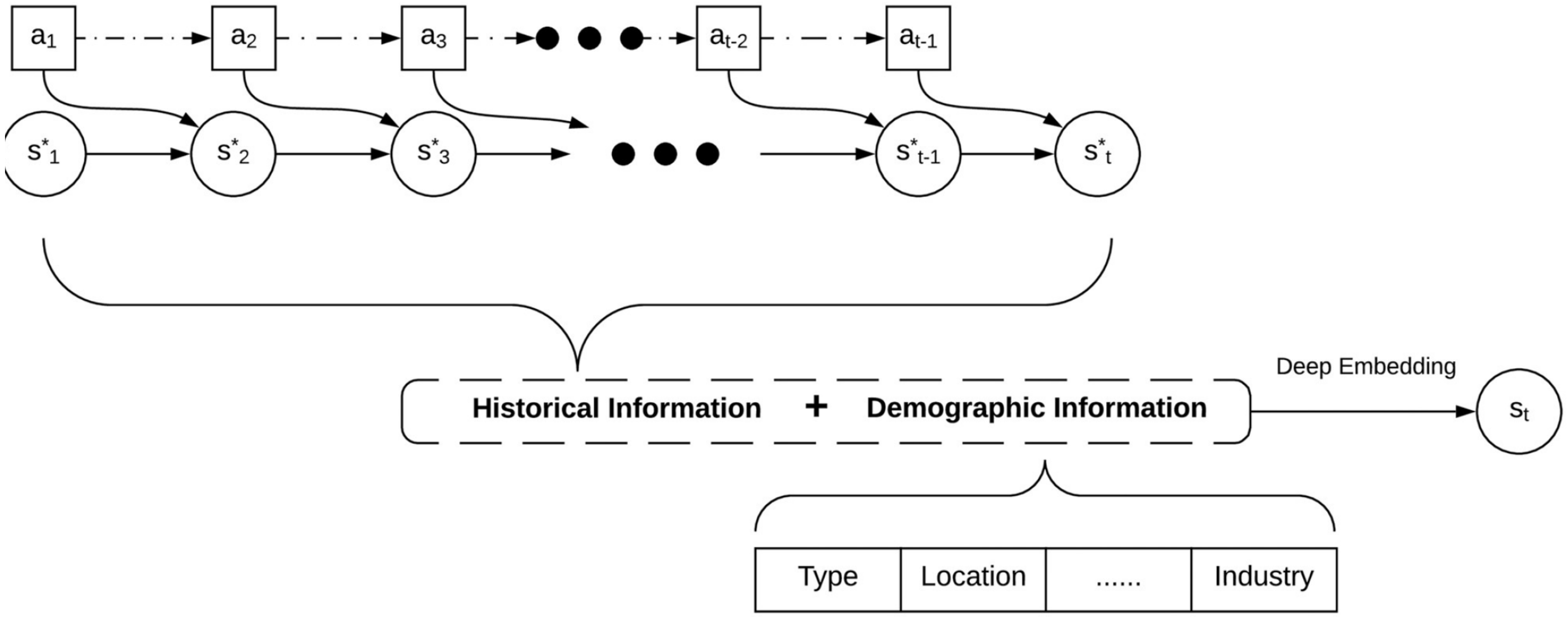
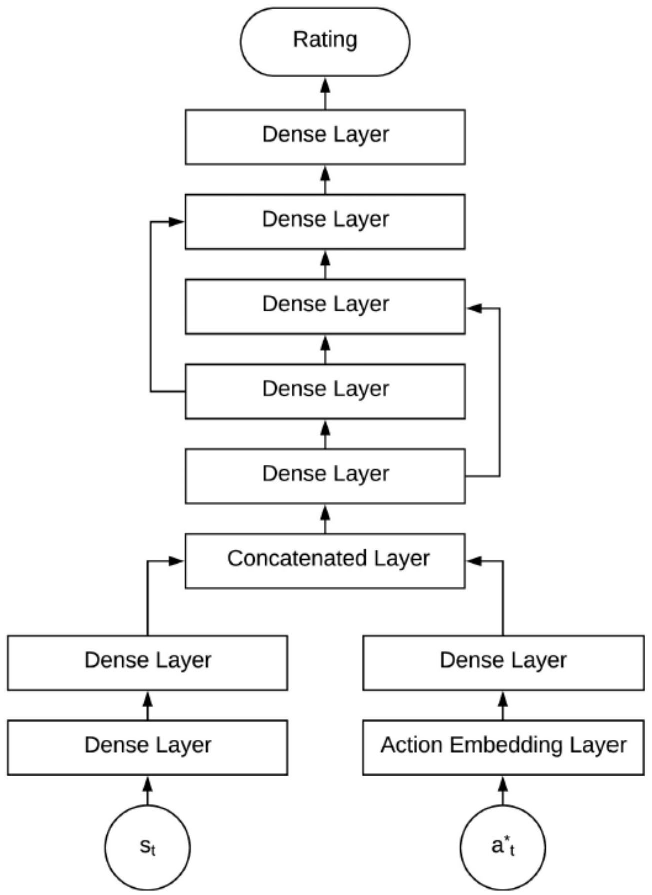


Fig 5. An example of representing clients by the reinforced coupled recurrent network.

NBA reward prediction



client state vector $C_t \rightarrow \mathbf{s}_t$

each decision action $a_t^j \rightarrow \mathbf{a}_t^j \in A_t^*$

action rating $r_{\theta}(C_t, a_t^j)$

next-best actions $\hat{A}_t^* \subseteq A_t^*$

Fig 6. Reward prediction for the next-best action on a client's state.

Case studies

- Non-Markovian NBA recommendation

Table 2. Average reward lift for 10 actions recommended by 11 deep models over the review measured by domain-driven debt collection rules.

Model	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Total_Avg	Action_Avg
CRN_IMB	5	4	3.0534	2.8752	6.8	2.1415	2.6984	3.3567	1.6772	2.9969	2.5569	3.4599
CRN	2.1957	3.5383	2.2068	2.6616	3.216	2.074	2.326	2.6277	1.7654	2.3425	2.1942	2.4954
WD	2.604	1.5992	2.0979	2.2798	3.2239	1.9824	2.2629	2.6967	0.9899	2.312	2.1089	2.2049
LSTM	0.9722	1.0987	0.9391	0.974	1.1272	1.0159	0.897	1.1097	1.1024	1.0847	1.0013	1.0321
WD_LSTM	2.0471	1.2731	1.9709	2.4755	2.2217	1.8129	2.0816	2.1909	1.1405	2.105	1.9198	1.9319
WD_Res_LSTM	1.7247	0.8219	1.7007	1.9816	2.4985	1.8164	1.9851	2.0921	0.8285	1.967	1.8488	1.7416
WD_Multi_LSTM	1.684	1.0468	1.6591	1.774	1.6924	1.7083	1.671	2.1678	1.2222	1.8098	1.7161	1.6435
GRU	0.5783	0.0865	0.9852	1.1201	1.5022	0.9154	0.861	0.9463	1.0347	1.0416	0.9345	0.9071
WD_GRU	1.0049	0.6397	1.3454	1.7369	2.1271	1.6489	1.6049	2.1562	0.665	1.6602	1.611	1.4589
WD_Res_GRU	1.4488	1.1333	1.7364	1.3479	2.2259	1.6932	1.7091	1.9582	1.2507	1.8869	1.7248	1.6391
WD_Multi_GRU	1.6329	1.8399	1.9114	1.7949	1.8781	1.8206	2.0276	1.7613	1.0508	2.2347	1.8959	1.7952
Δ_IMB	92.01%	117.40%	45.55%	16.15%	110.92%	8.03%	19.25%	24.47%	34.10%	29.62%	21.24%	56.92%
Δ	-15.68%	92.31%	5.19%	7.52%	-0.25%	4.62%	2.79%	-2.56%	41.15%	1.32%	4.04%	13.18%

Case studies

- Non-Markovian NBA recommendation

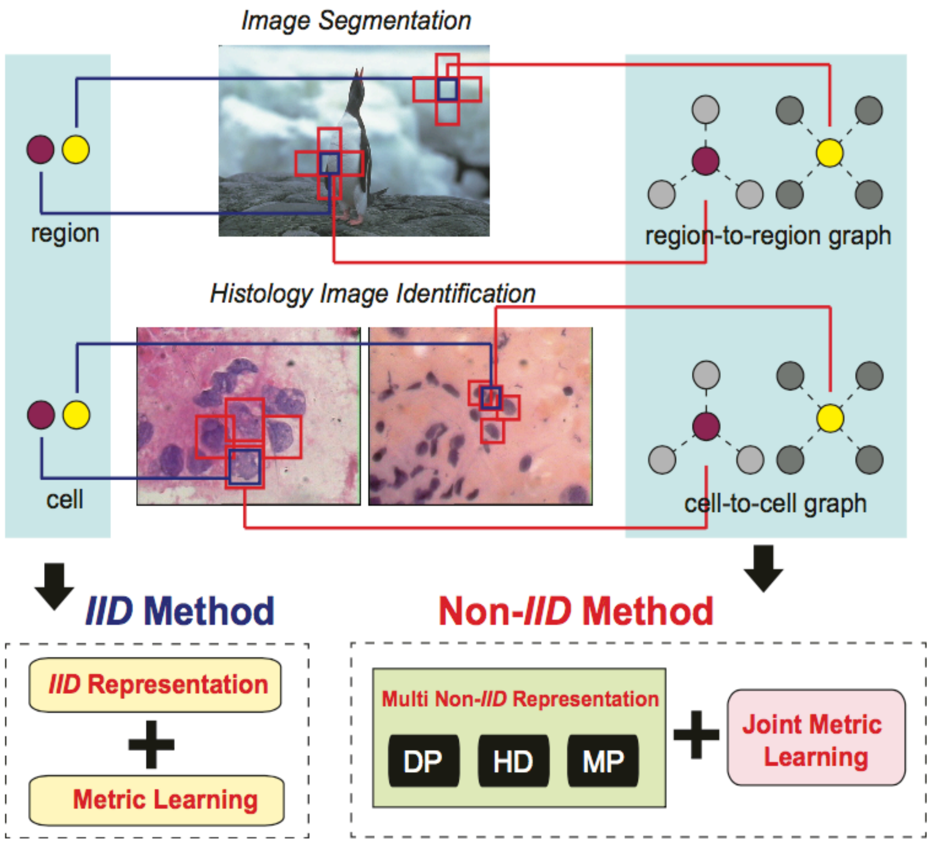
Table 4. The reward mean squared error (MSE) per action between the reward made by the domain-driven debt collection rules and that recommended by 10 deep models.

Model	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Total_Avg	Action_Avg
CRN	0.0266	0.055	0.0462	0.094	0.0222	0.0937	0.0733	0.0384	0.1077	0.056	0.0777	0.0613
WD	0.0271	0.0631	0.0491	0.1038	0.0263	0.0963	0.076	0.0384	0.1245	0.0565	0.0803	0.0661
LSTM	0.1219	0.1315	0.1129	0.1411	0.1286	0.131	0.1201	0.1216	0.1256	0.1166	0.1253	0.1251
WD_LSTM	0.2361	0.2395	0.2167	0.2188	0.2539	0.2163	0.2146	0.2352	0.1757	0.2108	0.2165	0.2218
WD_Res_LSTM	0.2188	0.2333	0.2187	0.2128	0.2363	0.2091	0.2078	0.2192	0.1776	0.2099	0.2108	0.2143
WD_Multi_LSTM	0.2429	0.2485	0.2203	0.2215	0.2616	0.2177	0.2161	0.2417	0.177	0.212	0.2185	0.2259
GRU	0.1011	0.1139	0.0957	0.1324	0.1035	0.1215	0.1076	0.103	0.1243	0.1021	0.1134	0.1105
WD_GRU	0.2299	0.2368	0.2211	0.2174	0.2417	0.213	0.2106	0.2261	0.1798	0.2174	0.2149	0.2194
WD_Res_GRU	0.2301	0.2384	0.2245	0.2168	0.2493	0.2142	0.2119	0.2304	0.1777	0.2156	0.2162	0.2209
WD_Multi_GRU	0.228	0.2354	0.2196	0.2195	0.2443	0.2157	0.2131	0.2279	0.1795	0.2136	0.2162	0.2197
Δ	1.85%	12.84%	5.91%	9.44%	15.59%	2.70%	3.55%	0.00%	13.35%	0.88%	3.24%	7.26%

Non-IID Vision Learning

Yinghuan Shi, Wenbin Li, Yang Gao, Longbing Cao, Dinggang Shen. Beyond IID: Learning to Combine Non-IID Metrics for Vision Tasks. AAAI2017.

Non-IID Metric Learning



- ❑ Three phases:
  - ✓ (non-*IID*) features
  - ✓ various non-*IID* representations
  - ✓ joint metric learning

★ Good adaptation with the best combination automatically learned

★ Easy to implement

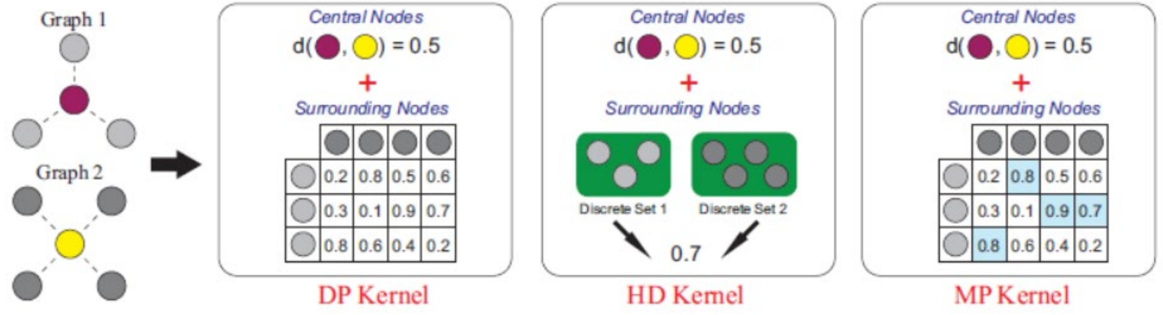
★ Many features, representations and classifiers can be integrated

Various Non-IID Representations

- **Core Idea:**

  Intra-node relation (within node) + Inter-node relations (between neighbored nodes)

- Capturing various data characteristics
  - ✓ Direct Product (DP)
  - ✓ Hausdorff Distance (HD)
  - ✓ Max Pooling (MP)



$$\mathbf{K}_{\mathrm{DP}}(i,j) = \underbrace{\boxed{f(\mathbf{x}_i,\mathbf{x}_j)}}_{\text{intra}} + \underbrace{\boxed{\frac{1}{m_i \cdot m_j}\sum_{p=1}^{m_i}\sum_{q=1}^{m_j} f(\tilde{\mathbf{x}}_{i,p},\tilde{\mathbf{x}}_{j,q}).}}_{\text{inter}}$$

$$\mathbf{K}_{\mathrm{HD}}(i,j) = \underbrace{\boxed{f(\mathbf{x}_i,\mathbf{x}_j)}}_{\text{intra}} + \underbrace{\boxed{\frac{1}{m_i \cdot m_j} h(\mathcal{X}_i,\mathcal{X}_j).}}_{\text{inter}}$$

$$\mathbf{K}_{\mathrm{MP}}(i,j) = \underbrace{\boxed{f(\mathbf{x}_i,\mathbf{x}_j)}}_{\text{intra}} + \underbrace{\boxed{\frac{1}{m_i}\sum_{p=1}^{m_i}\max_{q=1,\dots,m_j} f(\tilde{\mathbf{x}}_{i,p},\tilde{\mathbf{x}}_{j,q})}}_{\text{inter}}$$

$$+ \underbrace{\boxed{\frac{1}{m_j}\sum_{q=1}^{m_j}\max_{p=1,\dots,m_i} f(\tilde{\mathbf{x}}_{i,p},\tilde{\mathbf{x}}_{j,q}).}}_{\text{inter}}$$

Learning/combining Multiple Non-IID Representations

Objective function for combined non-IID metrics

$$\arg\min_{\boldsymbol{\Omega}, w^p} \mathcal{E}(\boldsymbol{\Omega}; \sum_p w^p \mathbf{K}^p) \quad \text{s.t.} \sum_p w^p = 1, w^p \geq 0$$

$$\arg\min_{w^p} \sum_{i,j} \psi_{ij} \|\boldsymbol{\Omega}\Big(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_j^p\Big)\|^2 +$$

Pair-wise Constraint

$$\lambda \sum_{i,j,l} \psi_{ij}(1 - y_{il}) h\Big[\|\boldsymbol{\Omega}\Big(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_j^p\Big)\|^2$$

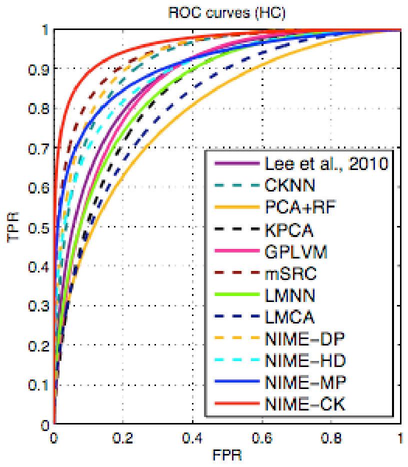
$$- \|\boldsymbol{\Omega}\Big(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_l^p\Big)\|^2 + 1\Big].$$

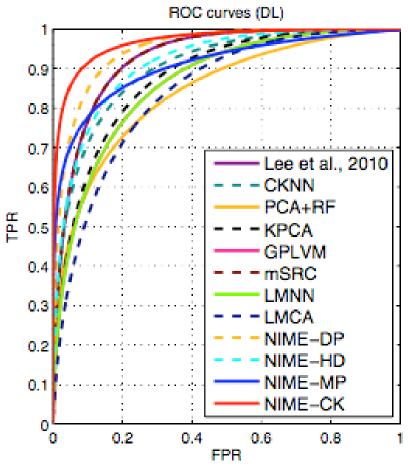
Triplet Constraint

$$\text{s.t.} \sum_p w^p = 1, w^p \geq 0.$$

Evaluation

Our methods outperform others in terms of AUC, Accuracy, Specificity, Sensitivity, F1 score





Method	(Lee 2010)	CKNN	PCA+RF	KPCA	GPLVM	mSRC	LMNN	LMCA	NIME-DP	NIME-HD	NIME-MP	NIME-MK
AC_{HC}	82.0	85.0	79.0	75.0	81.0	87.0	80.0	77.0	86.0	83.0	84.0	**89.0**
SP_{HC}	80.8	83.0	76.4	76.6	78.2	87.8	78.9	76.5	84.6	85.1	88.6	**91.5**
SE_{HC}	83.3	87.2	82.2	73.6	84.4	86.3	81.3	77.6	**87.5**	81.1	80.4	86.8
$F1_{HC}$	81.6	84.5	77.9	75.7	80.0	87.1	79.6	76.8	85.7	83.5	84.9	**89.3**
AUC_{HC}	87.9	91.6	84.2	79.1	86.8	93.8	85.3	81.6	92.7	89.1	90.6	**96.0**
AC_{DL}	86.0	84.0	82.0	79.0	81.0	86.0	81.0	79.0	88.0	85.0	84.0	**90.0**
SP_{DL}	89.1	84.0	83.3	76.4	81.6	89.1	81.6	80.9	**89.6**	85.7	79.3	88.5
SE_{DL}	83.3	84.0	80.8	82.2	80.4	83.3	80.4	77.4	86.6	84.3	90.5	**91.7**
$F1_{DL}$	86.5	84.0	82.4	77.9	81.2	86.5	81.2	79.6	88.2	85.2	82.6	**89.8**
AUC_{DL}	92.8	90.3	87.9	84.2	86.6	92.8	86.6	84.1	95.0	91.5	90.8	**96.9**

Image Segmentation

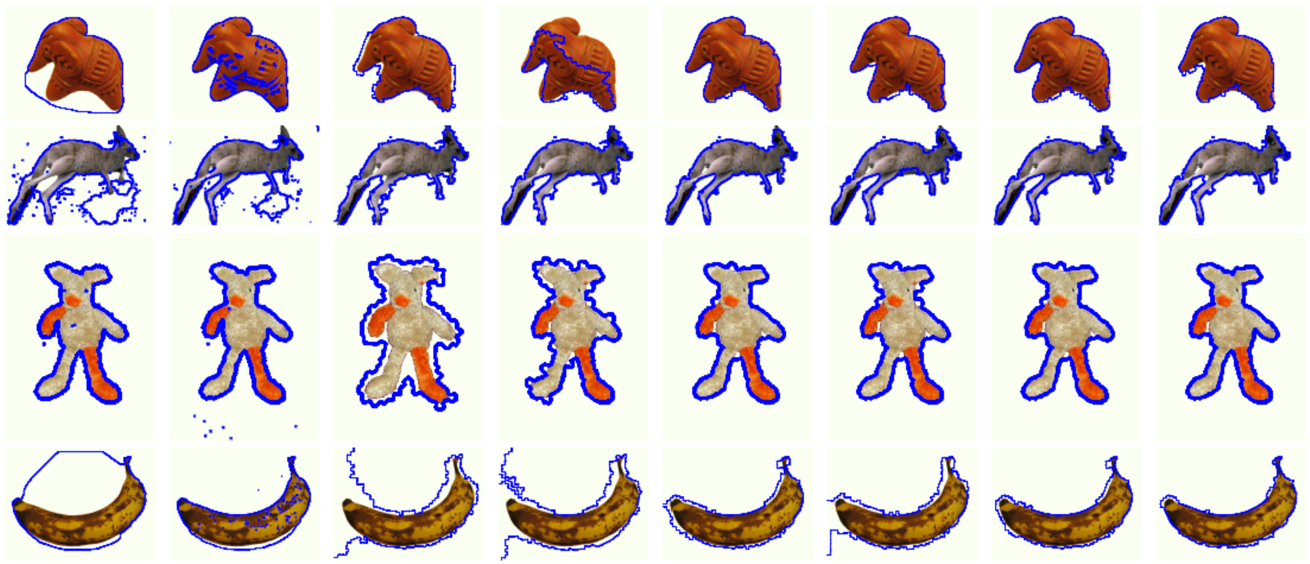


Figure 4: *Typical results. First to last columns: Graph Cut, Grab Cut, LMNN, LMCA, NIME-DP, NIME-HD, NIME-MP, NIME-CK.*

Non-IID Outlier Detection

Guansong Pang, Longbing Cao and Ling Chen. Homophily outlier detection in non-IID categorical data, Data Min. Knowl. Discov. 35(4): 1163-1224, 2021

Guansong Pang, Longbing Cao, Ling Cheny and Huan Liu. Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection. IJCAI2017

Guansong Pang, Hongzuo Xu, Longbing Cao and Wentao Zhao. Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data. CIKM2017

Multidimensional Data

- Multidimensional data
  - Data objects are characterized by two or more features
  - Information table
    - Rows -- data objects
    - Columns -- features

agegrp	density	Hispanic	bmi	count	cancer
0.888889	0.333333	0	0.333333	0.000517	0
0.888889	0.333333	0	0	0.000259	0
0.333333	0.333333	0	1	0.000517	0
0.777778	0.333333	0	0	0	0
0.888889	0	0	0	0	0
0.111111	0.333333	0	0	0	0
0.222222	0.666667	1	0.333333	0	0
0.333333	1	0	0	0	0
0.222222	0.666667	0	0.333333	0	0
0.222222	1	1	0	0	0

Traditional Outlier Detection

- Statistical/probabilistic-based approach
  - Statistical test-based –> *deviation from distribution*
  - Depth-based –> *data depth*
  - Deviation-based –> *sensitivity or uncertainty*
- Proximity-based approach
  - Distance-based –> *nearest neighbor distances*
  - Density-based –> *local density*
  - Clustering-based –> *distance to cluster centers*

Kriegel, H. P., Kröger, P., & Zimek, A. (2010). Outlier detection techniques. *Tutorial at KDD10*.

Aggarwal, C. C. (2017). Outlier analysis. Springer.

The IID Assumption

- Common assumptions
  - Values/features/objects from **homogeneous** distributions, mechanisms
  - They are **independent** to each other
    - E.g., implicit IID assumption in **Euclidean distance**

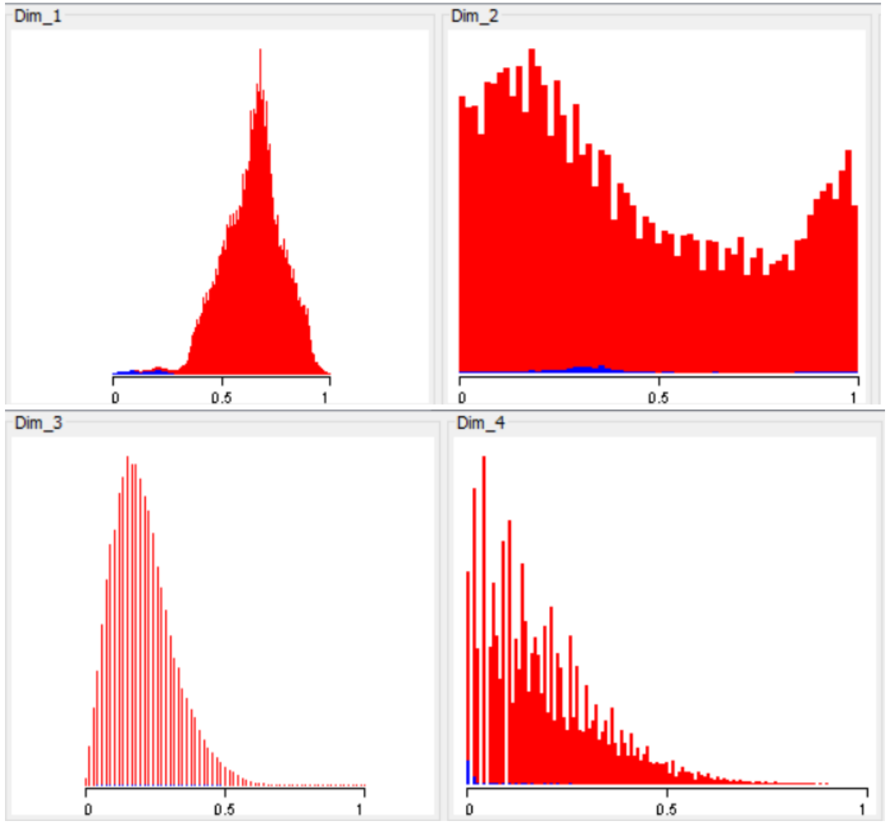
agegrp	density	Hispanic	bmi	count	cancer
0.888889	0.333333	0	0.333333	0.000517	0
0.888889	0.333333	0	0	0.000259	0
0.333333	0.333333	0	1	0.000517	0
0.777778	0.333333	0	0	0	0
0.888889	0	0	0	0	0
0.111111	0.333333	0	0	0	0
0.222222	0.666667	1	0.333333	0	0
0.333333	1	0	0	0	0
0.222222	0.666667	0	0.333333	0	0
0.222222	1	1	0	0	0

Non-IID Real-life Data

Couplings

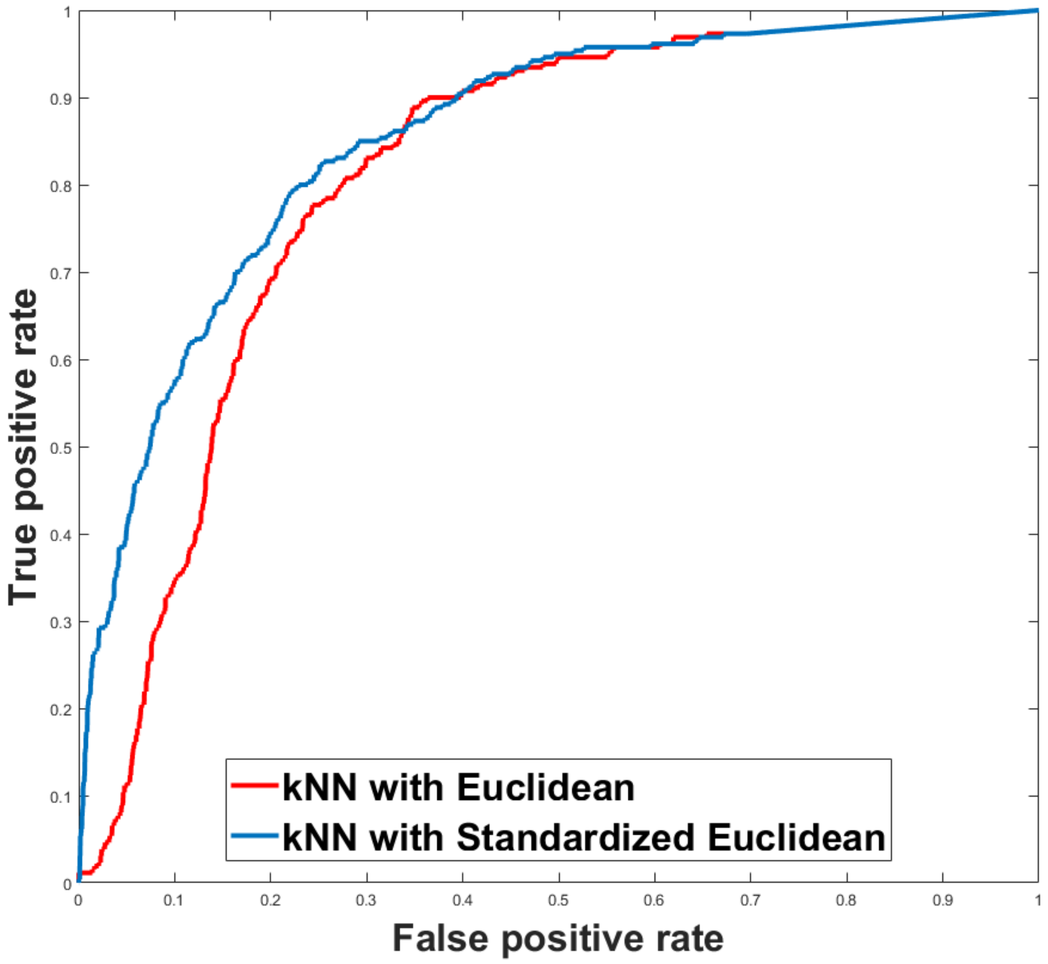
Source: http://www.diabeticrockstar.com

Heterogeneity



Four features from the *CoverType* data set

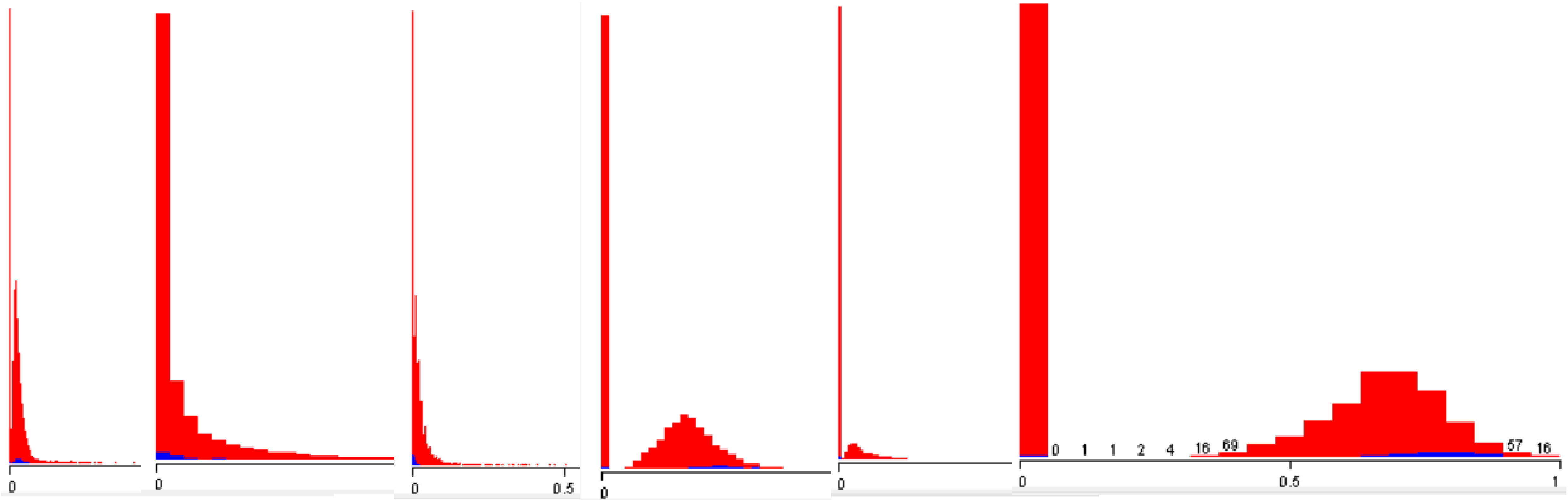
IID vs. Non-IID Outlier Detection – example



- ***Data: Mammography***
- Euclidean - AUC: 0.81
- Standardized Euclidean - AUC: 0.86

6.17% improvement

The *Mammography* Data Set



Non-IID Value-based Approach

Guansong Pang, Longbing Cao, Ling Chen. Outlier Detection in Complex Categorical Data by Modelling the Feature Value Couplings. IJCAI2016

Motivation

- **Value heterogeneity**
  - Semantic differs in different contexts

Values of the same frequency may indicate different outlierness

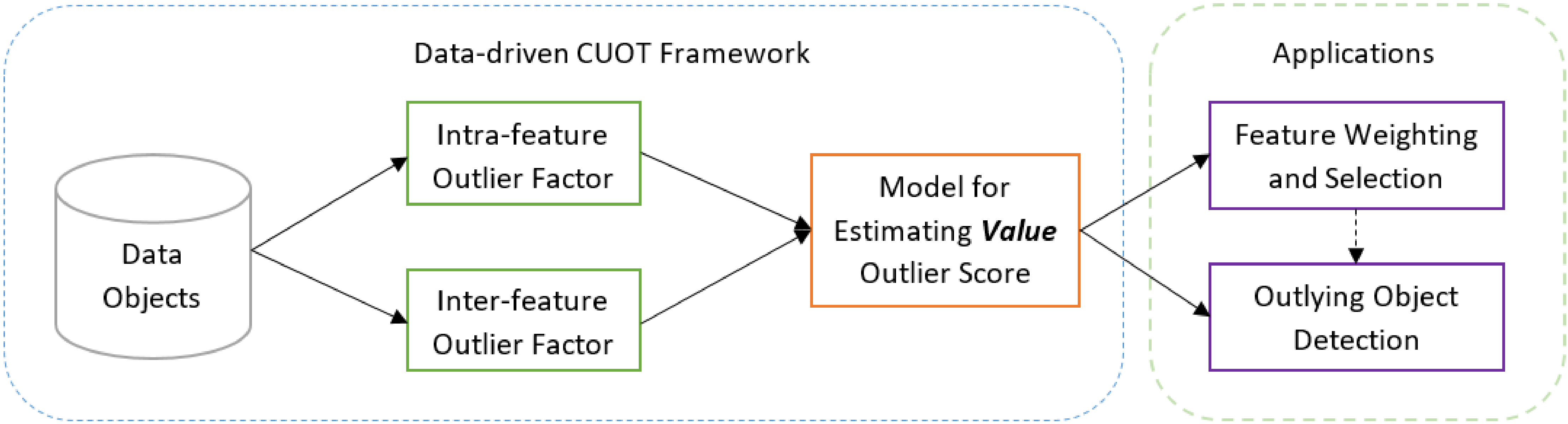
?

The outlierness of a value is dependent on its accompany values

- **Value coupling** – Guilt-by-association
  - *"A man is known by the company he keeps"*
    - Homophily couplings in outlying behaviors (values)
  - Concurrent outlying behaviors
    - E.g., thirsty, weight loss, dryness, urination in diabetes
    - E.g., Feel alienated, violence against the society is not immoral, etc. in terrorist characteristics

Our Framework

- Learning value outlierness from data with non-IID values



CBRW: Intra-feature Outlier Factor

- **Intra-feature** outlier factor for addressing heterogeneity
  - A value of **the same frequency** in different features can have very **different semantic**
  - Given a value $v \in dom(f)$

$$\sigma(v) = \frac{1}{2}[base(m) + dev(v)]$$

where m is the mode in the feature f, $base(m) = 1 - freq(m)$, $dev(v) = \frac{freq(m) - freq(v)}{freq(m)}$

CBRW: Inter-feature Outlier Factor

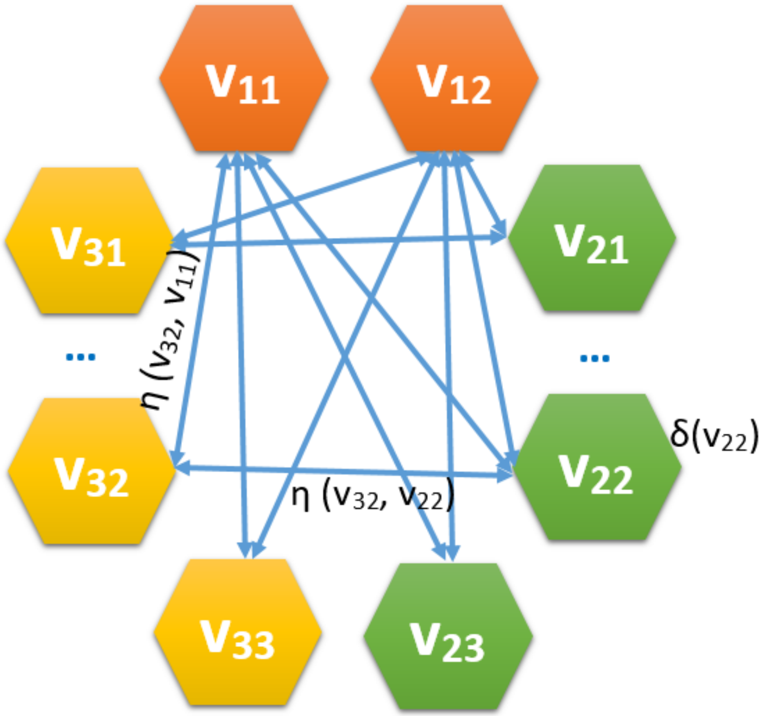
- **Inter-feature** outlier factor capturing the homophily value couplings
  - ***Concurrent rare*** values have high mutual conditional probabilities

$$\boldsymbol{q}_v = [\eta(u,v), \dots, \eta(w,v)]^\intercal = [\frac{freq(u,v)}{freq(v)}, \dots, \frac{freq(w,v)}{freq(v)}]^\intercal, \forall u,w \in V \backslash v$$

where V is the set of all values.

CBRW: Integrating the Two Outlier Factors

- Learning value outlierness from data with non-IID values
  - Map two outlier factors into a value-value graph
  - Stationary probabilities of random walks at value nodes as value outlierness



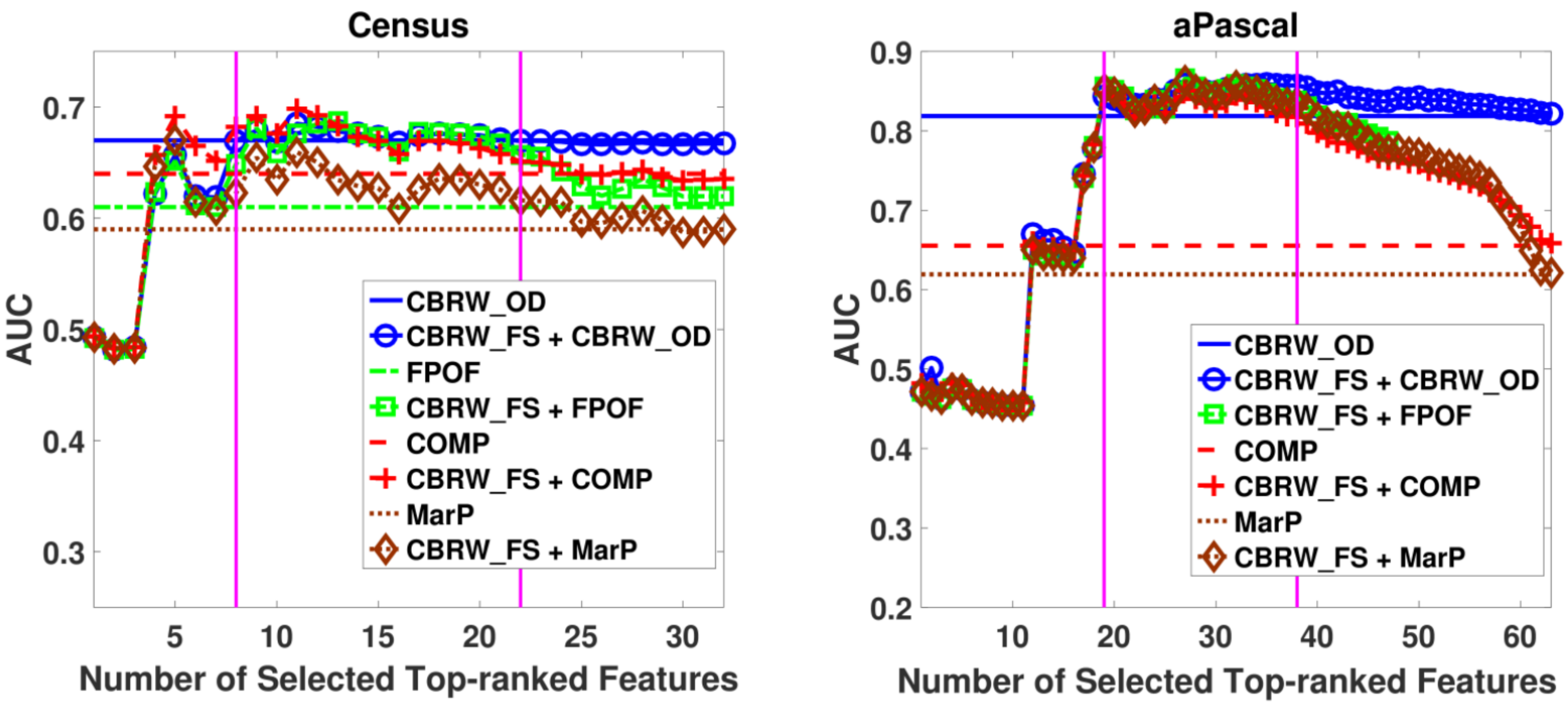
$$W_b(v_{32}, \boldsymbol{v_{22}}) = \frac{\boldsymbol{\delta(v_{22})\eta(v_{32}, v_{22})}}{\delta(v_{22})\eta(v_{32}, v_{22}) + \delta(v_{11})\eta(v_{32}, v_{11})}$$

$$W_b(v_{32}, \boldsymbol{v_{11}}) = \frac{\boldsymbol{\delta(v_{11})\eta(v_{32}, v_{11})}}{\delta(v_{22})\eta(v_{32}, v_{22}) + \delta(v_{11})\eta(v_{32}, v_{11})}$$

Direct Outlier Detection Performance

Data	CBRW	CBRWie	CBRWia	MarP$^+$	MarP	FPOF	COMP	FORE
BM	0.6287	**0.6566**	0.5999	0.5778	0.5584	0.5466	0.6267	0.5762
Census	0.6678	0.6579	**0.6832**	0.6033	0.5899	0.6148	0.6352	0.5378
AID362	**0.6640**	0.6324	0.6034	0.6152	0.6270	○	0.6480	0.6485
w7a	0.6484	**0.7338**	0.4453	0.4565	0.4723	○	0.5683	0.4053
CMC	**0.6339**	0.6323	0.6179	0.5623	0.5417	0.5614	0.5669	0.5746
APAS	0.8190	0.8624	**0.8739**	0.6208	0.6193	○	0.6554	0.4792
CelebA	0.8462	**0.9108**	0.7135	0.7352	0.7358	0.7380	0.7572	0.6797
Chess	**0.7897**	0.4058	0.7766	0.6854	0.6447	0.6160	0.6387	0.6124
AD	0.7348	**0.8270**	0.7250	0.7033	0.7033	○	•	0.7084
SF	0.8812	0.8833	**0.8867**	0.8469	0.8446	0.8556	0.8526	0.7865
Probe	0.9906	**0.9907**	0.9434	0.9795	0.9800	0.9867	0.9790	0.9762
U2R	0.9651	0.9640	0.8817	0.8848	0.8848	0.9156	**0.9893**	0.9781
LINK	0.9976	0.9976	0.9976	0.9977	0.9977	**0.9978**	0.9973	0.9917
R10	**0.9905**	0.9903	0.9823	0.9866	0.9866	○	0.9866	0.9796
CT	0.9703	0.9703	0.9388	0.9770	**0.9773**	0.9772	0.9772	0.9364
Avg.(Top-10)	0.7314	0.7202	0.6925	0.6407	0.6337	0.6554	0.6610	0.6009
Avg.(All)	0.8152	0.8077	0.7779	0.7488	0.7442	0.7810	0.7770	0.7247
	CBRW vs.	0.7959	0.0392	0.0012	0.0008	0.0115	0.0147	0.0040
p-value		CBRWie vs.	0.4225	0.0969	0.0592	0.4316	0.3167	0.0446
			CBRWia vs.	0.1460	0.1223	0.2886	0.8490	0.0979

Outlying Feature Selection Performance



Conclusions

- Learning value outlierness from data with non-IID values
  - Intra-feature and inter-feature outlier factors
- Different applications
  - Direct outlier detection: Significantly outperform other detectors in complex data
  - Feature selection: Substantially improve AUC and efficiency performance of existing OD methods

Non-IID Value-to-Feature-based Approach II

Guansong Pang, Longbing Cao, Ling Chen, Huan Liu. Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection. IJCAI 2017.

Motivation (1/2)

- Outliers are masked by **noisy features**

ID	…	Education	Income	Cheat?
1	**…**	**master**	**low**	**yes**
2	…	master	medium	no
3	…	master	high	no
4	…	master	medium	no
5	…	master	high	no
6	**…**	**PhD**	**high**	**no**
7	**…**	**bachelor**	**high**	**no**

Noisy features

Relevant features

Motivation (2/2)

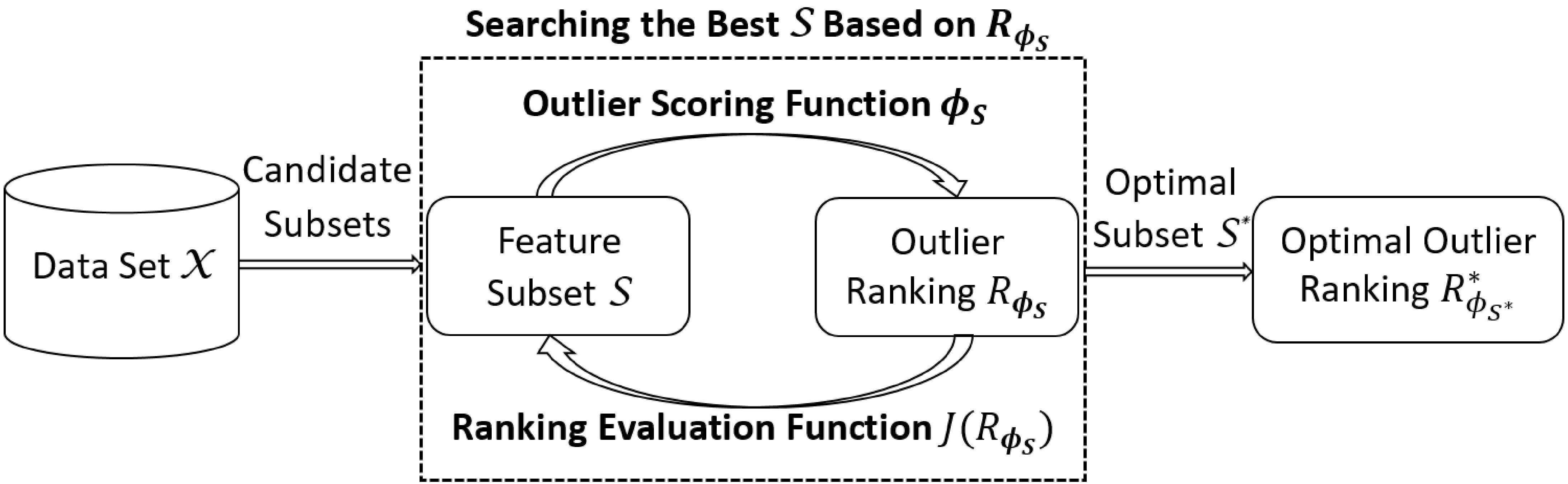
- Existing solutions: subspace/feature selection + OD
- Subspace/feature selection is independent from OD
  - Noisy features bias the subspace/feature search
  - Not optimal w.r.t. subsequent OD method

Filter approach

- Our solution: Simultaneous feature selection and outlier detection
  - **Wrapper approach** for this joint optimization

WrapperOD Framework

Wrapper approach for joint optimization of feature selection and OD



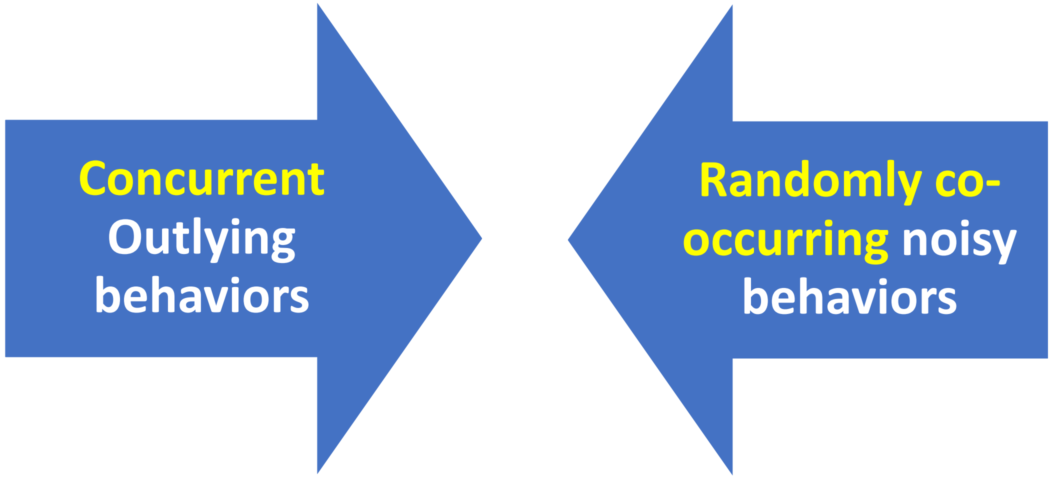
Challenge 1: how to ensure the outlier scoring efficacy

Challenge 2: how to evaluate the outlier ranking without class labels

The WrapperOD Instance: HOUR Scoring Function (1/3)

- The scoring function should at least be
  - Sufficiently resilient to noisy features
  - Very efficient

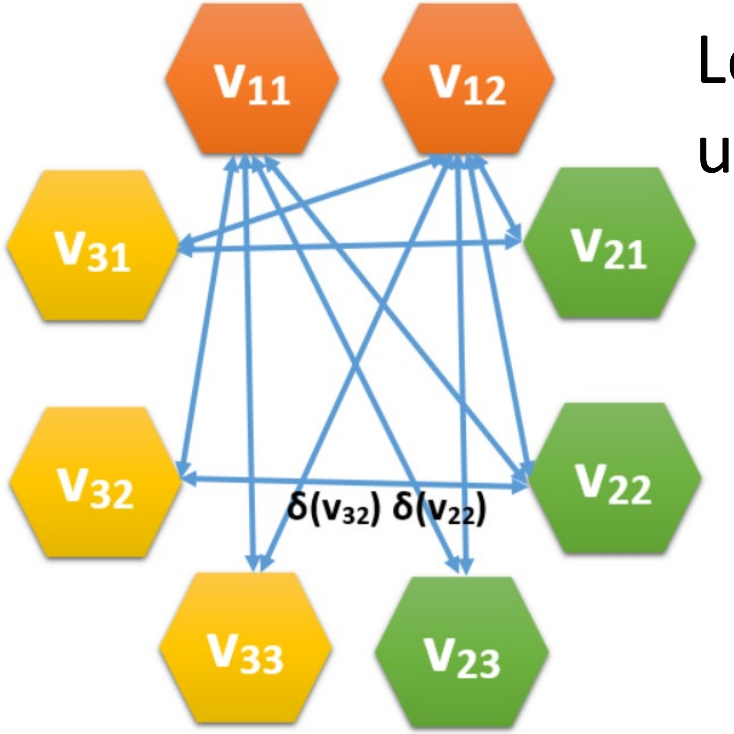
- Homophily couplings between outlying values



The WrapperOD Instance: HOUR Scoring Function (2/3)

Simplified CBRW:

$$\delta(v_{22})\eta(v_{32}, v_{22}) \rightarrow \delta(v_{32})\delta(v_{22})$$



Leading to random walks on undirected value graph

- Efficient closed-form solution

$$\tau(v) = \frac{\sum_{u \in \mathcal{N}_v} \delta(v)\delta(u)}{\sum_{v \in \mathcal{V}} \sum_{u \in \mathcal{N}_v} \delta(v)\delta(u)}$$

The WrapperOD Instance: HOUR Scoring Function (3/3)

- Homophily coupling learning – stage I

$$\tau(v) = \frac{\sum_{u \in \mathcal{N}_v} \delta(v)\delta(u)}{\sum_{v \in \mathcal{V}} \sum_{u \in \mathcal{N}_v} \delta(v)\delta(u)}$$

- Homophily coupling learning – stage II

$$\psi(v) = \sum_{u \in \mathcal{N}_v} \rho(u, v)\tau(u)$$

The WrapperOD Instance: HOUR Outlier Ranking Quality Evaluation

- Average outlierness margin between top-*k* objects and the rest of objects

$$J(R_{\phi_{\mathcal{S}}}, k) = \frac{\Delta_{\mathcal{S}}}{|\mathcal{S}|} = \frac{1}{k|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{O}} [\phi_{\mathcal{S}}(\boldsymbol{x}) - \phi_{\mathcal{S}}(\boldsymbol{x}')]$$

where *x'* is the data object ranked in the median position in the rest of (*N* - *k*) objects

Recursive backward feature elimination is used for generating the feature subset *S*

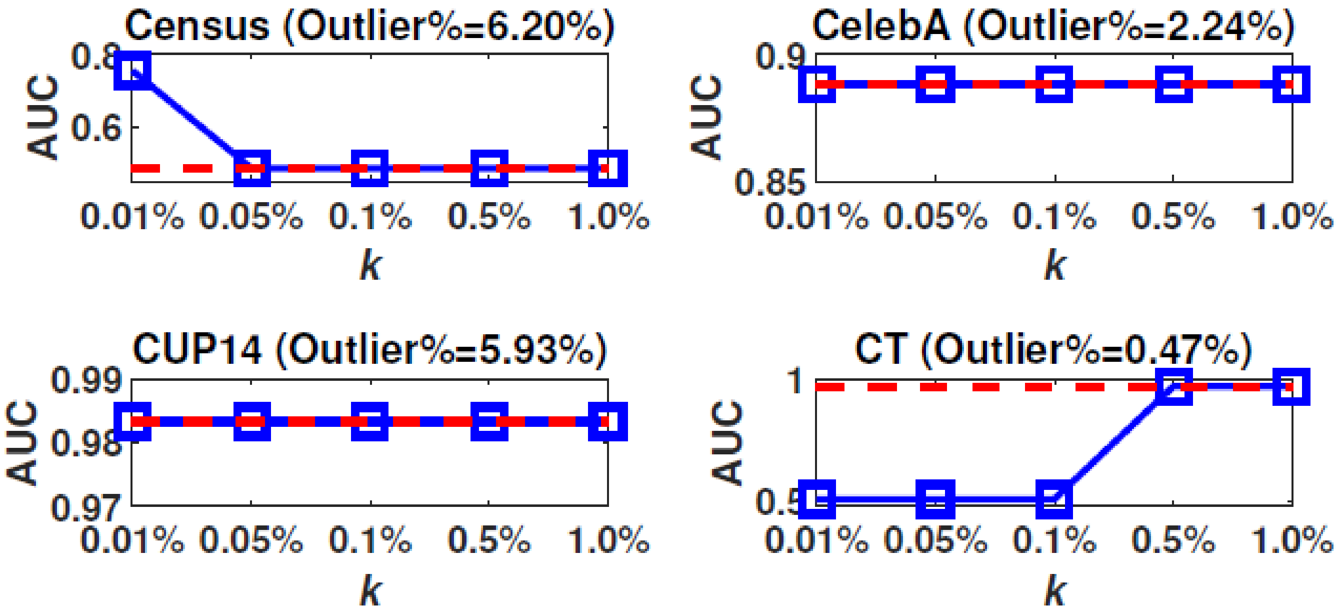
Comparing to State-of-the-art Detectors

					AUC				$P@n$			
Data	N	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{S}\rvert(\triangledown)$	*fnl*	HOUR	CBRW	COMP	FPOF	HOUR	CBRW	COMP	FPOF
SylvaA	14,395	172	16(91%)	91%	**0.9829**	0.9353	0.8855	NA	**0.7483**	0.5914	0.3770	NA
BM	41,188	10	5(50%)	90%	**0.6939**	0.6287	0.6267	0.5466	**0.3265**	0.2474	0.2565	0.1369
AID362	4,279	114	8(93%)	86%	0.5147	**0.6640**	0.6480	NA	**0.0833**	0.0500	0.0167	NA
APAS	12,695	64	13(80%)	81%	**0.9065**	0.8190	0.6554	NA	0.0000	0.0000	0.0000	NA
SylvaP	14,395	87	15(83%)	78%	**0.9725**	0.9715	0.9537	NA	**0.6907**	0.6151	0.5700	NA
Census	299,285	33	3(91%)	58%	0.4867	**0.6678**	0.6352	0.6148	0.0616	**0.0677**	0.0675	0.0637
CelebA	202,599	39	12(69%)	49%	**0.8879**	0.8462	0.7572	0.7380	**0.2085**	0.1748	0.1533	0.1256
CUP14	619,326	7	3(57%)	43%	**0.9833**	0.9420	0.9398	0.6041	**0.6730**	0.2671	0.2671	0.0000
Alcohol	1,044	32	3(91%)	38%	**0.9365**	0.9254	0.8919	0.5468	**0.3889**	0.3333	**0.3889**	0.0556
CMC	1,473	8	4(50%)	38%	**0.6647**	0.6339	0.5669	0.5614	0.0345	0.0345	0.0345	**0.1034**
CT	581,012	44	3(93%)	34%	0.9688	0.9703	**0.9772**	0.9770	0.0499	0.0386	**0.0688**	0.0644
Chess	28,056	6	3(50%)	33%	**0.8507**	0.7897	0.6387	0.6160	0.0000	0.0000	0.0000	0.0000
Turkiye	5,820	32	21(34%)	25%	**0.5256**	0.5116	0.5101	0.4746	**0.0776**	0.0746	0.0687	0.0597
Credit	30,000	9	6(33%)	11%	**0.7204**	0.5804	0.6543	0.6428	**0.4875**	0.2215	0.3502	0.3333
Probe	64,759	6	2(67%)	0%	0.9661	**0.9906**	0.9790	0.9867	0.8440	**0.8579**	0.7928	0.8548
Average	128,022	44	8(69%)	50%	0.8041	0.7918	0.7546	0.6644	0.3116	0.2383	0.2275	0.1634
				p-value		0.1876	0.0730	0.0322		0.0068	0.0068	0.1055

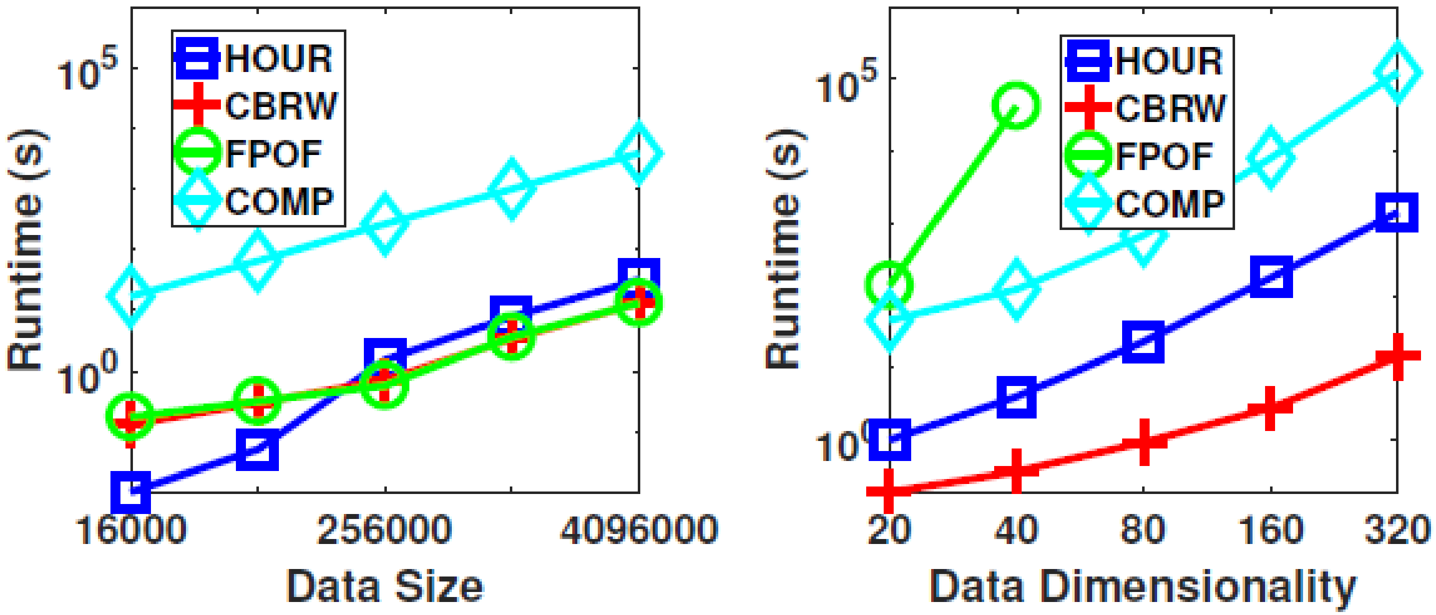
Comparing to State-of-the-art FS + Detectors

	AUC				
Data	HOUR	CBRW†	CBRW‡	COMP†	COMP‡
SylvaA	**0.9829**	0.8793	0.9381	0.8726	0.8858
BM	**0.6939**	0.6104	0.6114	0.6239	0.6239
AID362	0.5147	0.4659	**0.6518**	0.4982	0.6342
APAS	**0.9065**	0.6621	0.8807	0.6532	0.8771
SylvaP	**0.9725**	0.9582	0.9707	0.9307	0.9628
Census	0.4867	0.4844	0.6999	0.4841	**0.7135**
CelebA	**0.8879**	0.8865	0.8502	0.8855	0.7594
CUP14	**0.9833**	0.9821	0.9358	0.9821	0.9618
Alcohol	**0.9365**	0.9264	0.9294	0.8919	0.8595
CMC	**0.6647**	0.6366	0.6444	0.6475	0.6586
CT	**0.9688**	0.9192	0.9673	0.9187	0.9670
Chess	**0.8507**	0.7268	0.7649	0.7529	0.6305
Turkiye	**0.5256**	0.5161	0.5108	0.5145	0.5119
Credit	**0.7204**	0.5712	0.5712	0.6566	0.6566
Probe	0.9661	0.9591	0.9591	**0.9794**	**0.9794**
Average	0.8041	0.7456	0.7924	0.7528	0.7788
p-value	-	0.0001	0.0730	0.0006	0.1070

Sensitivity Test



Scalability Test



Conclusions

- This the first wrapper approach for outlier detection
- The simultaneous optimization scheme enables HOUR to work well in very noisy scenarios
  - Significantly better top-k outlier detection
- Good stability and scalability
- Source code will be available at

https://sites.google.com/site/gspangsite/sourcecode

Out-of-Distribution Detection

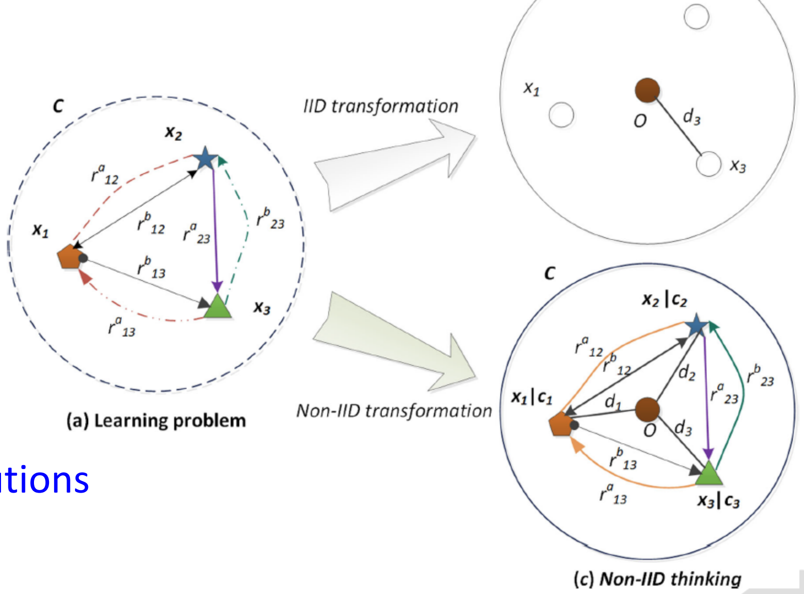
Conclusions & Prospects

Non-IID Learning: A Challenging Problem

- Data non-IIDness
- Data sampling
- Non-IID similarity/dissimilarity metrics/measures
- Non-IID representations
- Model structure
- Objective functions
- Result interpretation
- New perspectives

L. Cao. Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning, IEEE Intelligent Systems, 37:4, 3-15, 2022

O_1, O_2, O_3 are iid
$d_3 = ||O_3 - O||$



O_1, O_2, O_3 share different distributions
$d_3 = ||O_3 - O||$
$= || O_3(r_{13}, r_{23}) - O(d_1, d_2) ||$

FIGURE 1. IID thinking versus non-IID thinking. For example, from the machine learning perspective, a given learning problem (a) is either (b) IID transformed per the IID assumption (i.e., independent and identically distributed) and then solved by an IID learning system, or (c) non-IID transformed by characterizing its non-IIDness (i.e., heterogeneity and interaction) and then solved by a non-IID system.

IID to non-IID space

L. Cao. Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning, IEEE Intelligent Systems, 37:4, 3-15, 2022

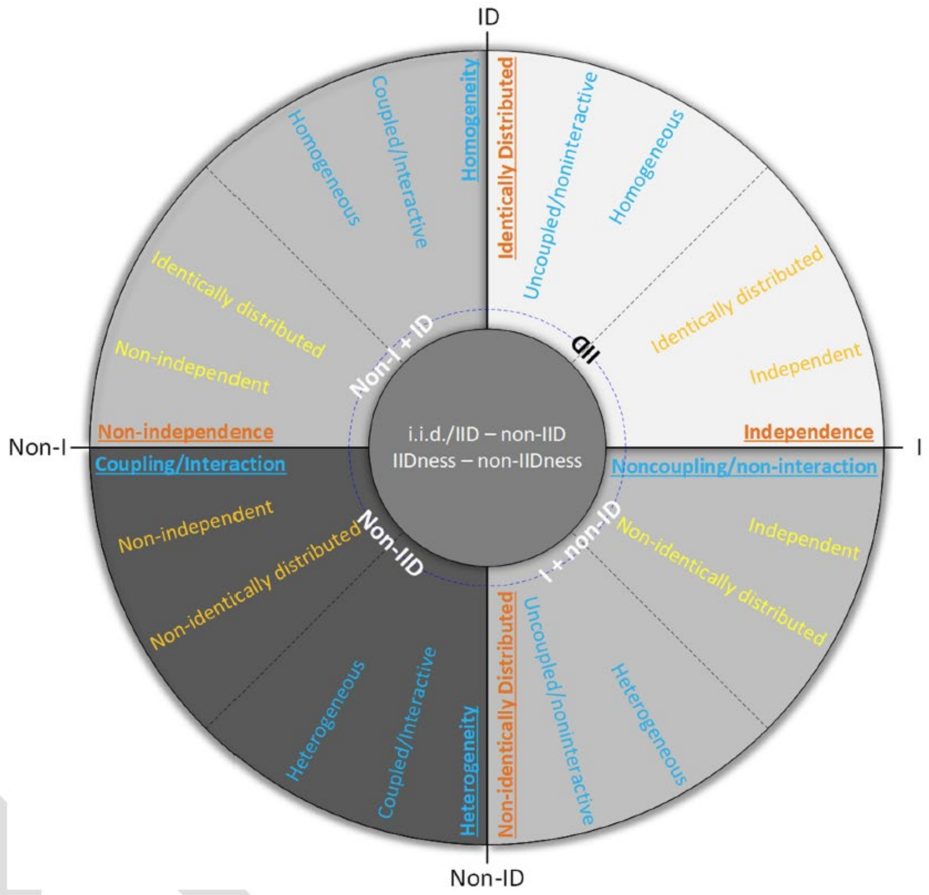


FIGURE 2. IID to non-IID space. Two sets of axes: classic independence/nonindependence-identical distribution/nonidentical distribution versus heterogeneity/homogeneity-coupling/interaction//noncoupling/noninteraction; generating four quadrants: IID, non-I + ID, non-IID, and I + non-ID.

Aspects of Non-IIDness

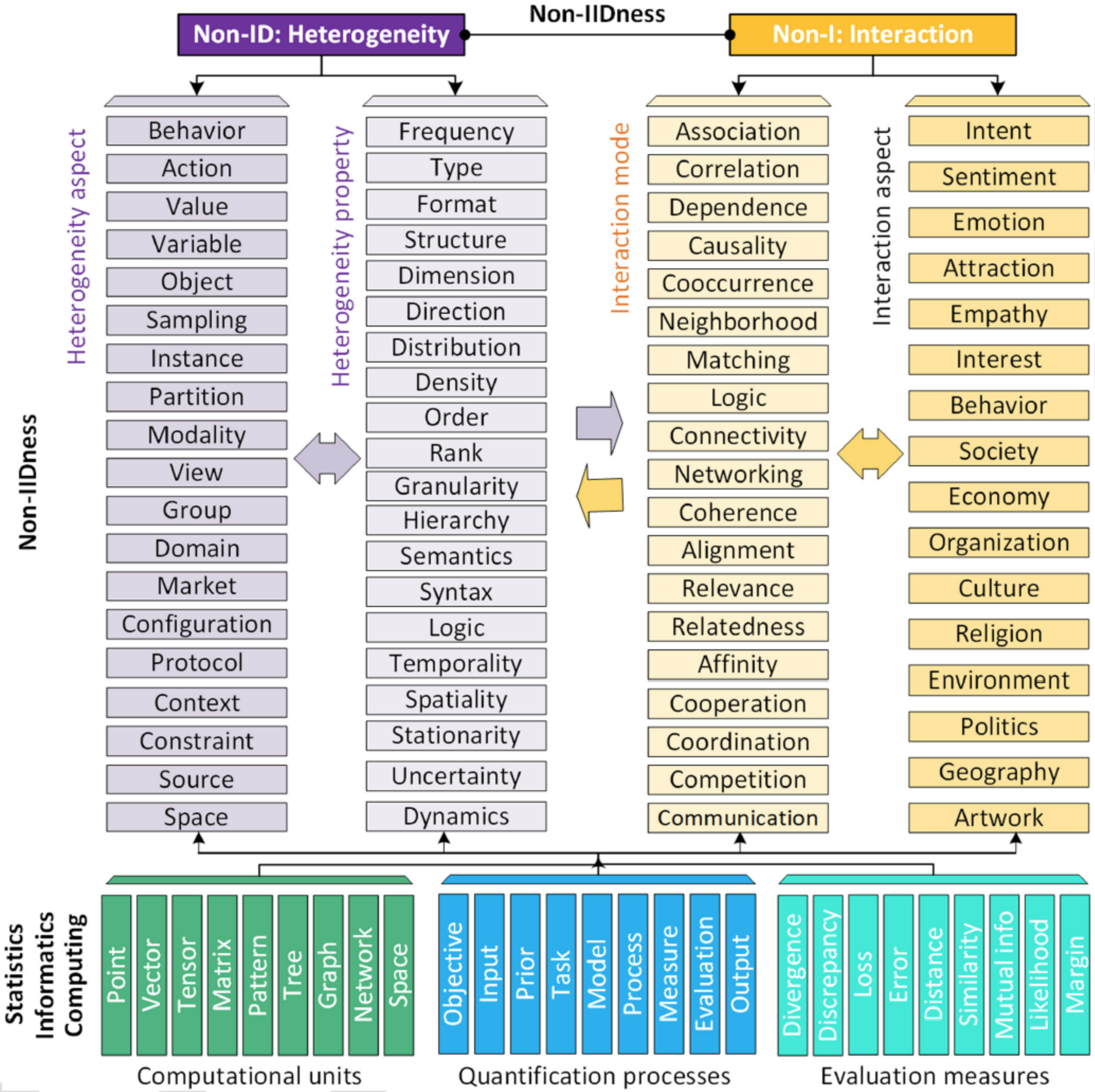
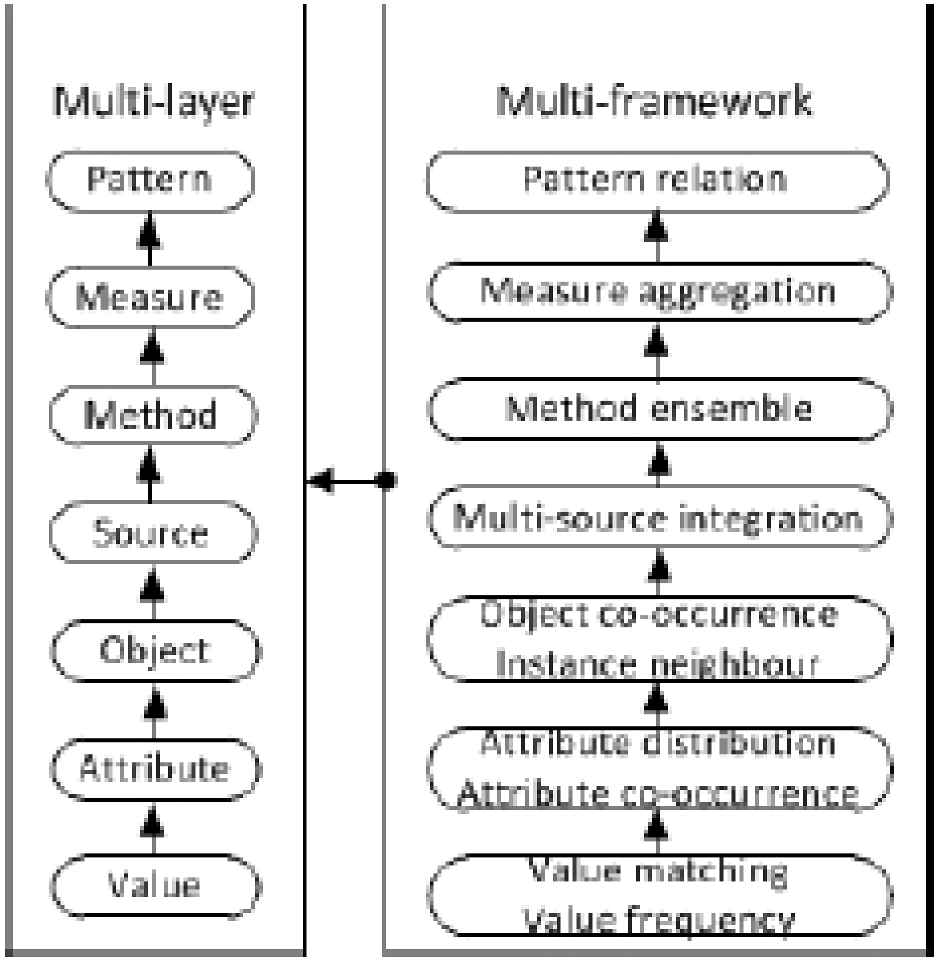


FIGURE 3. Terminology and conceptual map of non-IIDness: non-ID—heterogeneities, and non-I—interactions.

L. Cao. Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning, IEEE Intelligent Systems, 37:4, 3-15, 2022

Longbing Cao. Coupling Learning of Complex Interactions, Journal of Information Processing and Management, 51(2): 167-186 (2015)

Hierarchical Non-IIDness



Longbing Cao. Coupling Learning of Complex Interactions, Journal of Information Processing and Management, 51(2): 167-186 (2015)

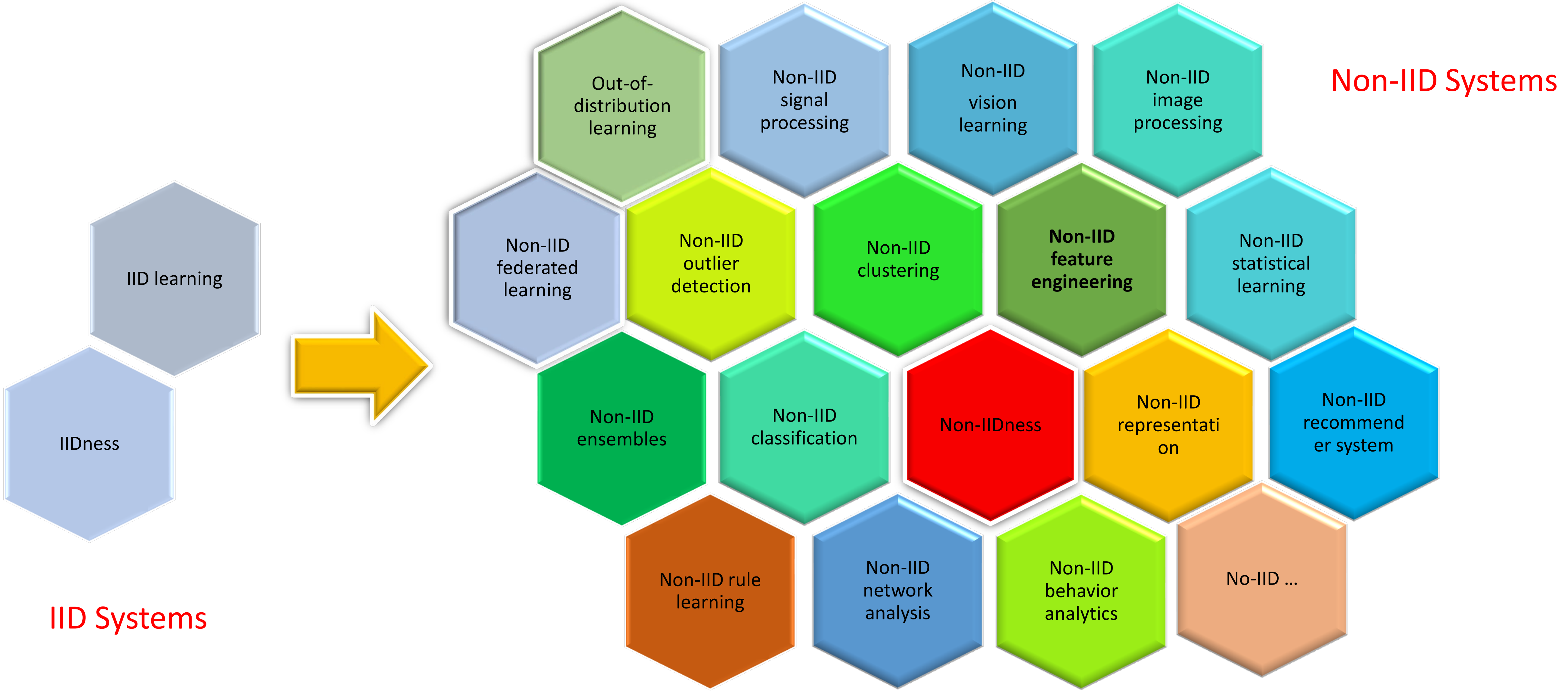
Some Fundamental Issues

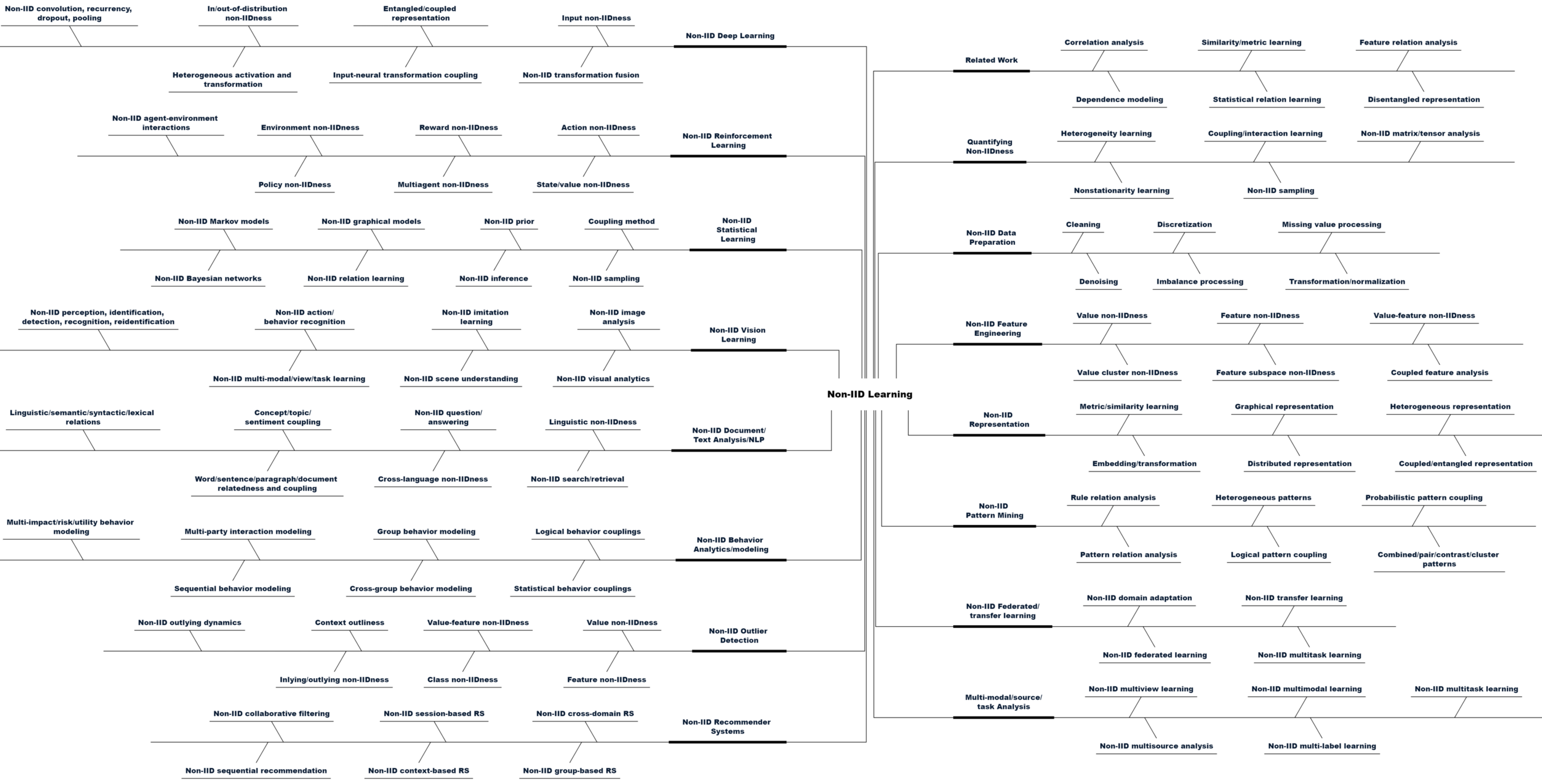
- How can we determine whether a dataset is IID or non-IID?
- Whether association, correlation, causality, dependency, uncertainty/randomness cover all relationships?
- Real-life problems often involve multiple sources (views, modals, tasks, etc.) of data, are they ID?
- What do we mean by ‘heterogeneity’? Does `identically distributed’ mean `homogeneity’?
- What do we mean by `independence’ in a broad sense?

Some Fundamental Issues

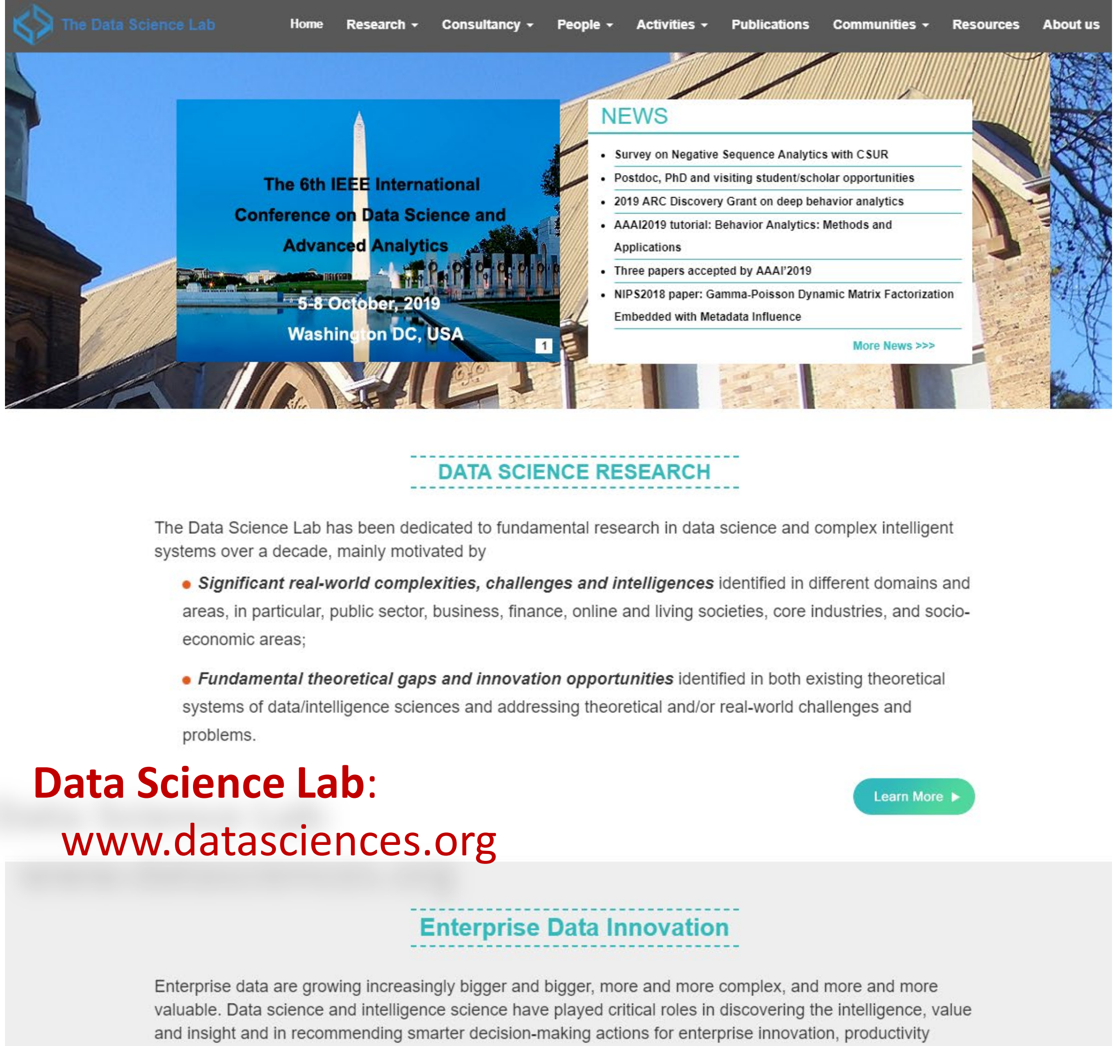
- Are KNN, SVM, decision tree, classic ensemble methods IID?
- Does classic transfer learning capture non-IIDness?
- In probabilistic graphical modeling, how non-IIDness is modelled?
- Do deep neural networks capture non-IIDness? To what extent?
- …

IID to Non-IID Learning Systems





Non-IID Learning
Non-IID Deep Learning
Non-IID convolution, recurrency, dropout, pooling
In/out-of-distribution non-IIDness
Entangled/coupled representation
Input non-IIDness
Heterogeneous activation and transformation
Input-neural transformation coupling
Non-IID transformation fusion
Non-IID Reinforcement Learning
Non-IID agent-environment interactions
Environment non-IIDness
Reward non-IIDness
Action non-IIDness
Policy non-IIDness
Multiagent non-IIDness
State/value non-IIDness
Non-IID Statistical Learning
Non-IID Markov models
Non-IID graphical models
Non-IID prior
Coupling method
Non-IID Bayesian networks
Non-IID relation learning
Non-IID inference
Non-IID sampling
Non-IID Vision Learning
Non-IID perception, identification, detection, recognition, reidentification
Non-IID action/ behavior recognition
Non-IID imitation learning
Non-IID image analysis
Non-IID multi-modal/view/task learning
Non-IID scene understanding
Non-IID visual analytics
Non-IID Document/ Text Analysis/NLP
Linguistic/semantic/syntactic/lexical relations
Concept/topic/ sentiment coupling
Non-IID question/ answering
Linguistic non-IIDness
Word/sentence/paragraph/document relatedness and coupling
Cross-language non-IIDness
Non-IID search/retrieval
Non-IID Behavior Analytics/modeling
Multi-impact/risk/utility behavior modeling
Multi-party interaction modeling
Group behavior modeling
Logical behavior couplings
Sequential behavior modeling
Cross-group behavior modeling
Statistical behavior couplings
Non-IID Outlier Detection
Non-IID outlying dynamics
Context outliness
Value-feature non-IIDness
Value non-IIDness
Inlying/outlying non-IIDness
Class non-IIDness
Feature non-IIDness
Non-IID Recommender Systems
Non-IID collaborative filtering
Non-IID session-based RS
Non-IID cross-domain RS
Non-IID sequential recommendation
Non-IID context-based RS
Non-IID group-based RS
Related Work
Correlation analysis
Similarity/metric learning
Feature relation analysis
Dependence modeling
Statistical relation learning
Disentangled representation
Quantifying Non-IIDness
Heterogeneity learning
Coupling/interaction learning
Non-IID matrix/tensor analysis
Nonstationarity learning
Non-IID sampling
Non-IID Data Preparation
Cleaning
Discretization
Missing value processing
Denoising
Imbalance processing
Transformation/normalization
Non-IID Feature Engineering
Value non-IIDness
Feature non-IIDness
Value-feature non-IIDness
Value cluster non-IIDness
Feature subspace non-IIDness
Coupled feature analysis
Non-IID Representation
Metric/similarity learning
Graphical representation
Heterogeneous representation
Embedding/transformation
Distributed representation
Coupled/entangled representation
Non-IID Pattern Mining
Rule relation analysis
Heterogeneous patterns
Probabilistic pattern coupling
Pattern relation analysis
Logical pattern coupling
Combined/pair/contrast/cluster patterns
Non-IID Federated/ transfer learning
Non-IID domain adaptation
Non-IID transfer learning
Non-IID federated learning
Non-IID multitask learning
Multi-modal/source/ task Analysis
Non-IID multiview learning
Non-IID multimodal learning
Non-IID multitask learning
Non-IID multisource analysis
Non-IID multi-label learning

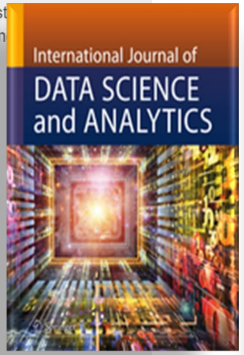


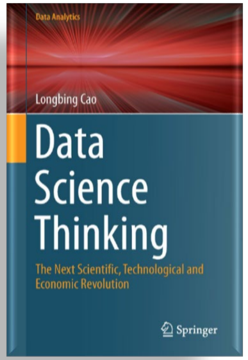
Data Science Lab:
www.datasciences.org

Thank You Very Much

Comments & suggestions:

Longbing.Cao@uts.edu.au





References

Not all references are listed here

https://datasciences.org/non-iid-learning/

References

Paper download: www.datasciences.org

- **Non-IID learning concepts**
  - L. Cao. Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning, IEEE Intelligent Systems, 37:4, 3-15, 2022
  - Longbing Cao, Philip S. Yu, Zhilin Zhao: Shallow and Deep Non-IID Learning on Complex Data. KDD 2022: 4774-4775
  - Longbing Cao: Non-IID Federated Learning. IEEE Intell. Syst. 37(2): 14-15 (2022)
  - Can Wang, Fosca Giannotti, Longbing Cao. Learning Complex Couplings and Interactions. IEEE Intell. Syst. 36(1): 3-5, 2021.
  - Longbing Cao. Non-IIDness Learning in Behavioral and Social Data, The Computer Journal, 57(9): 1358-1370 (2014).
  - Longbing Cao. Coupling Learning of Complex Interactions, Journal of Information Processing and Management, 51(2): 167-186 (2015).
  - Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery, 3(2): 140-155, 2013.
  - Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, IEEE Trans. SMC Part B, 41(3): 699 - 712, 2011.
- **Non-IID representation learning**
  - Chengzhang Zhu, Longbing Cao and Jianpin Yin. Unsupervised Heterogeneous Coupling Learning for Categorical Representation. IEEE Transaction on Pattern Recognition and Machine Intelligence, 44(1): 533-549, 2022
  - Songlei Jian, Liang Hu, Longbing Cao, and Kai Lu. Metric-based Auto-Instructor for Learning Mixed Data Representation. AAAI2018.
  - Songlei Jian, Longbing Cao, Guansong Pang, Kai Lu, Hang Gao. Embedding-based Representation of Categorical Data with Hierarchical Value Couplings, IJCAI 2017.
  - Chunming Liu, Longbing Cao, Philip S Yu. Coupled Fuzzy k-Nearest Neighbors Classification of Imbalanced Non-IID Categorical Data, IJCNN 2014.

References

Paper download: www.datasciences.org

- **Data discretization**
  - Can Wang, Mingchun Wang, Zhong She, Longbing Cao. CD: A Coupled Discretization Algorithm, PAKDD2012, 407-418
- **Non-IID K-Means**
  - Can Wang, Zhong She, Longbing Cao. Coupled Attribute Analysis on Numerical Data, IJCAI 2013.
  - Can Wang, Dong, Xiangjun; Zhou, Fei; Longbing Cao, Chi, Chi-Hung. Coupled Attribute Similarity Learning on Categorical Data, IEEE Transactions on Neural Networks and Learning Systems, 26(4): 781-797 (2015).
- **Non-IID K-Mode & Spectral clustering**
  - Can Wang, Longbing Cao, Minchun Wang, Jinjiu Li, Wei Wei, Yuming Ou. Coupled Nominal Similarity in Unsupervised Learning, CIKM 2011, 973-978.
  - Can Wang, Dong, Xiangjun; Zhou, Fei; Longbing Cao, Chi, Chi-Hung. Coupled Attribute Similarity Learning on Categorical Data (extension of the CIKM2011 paper), IEEE Transactions on Neural Networks and Learning Systems, 26(4): 781-797 (2015).
- **Non-IID KNN/classification**
  - Chunming Liu, Longbing Cao. A Coupled k-Nearest Neighbor Algorithm for Multi-label Classification, PAKDD2015, 176-187.
  - Chunming Liu, Longbing Cao, Philip S Yu. A Hybrid Coupled k-Nearest Neighbor Algorithm on Imbalance Data, IJCNN 2014.
  - Chunming Liu, Longbing Cao, Philip S Yu. Coupled Fuzzy k-Nearest Neighbors Classification of Imbalanced Non-IID Categorical Data, IJCNN 2014.

References

- **Non-IID ensemble clustering**
  - Can Wang, Zhong She, Longbing Cao. Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects, ICDE2013.
- **Group/Coupled behavior analysis with couplings**
  - Can Wang, Longbing Cao, Chi-Hung Chi: Formalization and Verification of Group Behavior Interactions. IEEE Trans. Systems, Man, and Cybernetics: Systems 45(8): 1109-1124 (2015)
  - Wei Cao, Liang Hu, Longbing Cao: Deep Modeling Complex Couplings within Financial Markets. AAAI 2015: 2518-2524
  - Wei Cao, Longbing Cao, Yin Song: Coupled market behavior based financial crisis detection. IJCNN 2013: 1-8
  - Yin Song, Longbing Cao, et al. Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation, KDD 2012, 976-984.
  - Yin Song and Longbing Cao. Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets, IJCNN 2012, 1-8.
  - Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Applications, IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).
  - Longbing Cao, Yuming Ou, Philip S YU, Gang Wei. Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors, KDD2010, 85-94.
- **Non-IID image processing**
  - Yonggang Huang, Yuying Liu, Longbing Cao, Jun Zhang, I Pan. Exploring Feature Coupling and Model Coupling for Image Source Identification, IEEE Transactions on Information Forensics & Security, 2018
  - Zhe Xu, Ya Zhang, Longbing Cao. Social Image Analysis from a Non-IID Perspective, IEEE Transactions on Multimedia.
  - Yinghuan Shi, Heung-Il Suk, Yang Gao, Dinggang Shen. Joint Coupled-Feature Representation and Coupled Boosting for Alzheimer's Disease Diagnosis, CVPR, 2014

References

- **Non-IID computer vision tasks**
  - Shi, Y., Li, W., Gao, Y., Cao, L., Shen, D. Beyond IID: Learning to combine non-iid metrics for vision tasks. *AAAI'17*
- **Statistical relation learning**
  - Trong Dinh Thac Do and Longbing Cao. Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence, NIPS2018.
  - Trong Dinh Thac Do and Longbing Cao. Metadata-dependent Infinite Poisson Factorization for Efficiently Modelling Sparse and Large Matrices in Recommendation, IJCAI2018
  - Trong Dinh Thac Do, Longbing Cao. Coupled Poisson Factorization Integrated with User/Item Metadata for Modeling Popular and Sparse Ratings in Scalable Recommendation. AAAI2018
  - Xuhui Fan, Richard Xu, Longbing Cao. Copula Mixed-Membership Stochastic Blockmodel. IJCAI2016.
  - Xuhui Fan, Richard Xu, Longbing Cao, Yin Song. Learning Nonparametric Relational Models by Conjugately Incorporating Node Information in a Network. IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2016.2521376.
  - Fan, Xuhui; Longbing Cao, Xu, Richard Yi Da. Dynamic Infinite Mixed-Membership Stochastic Blockmodel, IEEE Transactions on Neural Networks and Learning Systems, 26(9): 2072-2085 (2015).
  - Wei Cao, Liang Hu, Longbing Cao. Deep Modeling Complex Couplings within Financial Markets, AAAI2015, 2518-2524.
  - Liang Hu, Longbing Cao, Guandong Xu, Jian Cao, and Wei Cao. Bayesian Heteroskedastic Choice Modeling on Non-identically Distributed Linkages, ICDM2014.
  - Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu and Wei Cao. Deep Modeling of Group Preferences for Group-based Recommendation, AAAI 2014, 1861-1867.

References

- **Non-IID outlier detection/feature selection**
  - Guansong Pang, Longbing Cao, Ling Chen, Huan Liu. Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection, IJCAI2017
  - Guansong Pang, Hongzuo Xu, Longbing Cao and Wentao Zhao. Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data. CIKM2017
  - Guansong Pang, Longbing Cao, Ling Chen. Outlier Detection in Complex Categorical Data by Modelling the Feature Value Couplings. IJCAI2016.
  - Guansong Pang, Longbing Cao, Ling Chen. Unsupervised Feature Selection for Outlier Detection by Modelling Hierarchical Value-Feature Couplings. ICDM2016.
- **Pattern/rule relation analysis/combined pattern mining**
  - Wei Wang, Longbing Cao: Explicit and Implicit Pattern Relation Analysis for Discovering Actionable Negative Sequences. CoRR abs/2204.03571 (2022)
  - Shoujin Wang, Longbing Cao. Inferring Implicit Rules by Learning Explicit and Hidden Item Dependency. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017
  - Jinjiu Li, Can Wang, Longbing Cao, Philip S. Yu. Efficient Selection of Globally Optimal Rules on Large Imbalanced Data Based on Rule Coverage Relationship Analysis, SDM 2013.
  - Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang. Combined Pattern Mining: from Learned Rules to Actionable Knowledge, LNCS 5360/2008, 393-403, 2008.
  - Huaifeng Zhang, Yanchang Zhao, Longbing Cao and Chengqi Zhang. Combined Association Rule Mining, PAKDD2008.

References

- **Non-IID recommender systems**

- Quangui Zhang, Longbing Cao, Chengzhang Zhu, Zhiqiang Li and Jinguang Sun. CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering, IJCAI2018
- Longbing Cao. Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting. Engineering, 2: 212-224, doi:10.1016/J.ENG.2016.02.013., 2016.
- Liang Hu, Longbing Cao, Shoujin Wang, Guandong Xu, Jian Cao, Zhiping Gu. Diversifying Personalized Recommendation with User-session Context. In *IJCAI*. 2017
- Hu, L., Cao, L., Cao, J., Gu, Z., Xu, G., and Wang, J. Improving the Quality of Recommendations for Users and Items in the Tail of Distribution. ACM Trans. Inf. Syst., 2017
- Hu, L., Cao, L., Cao, J., Gu, Z., Xu, G., & Yang, D. (2016). Learning Informative Priors from Heterogeneous Domains to Improve Recommendation in Cold-Start User Domains. *ACM Transactions on Information Systems (TOIS)*, 35(2), 13.
- Hu, L., Cao, J., Xu, G., Cao, L., Gu, Z., & Cao, W. (2014, July). Deep Modeling of Group Preferences for Group-Based Recommendation. In *AAAI* (Vol. 14, pp. 1861-1867).
- Liang Hu, Wei Cao, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Bayesian Heteroskedastic Choice Modeling on Non-identically Distributed Linkages, ICDM 2014
- Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, Can Zhu: Personalized recommendation via cross-domain triadic factorization. WWW 2013
- Liang Hu, Jian Cao, Guandong Xu, Jie Wang, Zhiping Gu, Longbing Cao, Cross-Domain Collaborative Filtering via Bilinear Multilevel Analysis, IJCAI 2013
- Longbing Cao, Philip Yu. Non-IID Recommendation Theories and Systems. IEEE Intelligent Systems, 31(2), 81-84, 2016.
- Fangfang Li, Guandong Xu, Longbing Cao. Coupled Matrix Factorization within Non-IID Context, PAKDD2015, 707-719.
- Fangfang Li, Guandong Xu, Longbing Cao: Coupled Item-Based Matrix Factorization. WISE (1) 2014: 1-14
- Fangfang Li, Guandong Xu, Longbing Cao, Zhendong Niu. Coupled Group-based Matrix Factorization for Recommender System, WISE 2013.
- Yonghong Yu, Can Wang, Yang Gao, Longbing Cao, Qianqian Chen: A Coupled Clustering Approach for Items Recommendation. PAKDD (2) 2013

References

- **Non-IID document/text analysis**
  - Jinjin Guo, Longbing Cao, Zhiguo Gong: Recurrent Coupled Topic Modeling over Sequential Documents. ACM Trans. Knowl. Discov. Data 16(1): 8:1-8:32 (2022)
  - Shufeng Hao, Chongyang Shi, Longbing Cao, Zhendong Niu, Ping Guo: Learning deep relevance couplings for ad-hoc document retrieval. Expert Syst. Appl. 183: 115335, 2021
  - Shufeng Hao, Chongyang Shi, Zhendong Niu, Longbing Cao. Concept Coupling Learning for Improving Concept Lattice-based Document Retrieval. Engineering Applications of Artificial Intelligence, Volume 69, 65-75, 2018
  - Qianqian Chen, Liang Hu, Jia Xu, Wei Liu, Longbing Cao. Document similarity analysis via involving both explicit and implicit semantic couplings. DSAA 2015: 1-10.
  - Xin Cheng, Duoqian Miao, Can Wang, Longbing Cao. Coupled Term-Term Relation Analysis for Document Clustering, IJCNN2013.
- **Keyword query with couplings**
  - Xiangfu Meng, longbing Cao and Jingyu Shao. Semantic Approximate Keyword Query Based on Keyword and Query Coupling Relationship Analysis. CIKM2014
- **Non-IID similarity/metric learning**
  - Chengzhang Zhu, Longbing Cao, Qiang Liu, Jianpin Yin and Vipin Kumar. Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings. IEEE Transactions on Knowledge and Data Engineering, DOI: 10.1109/TKDE.2018.2791525, 2018
  - Songlei Jian, Longbing Cao, Kai Lu, Hang Gao. Unsupervised Coupled Metric Similarity for Non-IID Categorical Data. IEEE Transactions on Knowledge and Data Engineering, 2018
  - Can Wang, Chi-Hung Chi, Zhong She, Longbing Cao, Bela Stantic: Coupled Clustering Ensemble by Exploring Data Interdependence. TKDD 12(6): 63:1-63:38 (2018)
- **Open set/open domain learning**

References

- **Out-of-distribution detection and learning**

  - Zhilin Zhao, Longbing Cao, Yuan-Yu Wan: Coupling Online-Offline Learning for Multi-distributional Data Streams. CoRR abs/2202.05996 (2022)
  - Zhilin Zhao, Longbing Cao, Kun-Yu Lin: Supervision Adaptation Balances In-Distribution Generalization and Out-of-Distribution Detection. CoRR abs/2206.09380 (2022)
  - Zhilin Zhao, Longbing Cao, Chang-Dong Wang: Gray Learning from Non-IID Data with Out-of-distribution Samples. CoRR abs/2206.09375 (2022)
  - Zhilin Zhao, Longbing Cao, Kun-Yu Lin: Out-of-distribution Detection by Cross-class Vicinity Distribution of In-distribution Data. CoRR abs/2206.09385 (2022)
  - Zhilin Zhao, Longbing Cao: Label and Distribution-discriminative Dual Representation Learning for Out-of-Distribution Detection. CoRR abs/2206.09387 (2022)

- **Open set/open domain learning**

References

- Aggarwal, C. C. (2017). Outlier analysis. Springer.
- Anderson, C. 2006. *The long tail: Why the future of business is selling less of more*. Hachette Digital, Inc.
- Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. CoRR, abs/1511.06939, 2015.
- Charlin, L., Ranganath, R., McInerney, J., & Blei, D. M. (2015, September). Dynamic poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 155-162). ACM.
- Chau, D. H. P., Nachenberg, C., Wilhelm, J., Wright, A., & Faloutsos, C. (2011, April). Polonium: Tera-scale graph mining and inference for malware detection. In *Proceedings Of The 2011 Siam International Conference On Data Mining* (pp. 131-142). Society for Industrial and Applied Mathematics.
- Chen, T., Tang, L. A., Sun, Y., Chen, Z., & Zhang, K. (2016, July). Entity embedding-based anomaly detection for heterogeneous categorical events. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 1396-1403). AAAI Press.
- Fan, X., Da Xu, R. Y., & Cao, L. (2016, July). Copula Mixed-Membership Stochastic Blockmodel. In *IJCAI* (pp. 1462-1468)..
- Fan, X., Da Xu, R. Y., Cao, L., & Song, Y. (2017). Learning nonparametric relational models by conjugately incorporating node information in a network. *IEEE transactions on cybernetics*, *47*(3), 589-599..
- Fan, X., Cao, L., & Da Xu, R. Y. (2015). Dynamic infinite mixed-membership stochastic blockmodel. *IEEE transactions on neural networks and learning systems*, *26*(9), 2072-2085.
- Huang, Y. A., Fan, W., Lee, W., & Yu, P. S. (2003, May). Cross-feature analysis for detecting ad-hoc routing anomalies. In *Proceedings. 23rd International Conference on Distributed Computing Systems* (pp. 478-487). IEEE.

References

- Jian, S., Cao, L., Pang, G., & Lu, K., Gao, H. (2017 August). Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- Kim, D. I., Hughes, M., & Sudderth, E. (2012). The nonparametric metadata dependent relational model. *arXiv preprint arXiv:1206.6414*.
- Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. *Statistics and computing*, *26*(5), 1079-1099.
- Kriegel, H. P., Kröger, P., & Zimek, A. Outlier detection techniques. *Tutorial at KDD10*.
- Masthoff, J. (2015). Group recommender systems: aggregation, satisfaction and group attributes. In *Recommender Systems Handbook* (pp. 743-776). Springer US.
- Noto, K., Brodley, C., & Slonim, D. (2012). FRaC: a feature-modeling approach for semi-supervised and unsupervised anomaly detection. *Data mining and knowledge discovery*, *25*(1), 109-133.
- Pan W., E. W. Xiang, N. N. Liu, and Q. Yang. 2010. Transfer learning in collaborative filtering for sparsity reduction. In Proceedings of the 24th AAAI Conference on Artificial Intelligence 2010.
- Pang, G., Cao, L., & Chen, L., Liu, H. Unsupervised Feature Selection for Outlier Detection by Modelling Hierarchical Value-Feature Couplings. In *ICDM 2016* (pp. 410-419). IEEE.
- Pang, G., Cao, L., & Chen, L. (2016, July). Outlier detection in complex categorical data by modelling the feature value couplings. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 1902-1908). AAAI Press.
- Pang, G., Cao, L., & Chen, L., Liu, H. (2017 August). Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

References

- Rajan, V., & Bhattacharya, S. (2016, July). Dependency Clustering of Mixed Data with Gaussian Mixture Copulas. In *IJCAI* (pp. 1967-1973).
- Singh A. P. and Gordon G. J.. 2008. Relational learning via collective matrix factorization. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA2008 ACM, 1401969, 650–658.
- Tamersoy, A., Roundy, K., & Chau, D. H. (2014, August). Guilt by association: large scale malware detection by mining file-relation graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1524-1533). ACM.
- Wang, C., Cao, L., Wang, M., Li, J., Wei, W., & Ou, Y. (2011, October). Coupled nominal similarity in unsupervised learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 973-978). ACM.
- Wang, C., Dong, X., Zhou, F., Cao, L., & Chi, C. H. (2015). Coupled attribute similarity learning on categorical data. *IEEE transactions on neural networks and learning systems*, *26*(4), 781-797..
- Wang, Y., Li, B., Wang, Y., & Chen, F. (2015, June). Metadata dependent Mondrian processes. In *International Conference on Machine Learning* (pp. 1339-1347).
- Zhang, K., Wang, Q., Chen, Z., Marsic, I., Kumar, V., Jiang, G., & Zhang, J. (2015, June). From categorical to numerical: Multiple transitive distance learning and embedding. In *Proceedings of the 2015 SIAM International Conference on Data Mining* (pp. 46-54). Society for Industrial and Applied Mathematics.

References

Out-of-distribution detection:

- Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban and Mohammad Sabokrou. A Unified Survey on Anomaly, Novelty, Open-Set, and Out-of-Distribution Detection: Solutions and Future Challenges. CoRR2021.
- Jingkang Yang and Kaiyang Zhou and Yixuan Li and Ziwei Liu. Generalized Out-of-Distribution Detection. CoRR2021.
- Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, ICLR2017.
- Shiyu Liang, Yixuan Li and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks, ICLR2018.
- Kimin Lee, Kibok Lee, Honglak Lee and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks, NeurIPS2018.
- Yen-Chang Hsu, Yilin Shen, Hongxia Jin and Zsolt Kira. Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data, CVPR2020.
- Rui Huang and Yixuan Li. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space, CVPR2021.
- Dan Hendrycks, Mantas Mazeika and Thomas G. Dietterich. Deep Anomaly Detection with Outlier Exposure, ICLR2018.
- Kimin Lee, Honglak Lee, Kibok Lee and Jinwoo Shin. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples, ICLR2018.