

Non-IID Learning of Complex Data and Behaviors

Longbing Cao

University of Technology Sydney, Australia

Data Science Lab: www.datasciences.org
Non-IID Learning: noniid.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

Slides and info about non-IID learning

- <http://noniid.datasciences.org/>
- KDD2017 tutorial on non-IID learning Youtube videos:
<https://www.youtube.com/watch?v=3RwyGoiYcLg>

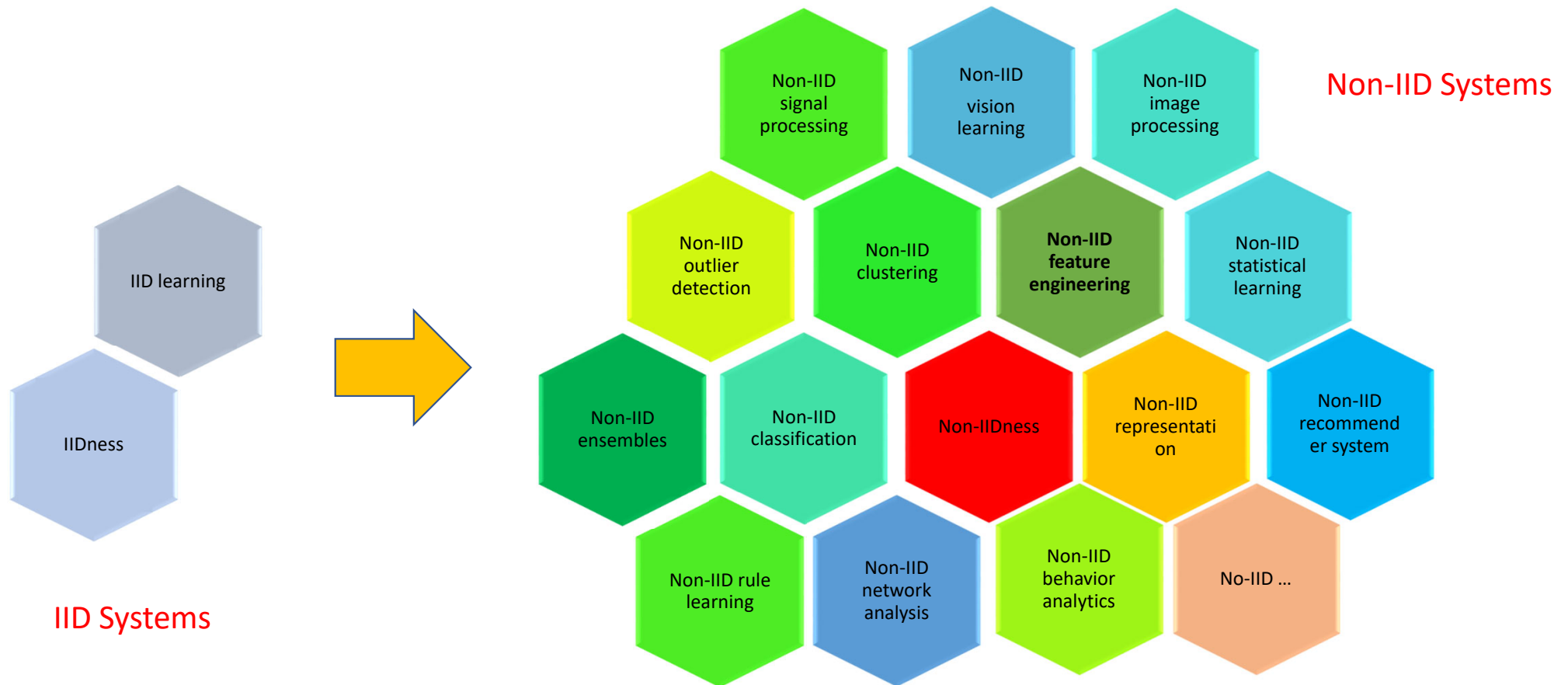
Agenda on non-IID Learning

Non-IID Learning



Related Work Overview

IID to Non-IID Learning Systems



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



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:

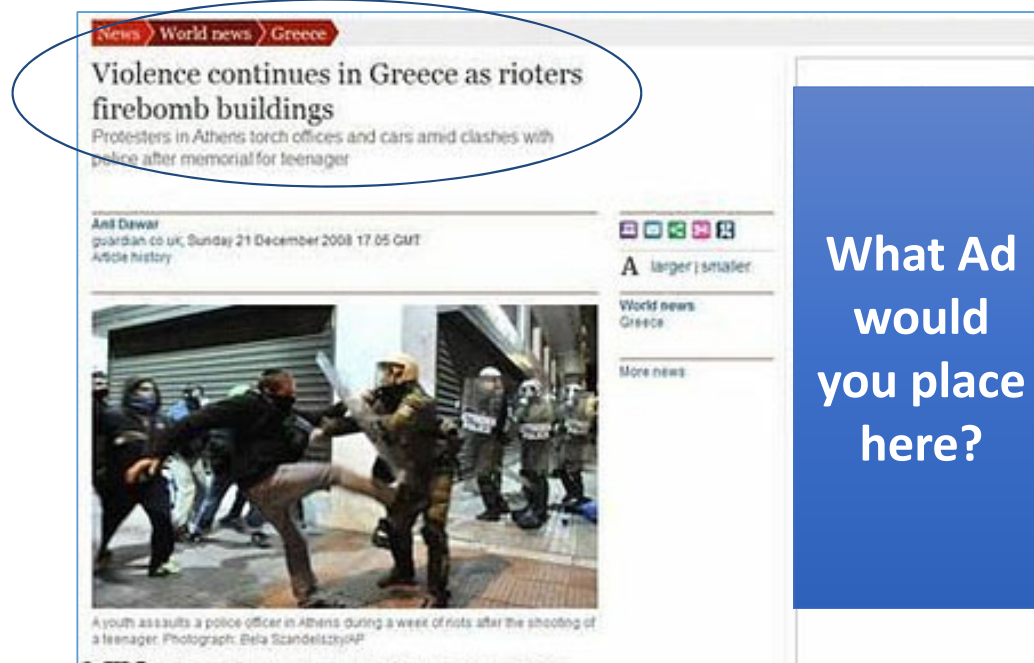
The topics covered under "Beyond IID" include but are not limited to the following:

- 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

IID Learning and Issues

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

Data Complexities: Challenge Existing Theories, Systems and Applications



**Irrelevant and
Damaging to Brand**

**What Ad
would
you place
here?**

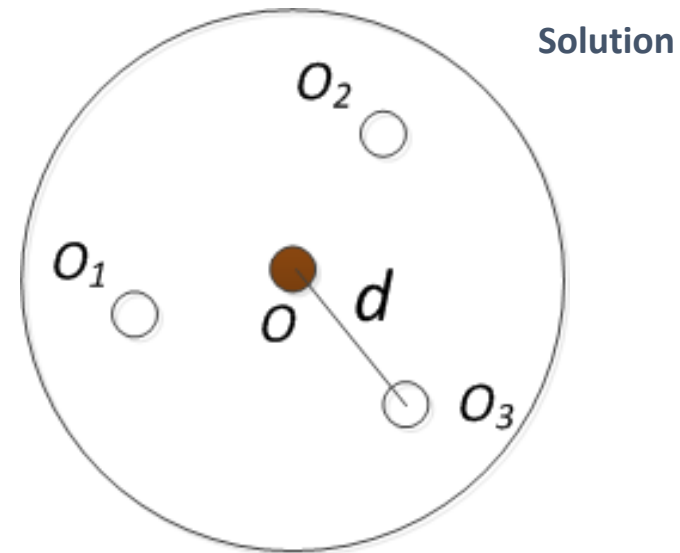
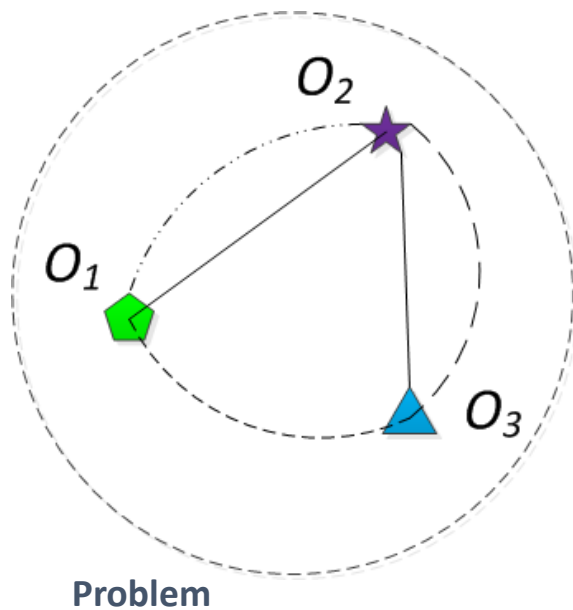
Why the Prediction Doesn't Work?

- 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, ...

Classic Assumption – IIDness & IID Learning

IID learning:

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



IIDness:

*Independence +
Identical Distribution*

O_1, O_2, O_3 are iid

$d_3 = ||O_3 - O||$

IID Learning

Traders are independent

Behaviors of a trader are totally or loosely independent

acct_id	trade_date	trade_time	sec_code	trade_price	trade_vol	trade_dir	seat_code	trade_bal
210266501	20090106	112138	600331	5.63	200	B	51721	200
315726605	20090106	92500	600477	7.4	400	B	73061	2000
315726605	20090106	92500	600477	7.4	1200	B	73061	3200
315726605	20090106	145838	600477	7.64	1600	S	73061	1600
315726605	20090107	93952	600477	7.67	1600	B	73061	3200
315726605	20090106	92500	600547	48	400	B	73061	1200
315726605	20090106	95552	600547	49.14	200	S	73061	1000
315726605	20090106	95756	600547	49.1	200	S	73061	800
783486703	20090106	92500	600547	49.14	200	S	73061	1000
783486703	20090106	92500	600547	49.14	200	S	73061	1000
783486703	20090106	92500	600547	49.14	200	S	73061	1000

Associations & frequent patterns

TID	Items
100	f, a, c, d, g, l, m, p
200	a, b, c, f, l, m, o
300	b, f, h, j, o
400	b, c, k, s, p
500	a, f, c, e, l, p, m, n

Foundation:

- Individual objects/behaviors
- Without coupling relationships (dependency) between objects/behaviors
- Focus on local features within an object/behavior

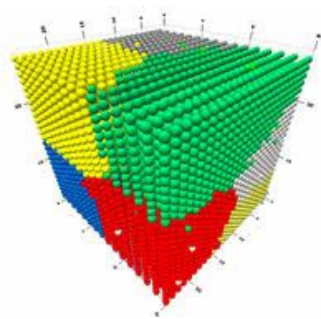
0.5-1.5 Q_0
Type C Type C Type B Type A Type B
>1.5 Q_0
Non-fissured

IID K-means

acct_id	trade_date	trade_time	sec_code	trade_price	trade_vol	trade_dir	seat_code	trade_bal
210266501	20090106	112138	600331	5.63	200	B	51721	200
315726605	20090106	92500	600477	7.4	400	B	73061	2000
315726605	20090106	92500	600477	7.4	1200	B	73061	3200
315726605	20090106	145838	600477	7.64	1600	S	73061	1600
315726605	20090107	93952	600477	7.67	1600	B	73061	3200
315726605	20090106	92500	600547	48	400	B	73061	1200
315726605	20090106	95552	600547	49.14	200	S	73061	1000
315726605	20090106	95756	600547	49.1	200	S	73061	800
783486703	20090106	92500	600001	3.32	1000	B	46451	6000
783486703	20090106	92500	600001	3.32	1000	B	46451	7000



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!

-FCM

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

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

Question:

- How about X_{j1} and X_{j2} dependent?

What Makes K-means IID?

Objective functions:

-K-means

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

- Object independency: \mathbf{x}_j do not consider interactions with other objects $\{\mathbf{x}_k\}$
- Object IIDness: assume \mathbf{x}_i for every cluster follows the same distribution
- Learning analytical goal: global \rightarrow local distribution
- Global mean \mathbf{x}_i

IID Decision Tree, KNN

Note:

- Dependence is on X_{ij} individual variables within an object (a branch represents an object)!
- Individual objects X

Question:

- How about if objects x_i and x_j are dependent?

acct_id	trade_date	trade_time	sec_code	trade_price	trade_vol	trade_dir	seat_code	trade_bal
210266501	20090106	112138	600331	5.63	200	B	51721	200
315726605	20090106	92500	600477	7.4	400	B	73061	2000
315726605	20090106	92500	600477	7.4	1200	B	73061	3200
315726605	20090106	145838	600477	7.64	1600	S	73061	1600
315726605	20090107	93952	600477	7.67	1600	B	73061	3200
315726605	20090106	92500	600547	48	400	B	73061	1200
315726605	20090106	95552	600547	49.14	200	S	73061	1000
315726605	20090106	95756	600547	49.1	200	S	73061	800
783486703	20090106	92500	600001	3.32	1000	B	46451	6000
783486703	20090106	92500	600001	3.32	1000	B	46451	7000

Objective functions:

-Decision tree

$$(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y)$$

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i) = \sum_{i=1}^m (f_i - f_i^2) = \sum_{i=1}^m f_i - \sum_{i=1}^m f_i^2 = 1 - \sum_{i=1}^m f_i^2$$

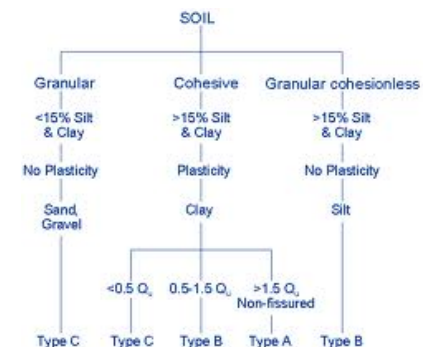
$$I_E(f) = - \sum_{i=1}^m f_i \log_2 f_i$$

-KNN

Euclidean distance: $d(x_1, x_2)$

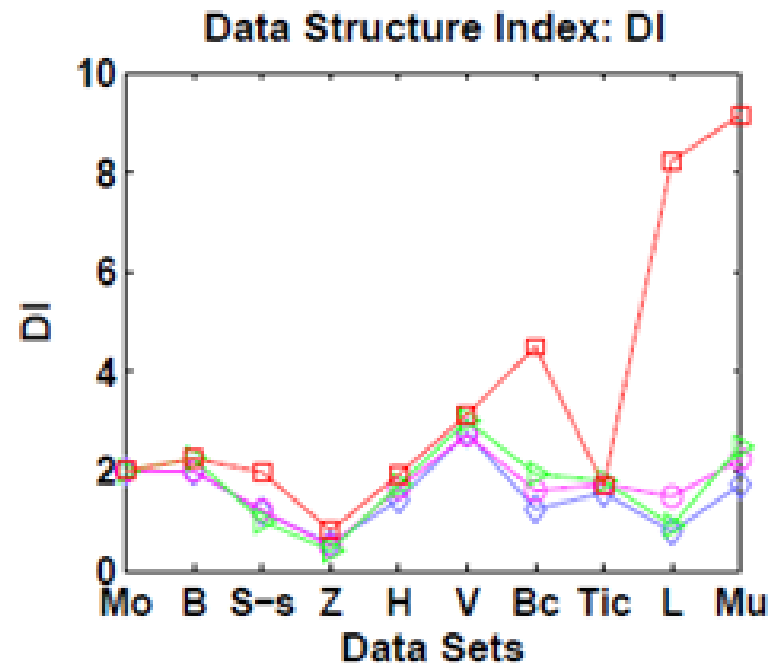
Hamming distance: $d(s_1, s_2)$

Classification



Potential Risk of IID Assumption

- Outcomes to be delivered by IID analytical/learning methods/algorithms on non-IID data could be:
 - incomplete
 - biased, or even
 - misleading



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)

Non-IIDness in Big and Small Data

- Heterogeneity:

- Data types, attributes, sources, aspects, ...
- Formats, structures, distributions, relations, ...
- Learning outcomes

Not identically distributed.

- Coupling relationships:

- Within and between values, attributes, objects, sources, aspects, ...
- Structures, distributions, relations, ...
- Methods, models, ...
- Outcomes, impact, ...

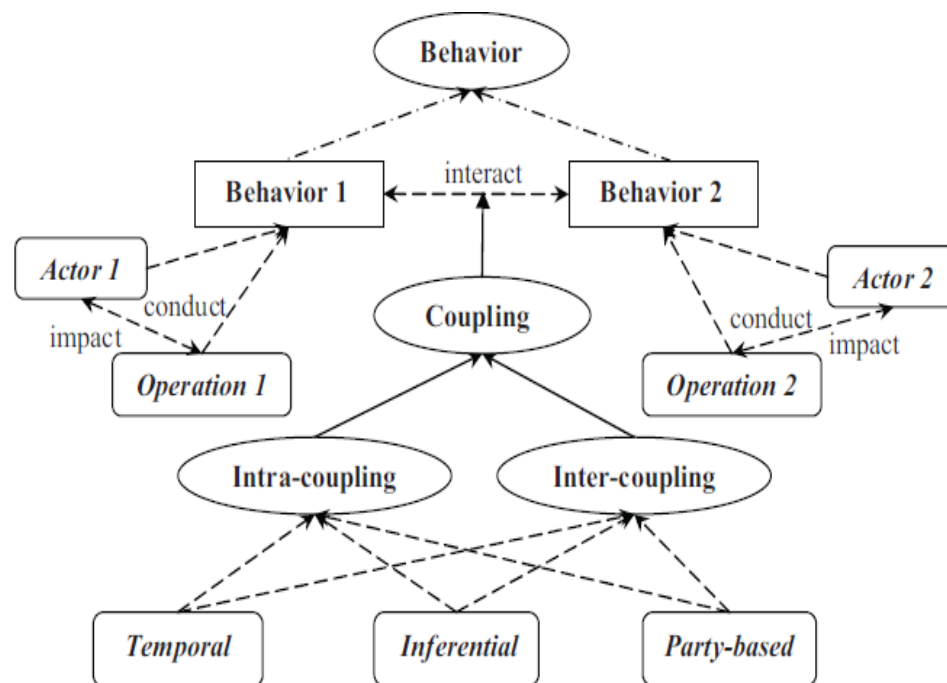
Not independent.

Non-IIDness

Couplings vs. Well Explored Relationships

- **Couplings**: numerical, categorical, textual, mixed-structure, syntactic, semantic, organizational, social, cultural, economic, uncertain, unknown/latent relation etc.
- **Coupling as a concept is much richer than existing terms including Dependence, Correlation, Association**
- **Dependence, Correlation, Association are much more specific, descriptive, explicit, etc.**
- **Coupling: explicit + implicit, qualitative + quantitative, descriptive + deep, specific + comprehensive, local + global, etc.**

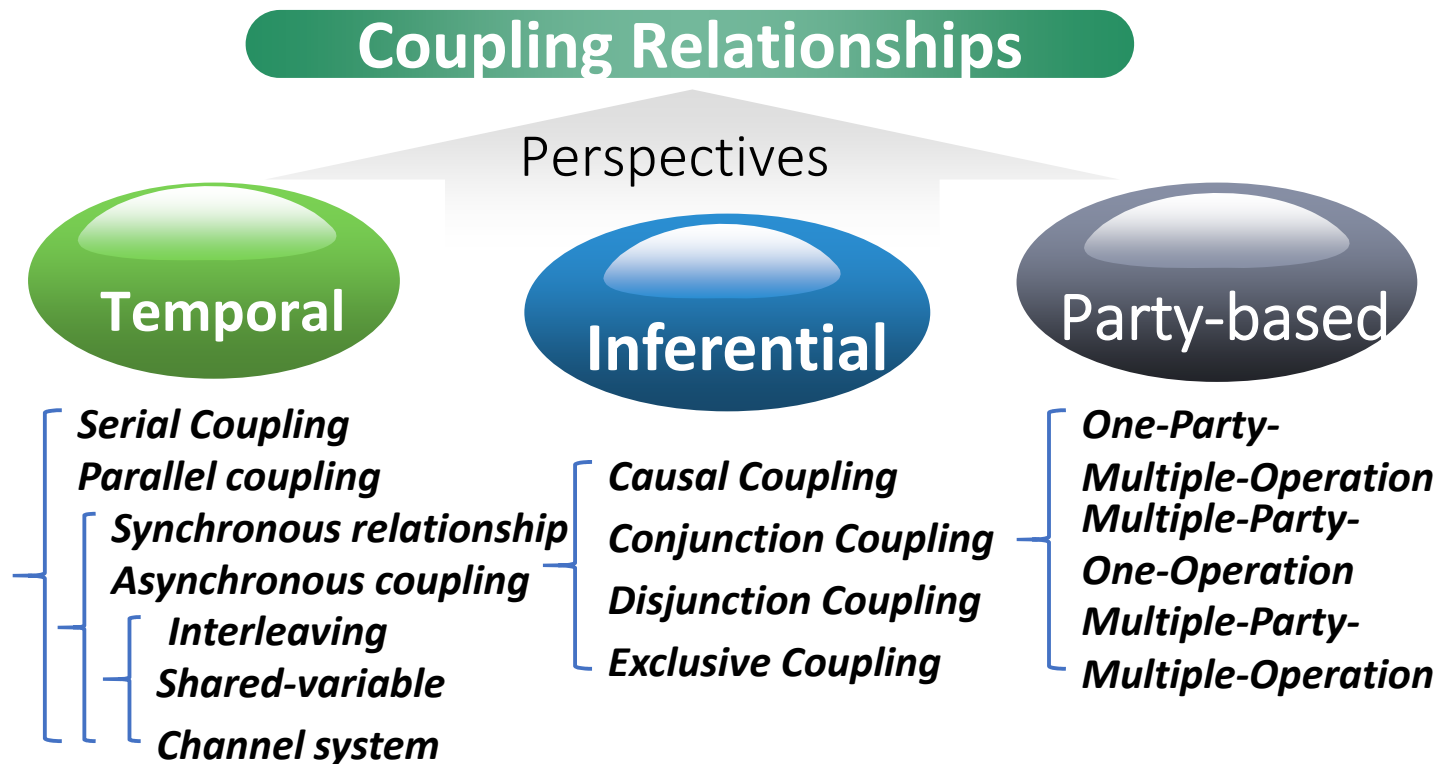
Example: Behavior Couplings



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

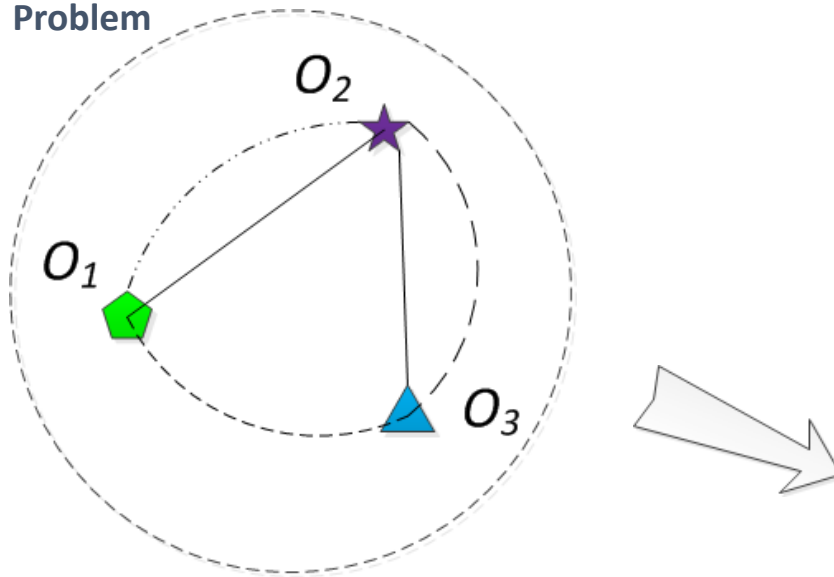
Can Wang, and Longbing Cao. [Modeling and Analysis of Social Activity Process](#), in Longbing Cao and Philip S Yu (eds) Behavior Computing, 21-35, Springer, 2012

Example: Couplings in Behavioral Data



A Foundational Issue: Non-IID Learning

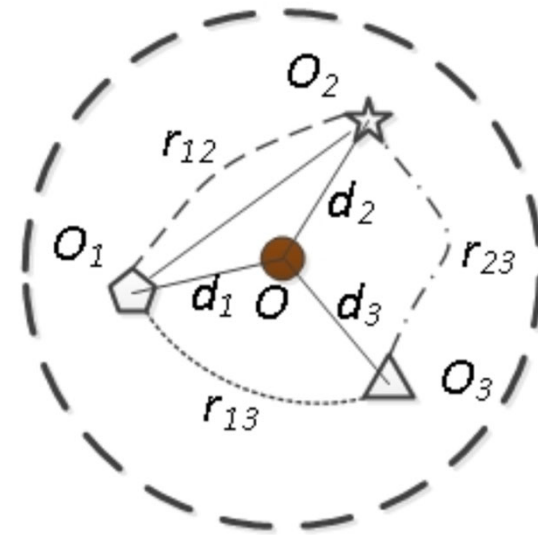
Problem



O_1, O_2, O_3 share different distributions

$$d_3 = ||O_3 - O||$$

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



Non-IID Similarity/Metric Learning

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

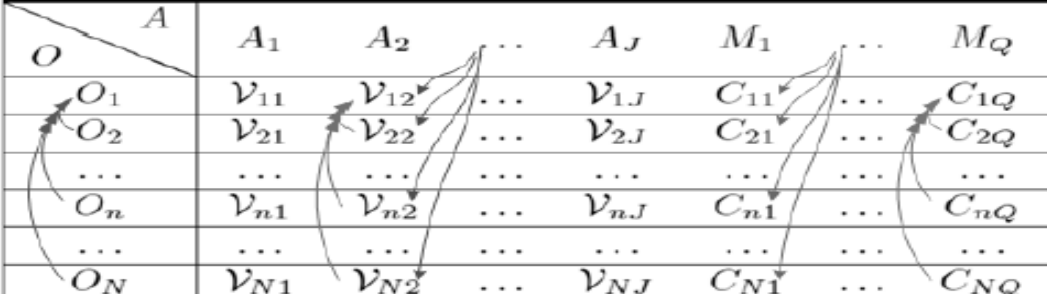


Why these two people
sit together at that
place at that
particular time?

Coupling Learning

TABLE 1. The Extended Information Table

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



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

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 Couplings

- Intra-attribute couplings

- indicate the involvement of attribute value occurrence frequency within one attribute
- how often the value occurs

- Inter-attribute couplings

- refer to the interaction between other attributes with this attribute
- reflect the extent of the value difference brought by other attributes

Hierarchical Coupling Relationships

$U \backslash A$	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

inter-coupled/inter-coupling

intra-coupled/intra-coupling

Set Information Function

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

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

$$f_j^*(\{u_{k_1}, \dots, u_{k_t}\}) = \{f_j(u_{k_1}), \dots, f_j(u_{k_t})\}, \quad (3.1)$$

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

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

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

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

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

$$f_2^*(\{u_1, u_2, u_3\}) = \{\dot{B}_1, B_2\}$$

$$g_2(B_1) = \{u_1, u_2\}$$

$$g_2^*(\{B_1, B_2\}) = \{u_1, u_2, u_3, u_6\}$$



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:*

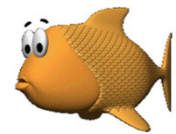
$$\delta_j^A(x, y) = \delta_j^{Ia}(x, y) \cdot \delta_j^{Ie}(x, y) \quad (4.1)$$

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

{ Intra-coupled Interaction:
Inter-coupled Interaction:

$\delta_j^{Ia}(x, y)$

$\delta_j^{Ie}(x, y)$



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

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

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

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

$$\delta_{j|k}^I = \sum_{v_k \in \cap} \min\{P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y)\}, \quad (4.8)$$

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:*

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

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-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:

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

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 $CASO(u_x, u_y)$:

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

Examples: Measuring Hierarchical Couplings

TABLE 4
Example of Computing Similarity Using *IRSP*

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

$U \backslash A$	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

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

Theoretical Analysis

- Computational Accuracy Equivalence:

THEOREM 5.1. *IRSP, IRSU, IRSJ and IRSI are all equivalent to one another.*²

IRSP \longleftrightarrow IRSU \longleftrightarrow IRSJ \longleftrightarrow IRSI

Complexity Analysis

- Computational Complexity Comparison:

Metric	Calculation Steps	Flops per Step	Complexity
<i>IRSP</i>	$nR(R-1)/2$	$2(n-1)2^R$	$O(n^2 R^2 2^R)$
<i>IRSU</i>	$nR(R-1)/2$	$2(n-1)R$	$O(n^2 R^2 R)$
<i>IRSJ</i>	$nR(R-1)/2$	$2(n-1)P$	$O(n^2 R^2 R)$
<i>IRSI</i>	$nR(R-1)/2$	$2(n-1)Q$	$O(n^2 R^2 R)$

$$2^R > R \geq P \geq Q$$



$$IRSP \geq IRSU \geq IRSJ \geq IRSI$$

R: The maximal number of attribute values.

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 $CASO(u_x, u_y)$.

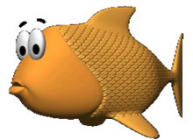
```
1 begin
  // Compute pairwise similarity for any
  // two values of the same attribute.
2  for attribute  $a_j, j = 1 : n$  do
3    for every value pair  $(v_j^x, v_j^y \in [1, |V_j|])$  do
4       $U_1 \leftarrow \{i | v_j^i == v_j^x\}, U_2 \leftarrow \{i | v_j^i == v_j^y\};$ 
      // Compute intra-coupled similarity
      // for two values  $v_j^x$  and  $v_j^y$ .
5       $\delta_j^{Ia}(v_j^x, v_j^y) = (|U_1| + |U_2|) / (|U_1| |U_2|);$ 
      // Compute coupled similarity for
      // two attribute values  $v_j^x$  and  $v_j^y$ .
6       $\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n) \leftarrow$ 
         $\delta_j^{Ia}(v_j^x, v_j^y) \cdot IeASV(v_j^x, v_j^y, \{V_k\}_{k \neq j});$ 

      // Compute coupled similarity between
      // two objects  $u_x$  and  $u_y$ .
7       $CASO(u_x, u_y) \leftarrow \text{sum}(\delta_j^A(v_j^x, v_j^y, \{V_k\}_{k=1}^n));$ 
8  end

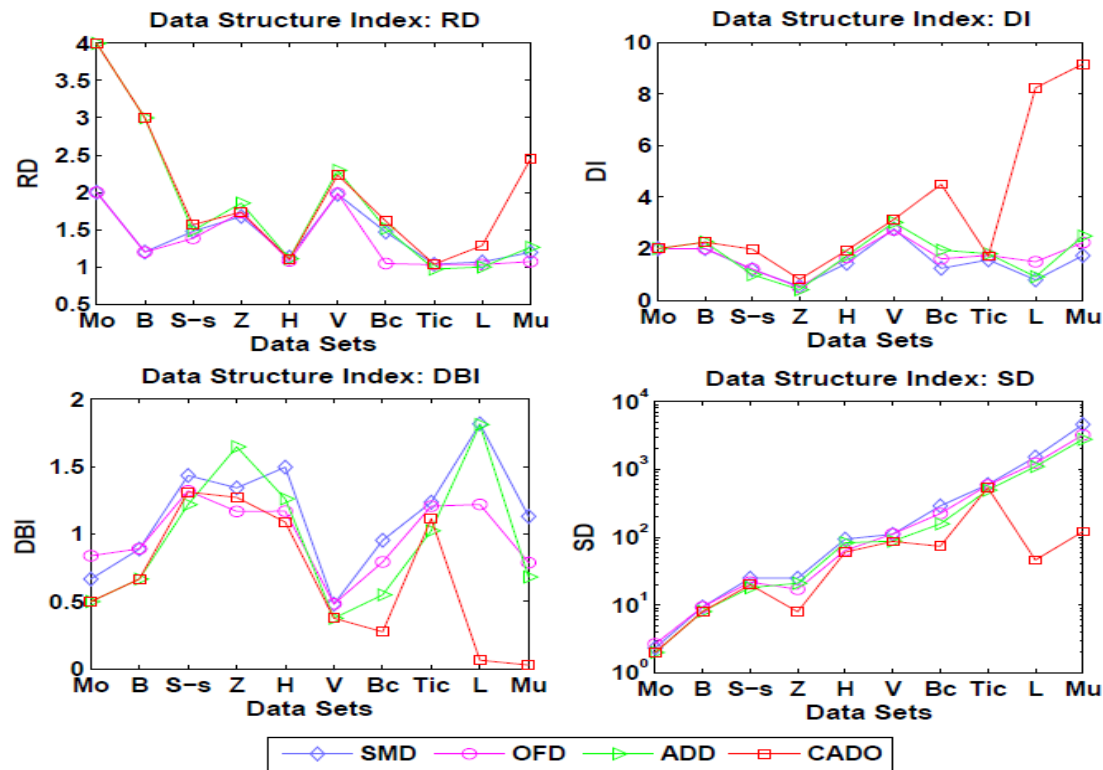
9 Function  $IeASV(v_j^x, v_j^y, \{V_k\}_{k \neq j})$ 
10 begin
  // Compute inter-coupled similarity for
  // two attribute values  $v_j^x$  and  $v_j^y$ .
11  for attribute  $(k = 1 : n) \wedge (k \neq j)$  do
12     $\{v_k^z\}_{z \in U_3} \leftarrow \{v_k^x\}_{x \in U_1} \cap \{v_k^y\}_{y \in U_2};$ 
13    for intersection  $z = U_3(1) : U_3(|U_3|)$  do
14       $U_0 \leftarrow \{i | v_k^i == v_k^z\};$ 
15       $ICP_x \leftarrow |U_0 \cap U_1| / |U_1|;$ 
16       $ICP_y \leftarrow |U_0 \cap U_2| / |U_2|;$ 
17       $Min_{(x,y)} \leftarrow \min(ICP_x, ICP_y);$ 
      // Compute  $IRSI$  for  $v_j^x$  and  $v_j^y$ .
18       $\delta_{j|k}^I(v_j^x, v_j^y, V_k) = \text{sum}(Min_{(x,y)});$ 
19       $\delta_j^{Ie}(x, y) = \text{sum}[\alpha(k) \times \delta_{j|k}^I(v_j^x, v_j^y, V_k)];$ 
20  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



Clustering performance indicator:

•Increasing

- Relative Dissimilarity (RD)
- Dunn Index (DI) [21]

•Decreasing:

- Davies-Bouldin Index (DBI) [20],
- Sum-Dissimilarity (SD)

Fig. 3. Data structure index comparison.

Applications – Clustering Performance

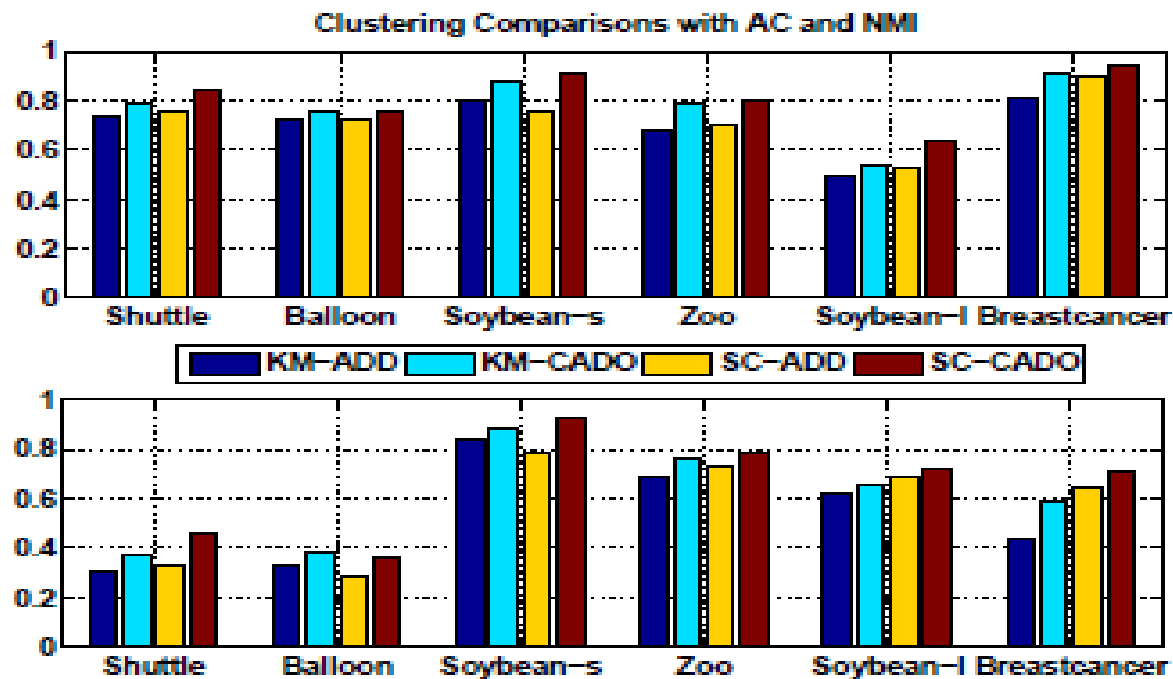
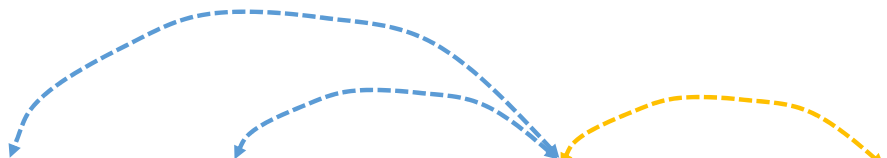


Fig. 4. Clustering evaluation on six data sets.

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



The diagram illustrates the motivation for a frequency-based distance metric. It shows a table with commitment levels H and I. A blue dashed arrow points from H to I, indicating a Hamming distance of 1. A yellow dashed arrow points from H to L, indicating a Hamming distance of 1. A blue solid arrow points from H to I, indicating a frequency-based distance of 0.

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

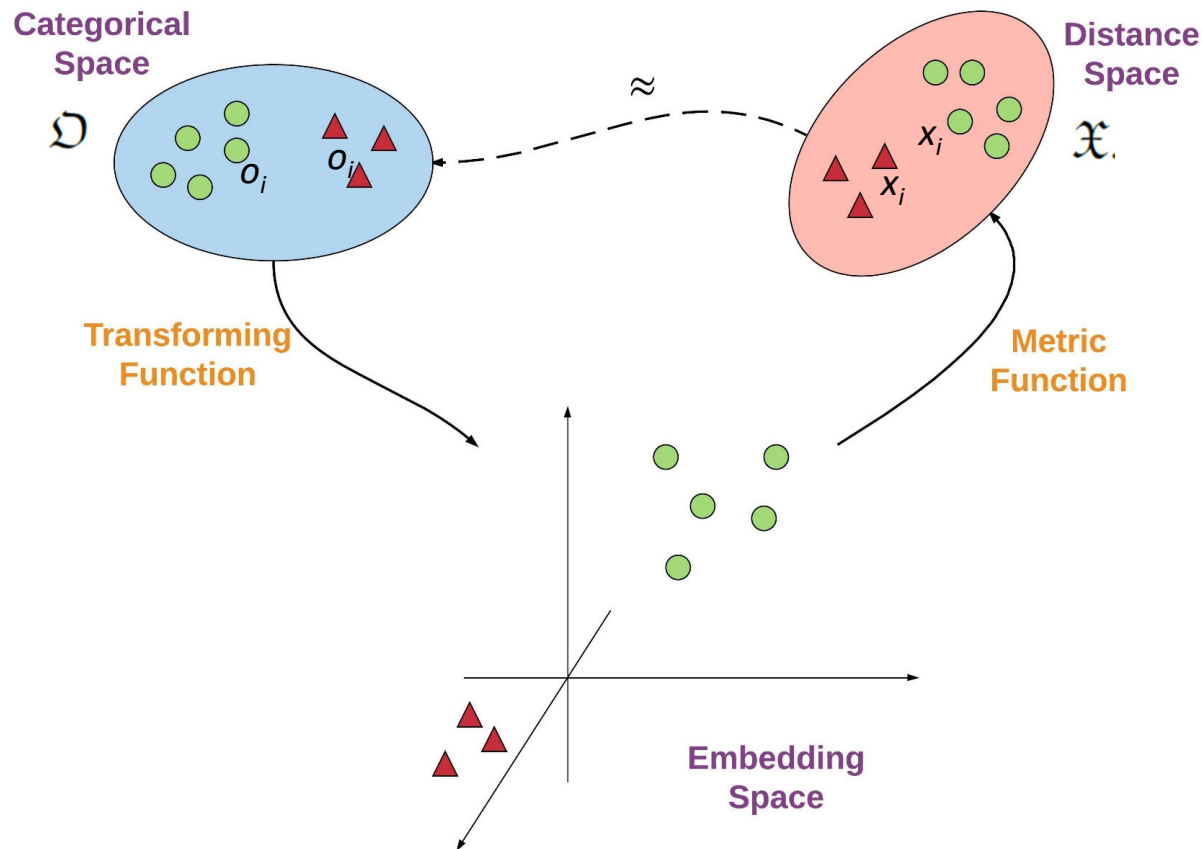
Hamming distance: $\text{Dis}(H, I) = \text{Dis}(H, L) = 1$

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

Frequency-based distance: $\text{Dis}(H, I) = 0$

H commitment is different from I.

Problem Statement

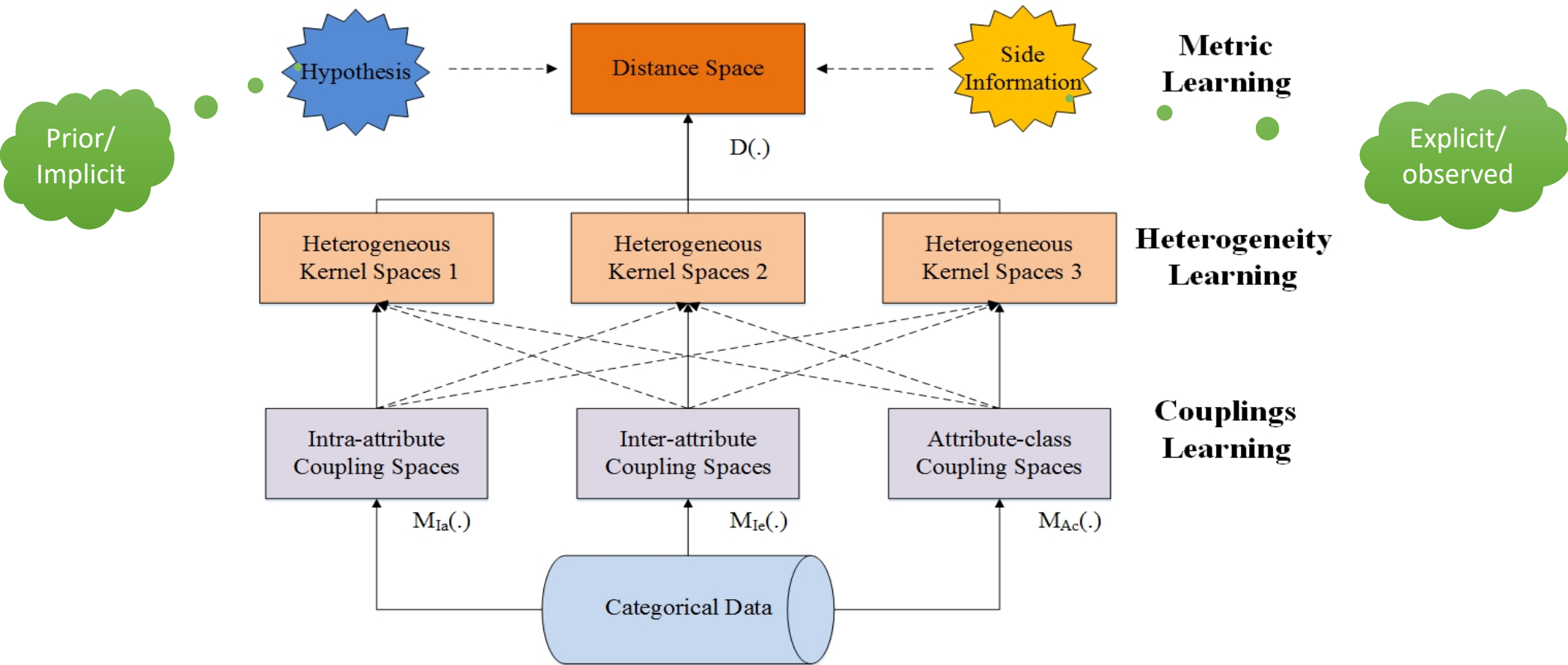


$$\begin{aligned}
 & \underset{\mathbf{x}}{\text{minimize}} && \widetilde{Div}(\mathcal{O} || \mathcal{X}) \\
 & \text{subject to} && \mathbf{o} \sim \mathcal{O} \\
 & && \mathbf{x} \sim \mathcal{X} \\
 & && d(\mathbf{o}_i, \mathbf{o}_j) = \mathbf{x}_i \odot \mathbf{x}_j.
 \end{aligned}$$

Distance metric $d(., .)$ satisfies:

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

HELIC Framework



HELIC: Heterogeneous Metric Learning with hierarchical Couplings

Learning Value-to-Class Couplings

Learning **Intra-attribute Couplings**

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

Capture value frequency

Learning **Inter-attribute Couplings**

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

Capture value co-occurrence

Learning **Attribute-class Couplings**

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

Capture value distribution in each class

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 $\{\mathbf{K}_1, \dots, \mathbf{K}_{n_k}\}$ can be obtained. Further, a set of transformation matrices $\{\mathbf{T}_1, \dots, \mathbf{T}_{n_k}\}$ 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 the \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 :

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

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

The distance can be represented as

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

$$\omega = \begin{bmatrix} \omega_1^{\text{diag}} & 0 & \cdots & 0 \\ 0 & \omega_2^{\text{diag}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{n_k}^{\text{diag}} \end{bmatrix}$$

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

Metric Learning

Objective function:

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

Selecting the kernels for their sensitive data distribution

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

Theoretical Analysis

Generalization Error Bound

$$\begin{aligned} \varepsilon(\boldsymbol{\omega}, b) - \varepsilon_{\mathcal{Z}}(\boldsymbol{\omega}, b) \leq & 2(1 + 1/\sqrt{\lambda})\sqrt{2\ln(1/\delta)/n_o} \\ & + \left(8 + 16\sqrt{e\ln(n_on_k)}\right) / \sqrt{n_o\lambda} + 12/\sqrt{n_o} \end{aligned}$$

Time Complexity

$$O(n_v(n_c + 1) + n_{mv}^2 n_a^2 + n_b n_\omega n_{step})$$

Space Complexity

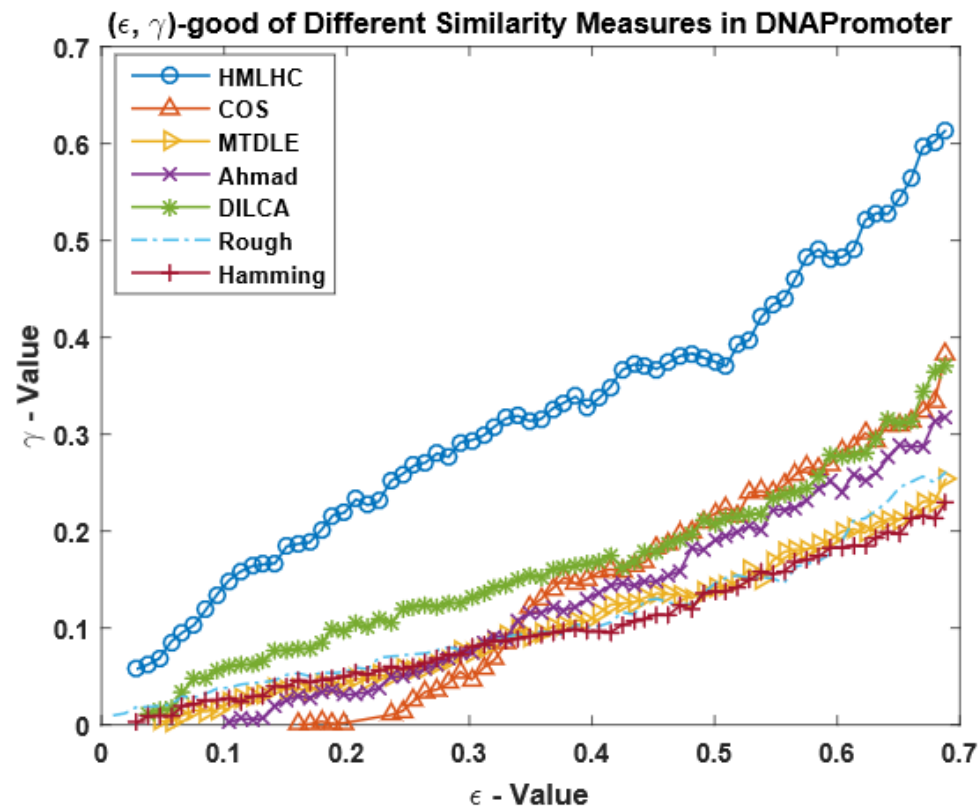
$$O(n_b n_\omega)$$

Representation Performance of HELIC

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

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

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

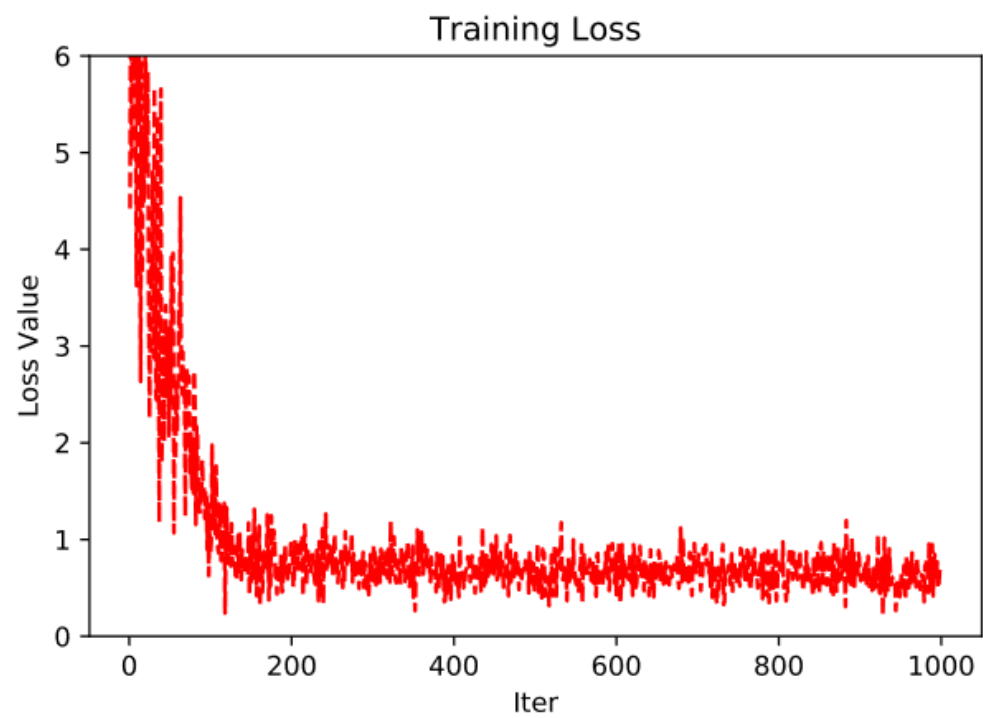
Flexibility of HELIC

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

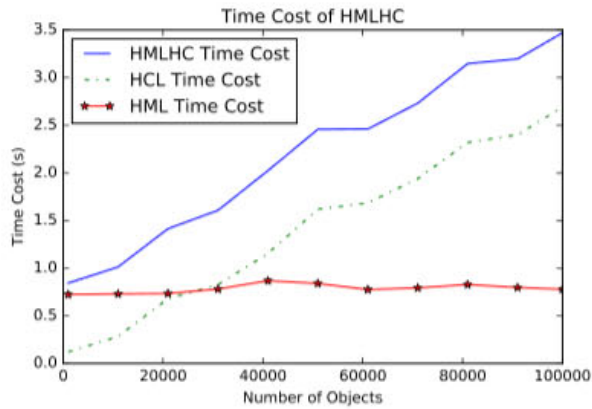
Data	HELIC-LR	MTDLE-LR	$\Delta\%$	HELIC-RF	MTDLE-RF	$\Delta\%$	HELIC-SVM	MTDLE-SVM	$\Delta\%$
Zoo	100	92.50 \pm 11.75	8.11%	100	99.64 \pm 1.63	0.36%	100	100	0%
DNAPromoter	98.48 \pm 3.70	89.84 \pm 10.89	9.62%	93.88 \pm 9.02	74.87 \pm 11.89	25.39%	97.98 \pm 4.15	89.88 \pm 10.35	9.01%
Hayesroth	83.56 \pm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

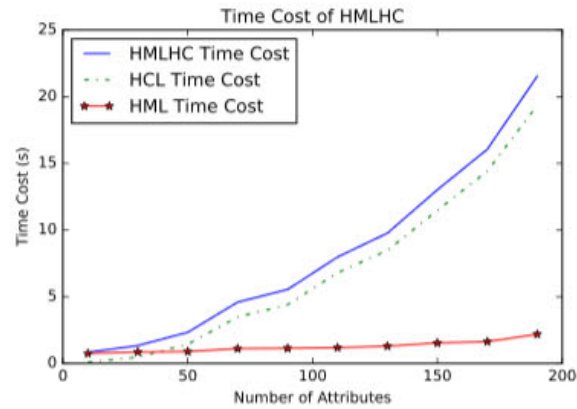
Scalability of HELIC



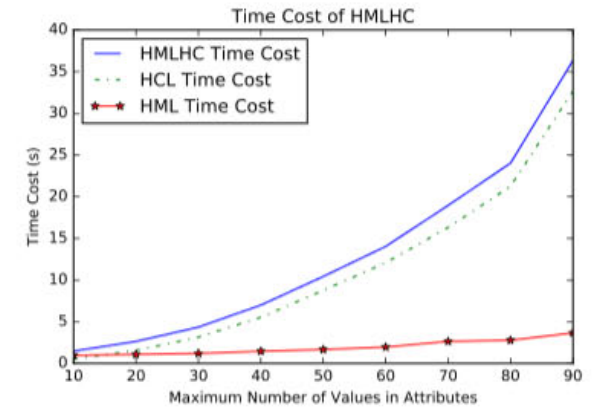
Scalability of HELIC



(a) Time Cost v.s. Number of Objects.



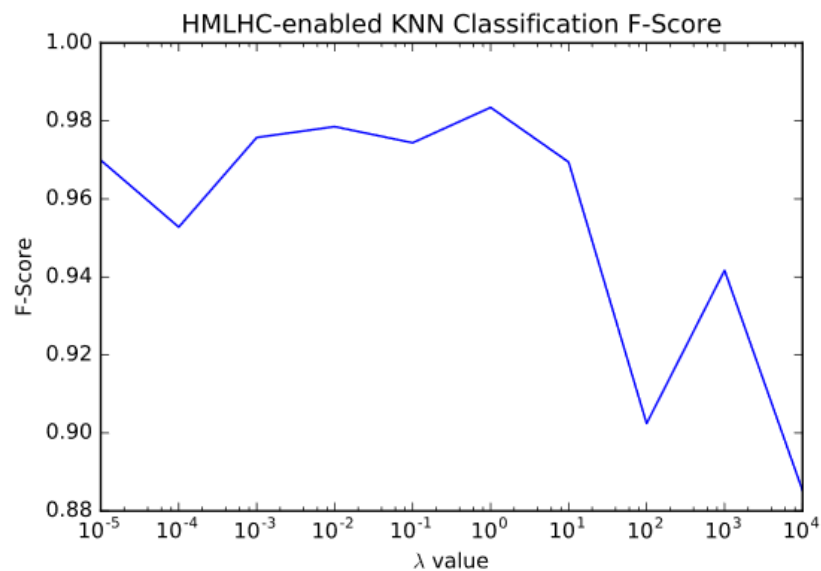
(b) Time Cost v.s. Number of Attributes.



(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.

Stability of HELIC



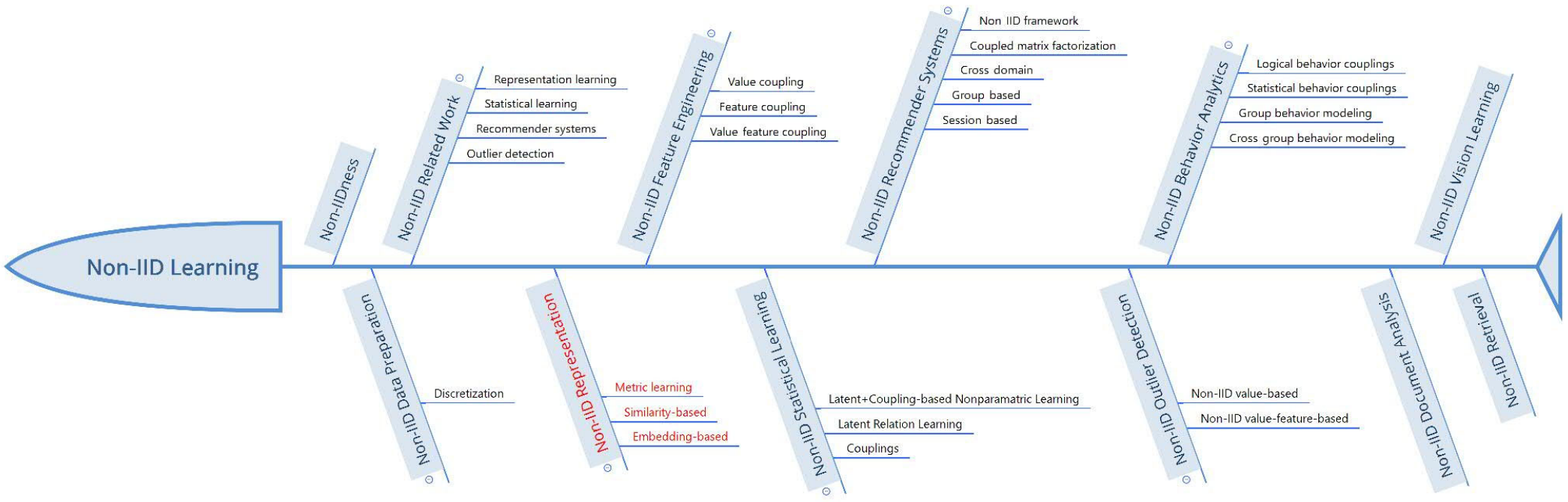
- ✓ The only parameter needs to tune in HELIC is λ .
- ✓ HELIC is stable for a large range of λ especially when λ is less than 1.

The HELIC-enabled KNN Classification F-score under Different Setting of Parameter λ .

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



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

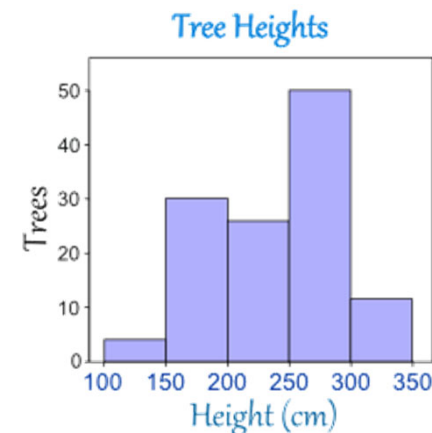
Representation of numerical features

- 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	$\frac{X - \bar{X}}{s}$
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 into categorical one
 - Discretization
- Transform categorical data into 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	$1 + \log f_{t,d}$	$\log\left(1 + \frac{N}{n_t}\right)$
3	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \cdot \log \frac{N}{n_t}$

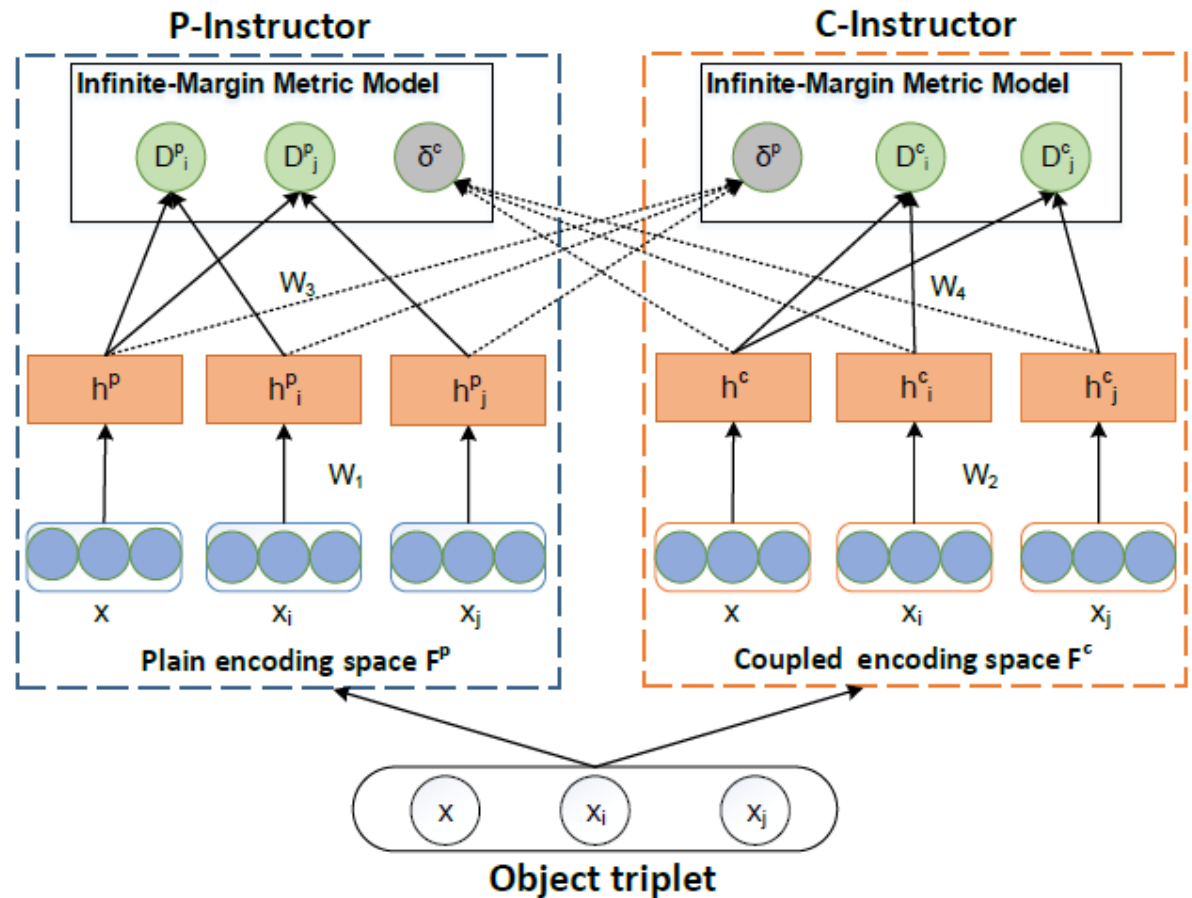
Name	Gender	Height
Alice	Female	1.75 m
Bob	Male	1.75 m

What is a good representation for mixed data?

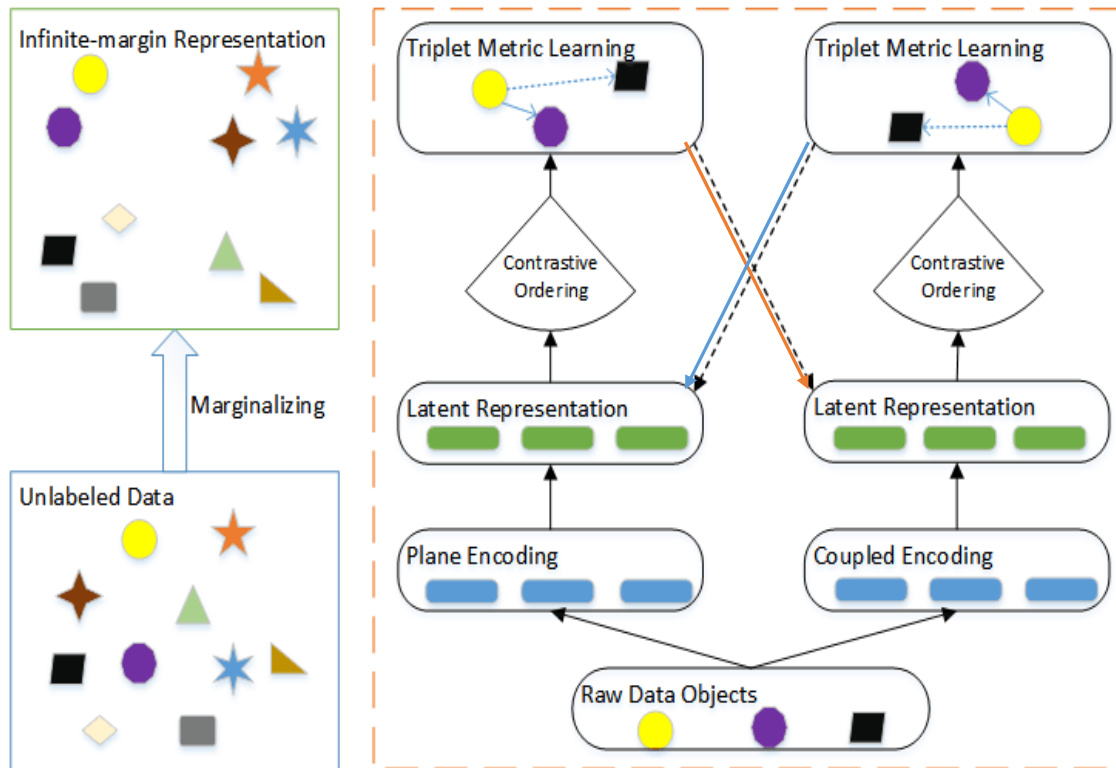
- At **the feature level**: capture the **heterogeneous coupling** (e.g., complex interactions, dependencies) between features
 - Couplings between categorical features
 - Couplings between numerical features
 - Couplings between categorical and numerical features
- At **the object level** , a good representation 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



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

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

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

Experiments

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

Table 1: Statistics of UCI datasets

Datasets	$ \mathcal{X} $	$ \mathcal{F}^c $	$ \mathcal{F}^n $	$ Class $
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 \pm 0.1033	0.1749 \pm 0.0444	0.1237 \pm 0.1147	0.2493 \pm 0.0207	0.3246\pm0.0000	0.3304\pm0.0000
Hepatitis	0.1453 \pm 0.0703	0.1761 \pm 0.0292	0.1532 \pm 0.0342	0.1689 \pm 0.0163	0.1848\pm0.0000	0.1905\pm0.0000
MPG	0.1490 \pm 0.0106	0.1477 \pm 0.0184	0.1373 \pm 0.0347	0.1536 \pm 0.0086	0.1831\pm0.0232	0.1770\pm0.0000
Heart	0.3130\pm0.0688	0.1439 \pm 0.0642	0.1037 \pm 0.1215	0.3302\pm0.0042	0.2632 \pm 0.0000	0.2774 \pm 0.0000
ACA	0.3204 \pm 0.1518	0.3433 \pm 0.1726	0.3182 \pm 0.0627	0.3477 \pm 0.0844	0.4258\pm0.0000	0.4258\pm0.0000
CRX	0.2322 \pm 0.1191	0.0836 \pm 0.1109	0.2714 \pm 0.1361	0.1445 \pm 0.1477	0.4267\pm0.0000	0.4267\pm0.0000
CMC	0.0293 \pm 0.0052	0.0269 \pm 0.0013	0.0333\pm0.0070	0.0292 \pm 0.0037	0.0327\pm0.0077	0.0303 \pm 0.0081
Income	0.1139 \pm 0.0361	0.1414\pm0.0291	0.1258 \pm 0.0658	0.1314 \pm 0.0000	0.1325\pm0.0000	0.1325\pm0.0000
Average	0.1853 \pm 0.0707	0.1547 \pm 0.0588	0.1583 \pm 0.0722	0.1944 \pm 0.0353	0.2467\pm0.0064	0.2488\pm0.0010

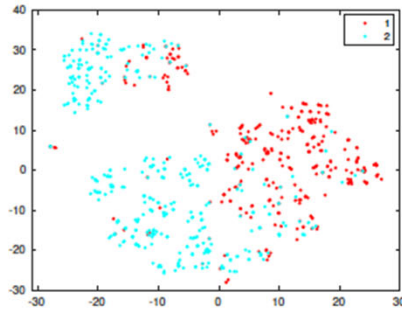
Table 3: DBSCAN clustering performance w.r.t. AMI/Clusters.

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

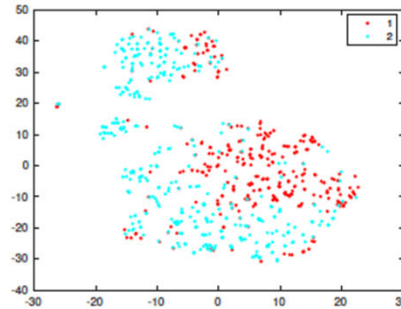
Table 4: Calinski-Harabasz index on representation w.r.t. the Euclidean distance for ground-truth labels

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

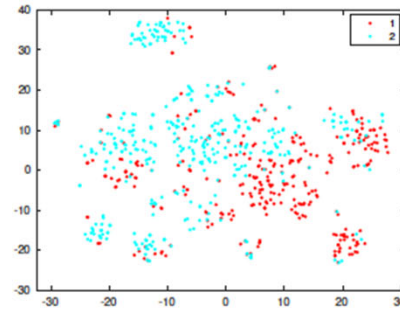
Visualization



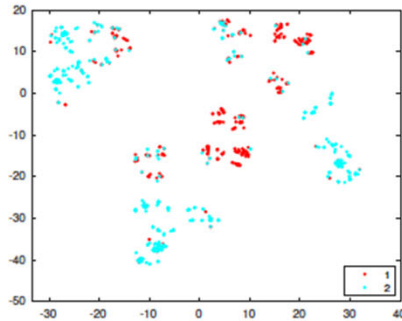
(a) Plain encoding



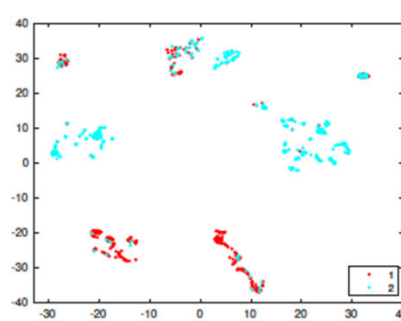
(b) Coupled encoding



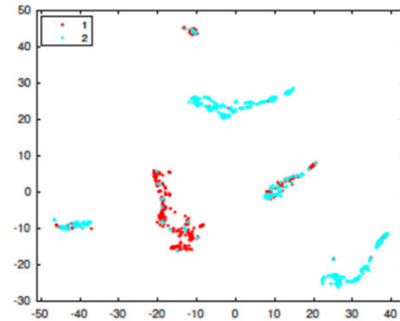
(c) CoupledMC



(d) Autoencoder



(e) MAI-F



(f) MAI-D

Conclusion

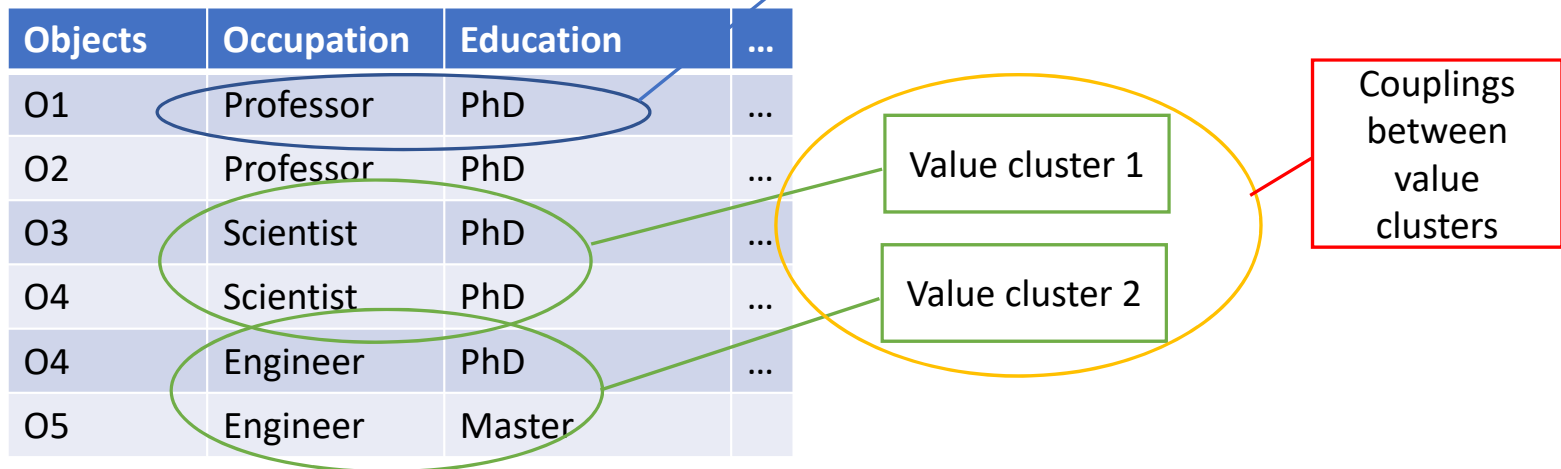
- A comprehensive representation for mixed data simultaneously learns the couplings at feature level and the discrimination between objects at the object level.
- A metric-based auto-instructor (MAI) model with two collaborative instructors learns more discriminative representation between objects by 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.

Embedding-based Representation

Songlei Jian, Longbing Cao, Guansong Pang, Kai Lu, Hang Gao. Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning. IJCAI 2017

Motivation

- Hierarchical value couplings in data
 - Pairwise value couplings
 - Multi-granularity value clusters
 - Couplings between value clusters



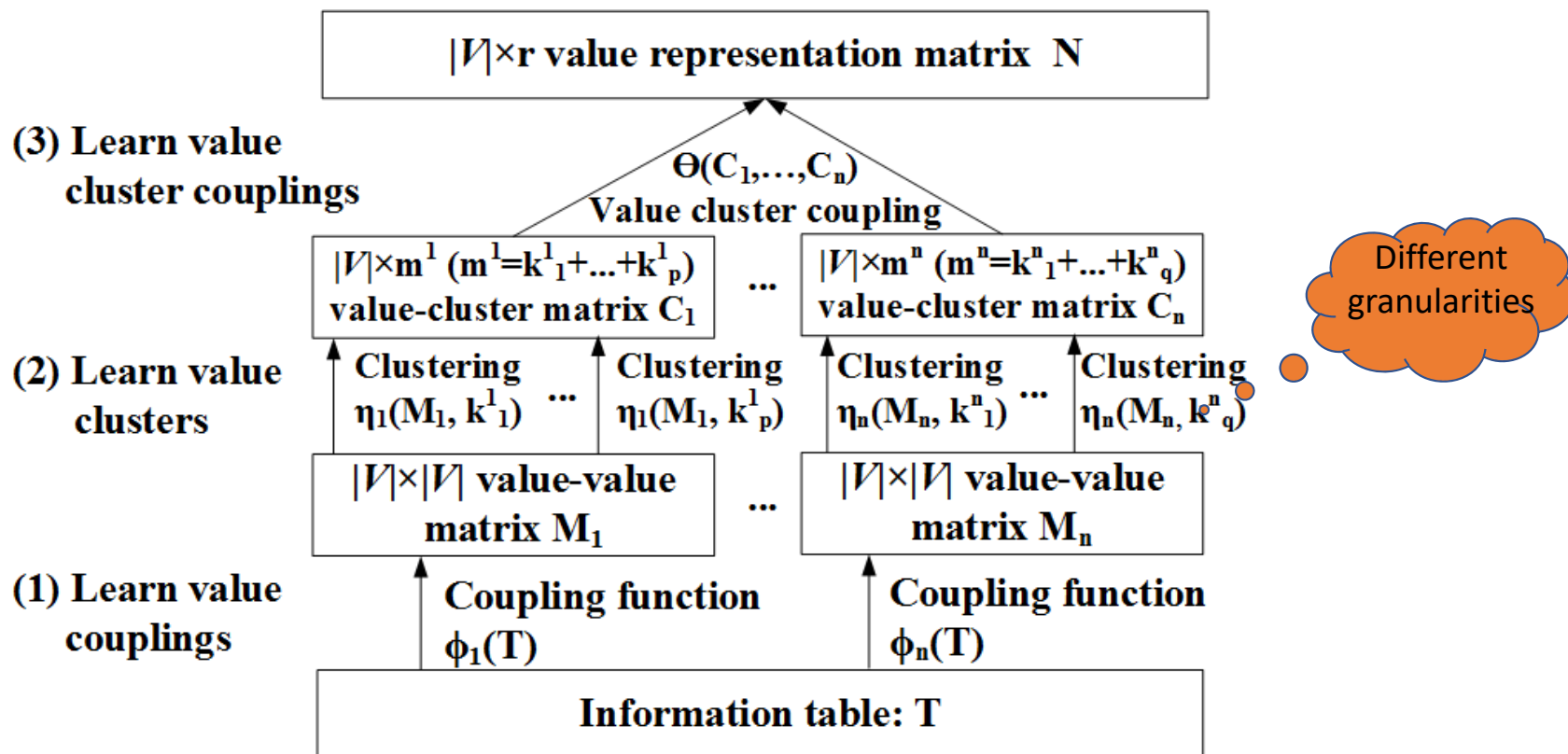
Related work

- Representation for categorical data
 - Embedding-based representation
 - One-hot encoding
 - IDF encoding
 - Similarity-based representation
 - Pairwise couplings based methods
- Gaps for representation
 - Ignore the intrinsic data dependency and interactions within values

The CURE Framework

- A novel Coupled Unsupervised Representation framework (CURE for short) to capture the hierarchical value couplings in data representation
- We instantiate CURE into an Coupled Data Embedding (CDE) method for clustering.

The CURE Framework



Learning Complementary Value Couplings

- Occurrence-based Value Influence Matrix
- Co-occurrence-based Value Influence Matrix

$$\mathbf{M}_o = \begin{bmatrix} \phi_o(v_1, v_1) & \dots & \phi_o(v_1, v_l) \\ \vdots & \ddots & \vdots \\ \phi_o(v_l, v_1) & \dots & \phi_o(v_l, v_l) \end{bmatrix}$$

Coupling function:

$$\phi_o(v_i, v_j) = \psi(f^i, f^j) \times \frac{p(v_j)}{p(v_i)}$$

$$\mathbf{M}_c = \begin{bmatrix} \phi_c(v_1, v_1) & \dots & \phi_c(v_1, v_l) \\ \vdots & \ddots & \vdots \\ \phi_c(v_l, v_1) & \dots & \phi_c(v_l, v_l) \end{bmatrix}$$

Coupling function:

$$\phi_c(v_i, v_j) = \frac{p(v_i, v_j)}{p(v_i)}$$

The Main Idea in CDE

- Build two value coupling matrices
 - Occurrence-based Value Influence Matrix
 - Co-occurrence-based Value Influence Matrix
- Generate value clusters with different granularities on value coupling matrices
 - K-means clustering with different parameters
- Learn correlation between different value clusters
 - Use PCA to learn linear correlation

Algorithm

Algorithm 1 *Value Embedding* ($\mathcal{D}, \alpha, \beta$)

Input: \mathcal{D} - data set, α - proportion factor, β - dimension reducing factor

Output: \mathbf{N} - the numerical representation of all values

```
1: Generate  $\mathbf{M}_o$  and  $\mathbf{M}_c$ 
2: Initialize  $\mathbf{I} = \emptyset$ 
3: for  $\mathbf{M} \in \{\mathbf{M}_o, \mathbf{M}_c\}$  do
4:   Initialize  $k = 2$ 
5:    $rm = \emptyset$ 
6:   repeat
7:      $\mathbf{I} = [\mathbf{I}; kmeans(\mathbf{M}, k)]$ 
8:     Remove the cluster with only one value and store
       the remove cluster in  $rm$ 
9:      $k+ = 1$ 
10:  until  $length(rm) \geq \lceil \frac{k}{\alpha} \rceil$ 
11: end for
12:  $\mathbf{X} = \mathbf{I} - mean(\mathbf{I})$ 
13: Calculate the covariance matrix  $\mathbf{S}$  of  $\mathbf{X}$ 
14:  $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = SVD(\mathbf{S})$ 
15:  $\mathbf{N} = \mathbf{XV}^T$ 
16: Remove the columns whose maximum Euclidean distance
   of any two elements is less than  $\beta$  from  $\mathbf{N}$ 
17: return  $\mathbf{N}$ 
```

- \mathbf{N} : Value embedding

$$\mathbf{N} = \mathbf{XV}^T,$$

- \mathbf{X} : Centralized matrix of indicator matrix \mathbf{I}
- \mathbf{V} : principal component matrix from SVD of \mathbf{S}
- \mathbf{S} : Covariance matrix from \mathbf{X}

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}.$$

Experiments

- Comparison with Embedding Methods

Basic data info. & Data Factor				F-score			
Data	O	V	FCI	CDE	0-1	0-1P	IDF
Wisconsin	683	89	0.212	0.967	0.946	0.946	0.943
Soybeansmall	47	58	0.180	0.915	0.829	0.854	0.763
Mushroom	5644	97	0.148	0.731	0.709	0.694	0.506
Mammographic	830	20	0.116	0.809	0.793	0.815	0.517
Zoo	101	30	0.110	0.647	0.596	0.607	0.537
Dermatology	366	129	0.089	0.670	0.598	0.606	0.616
Hepatitis	155	36	0.085	0.680	0.681	0.667	0.535
Adult	30162	98	0.060	0.654	0.585	0.588	0.479
Lymphography	148	59	0.057	0.418	0.381	0.379	0.561
Primarytumor	339	42	0.020	0.240	0.230	0.238	0.190
Average				0.673	0.635	0.640	0.565
				p-value	0.003	0.003	0.020

CDE has an approximate 9%, 5% and 19% improvement over 0-1, 0-1P and IDF.

FCI is data indicator which measures the average correlation strength between features.

For most data sets with higher FCI, CDE outperforms the other embedding methods

Experiments

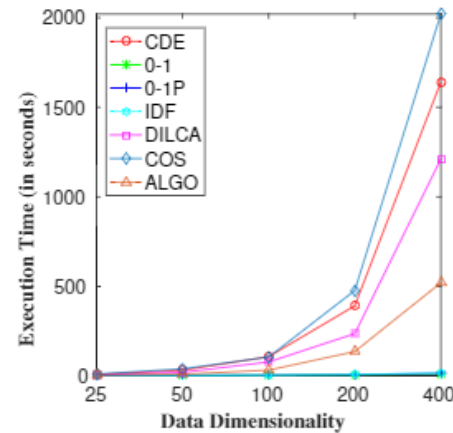
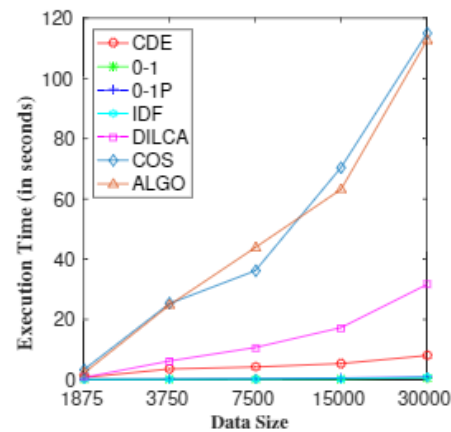
- Comparison with Similarity Measures

Clustering Info & Data Factor			F-score			
Data	$ C $	VCI	CDE-G	COS	DILCA	ALGO
Primarytumor	21	0.873	0.242	0.196	0.224	0.209
Zoo	7	0.733	0.644	0.538	0.583	0.547
Soybeansmall	4	0.712	1.000	0.893	0.910	0.911
Lymphography	4	0.699	0.397	0.395	0.353	0.366
Dermatology	6	0.664	0.784	0.730	0.808	0.710
Mushroom	2	0.310	0.828	0.825	0.826	0.826
Wisconsin	2	0.237	0.962	0.973	0.921	0.971
Hepatitis	2	0.141	0.667	0.463	0.679	0.662
Mammographic	2	0.071	0.817	0.828	0.826	0.818
Adult	2	0.032	0.676	NA	NA	NA
Average			0.762	0.706	0.738	0.726
			p-value	0.050	0.100	0.032

CDE has an approximate 8%, 3% and 5% improvement over COS, DILCA and ALGO respectively in terms of F-score. VCI is data indicator which reflects the discriminative ability of the value clusters in object classes. For most data sets with higher VCI, CDE outperforms the other similarity methods.

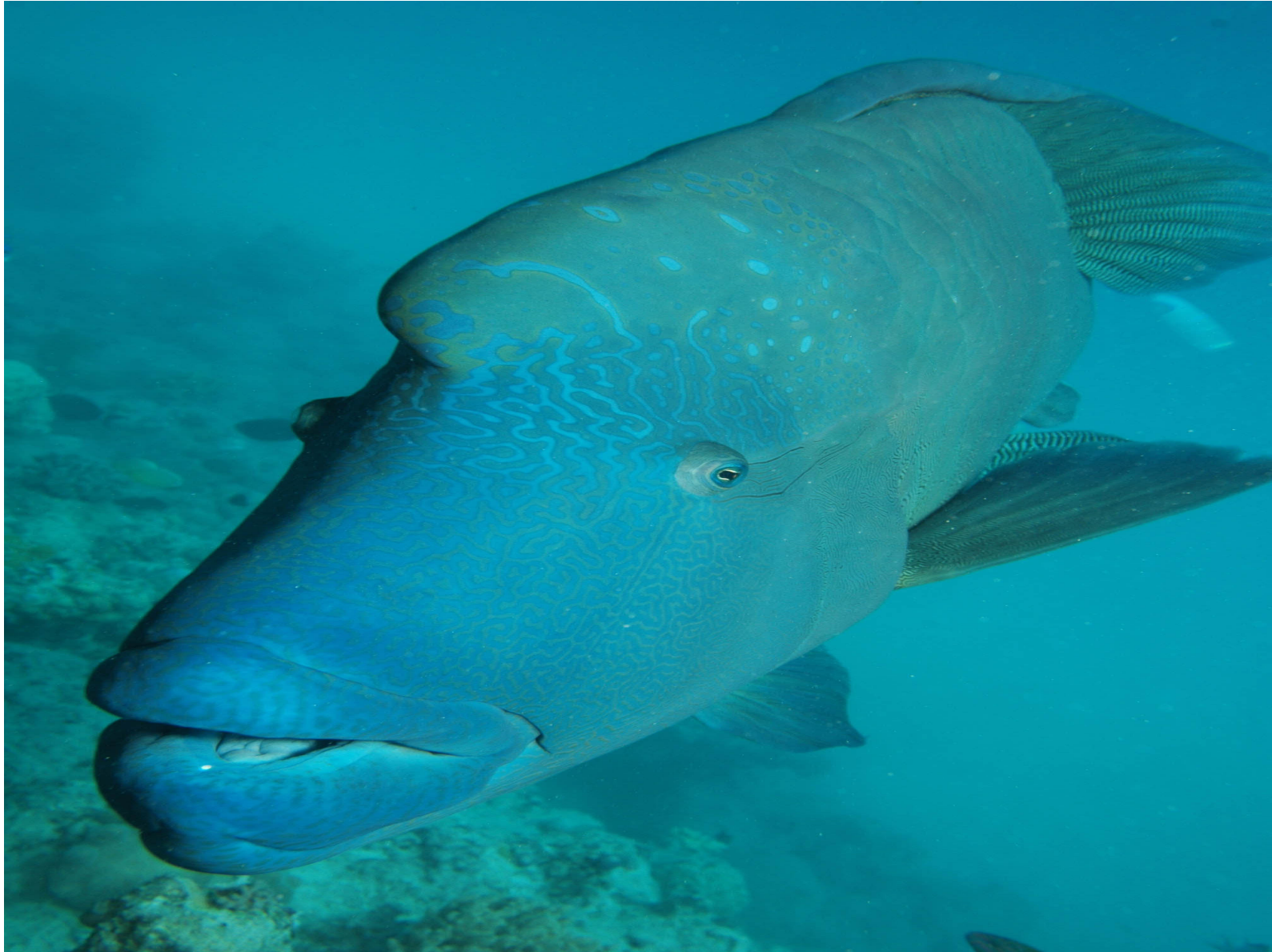
Experiments

- Good scalability w.r.t. data size and dimensionality
 - Linear with data size and quadratic with dimensionality



Conclusions

- Different from existing encoding-based embedding and feature correlation-based similarity measures, a novel unsupervised representation framework (CURE) and its instantiation (CDE) are introduced in this paper, which model hierarchical value couplings in terms of feature interactions and value clustering.
- Extensive experiments show that CDE significantly outperforms typical embedding methods and similarity measures in capturing feature value interactions. In addition, two proposed data factors further indicate the feature value couplings and value clusters in data sets.



Non-IID Ensemble Clustering

Can Wang, Zhong She, Longbing Cao. [Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects](#), ICDE2013.

Introduction

- **Clustering ensemble** has exhibited great potential in enhancing the clustering accuracy, robustness and parallelism by combining results from various clustering methods.
- The whole **process of clustering ensemble**
 - building base clusterings
 - aggregating base clusterings
 - post-processing clustering.

Problems

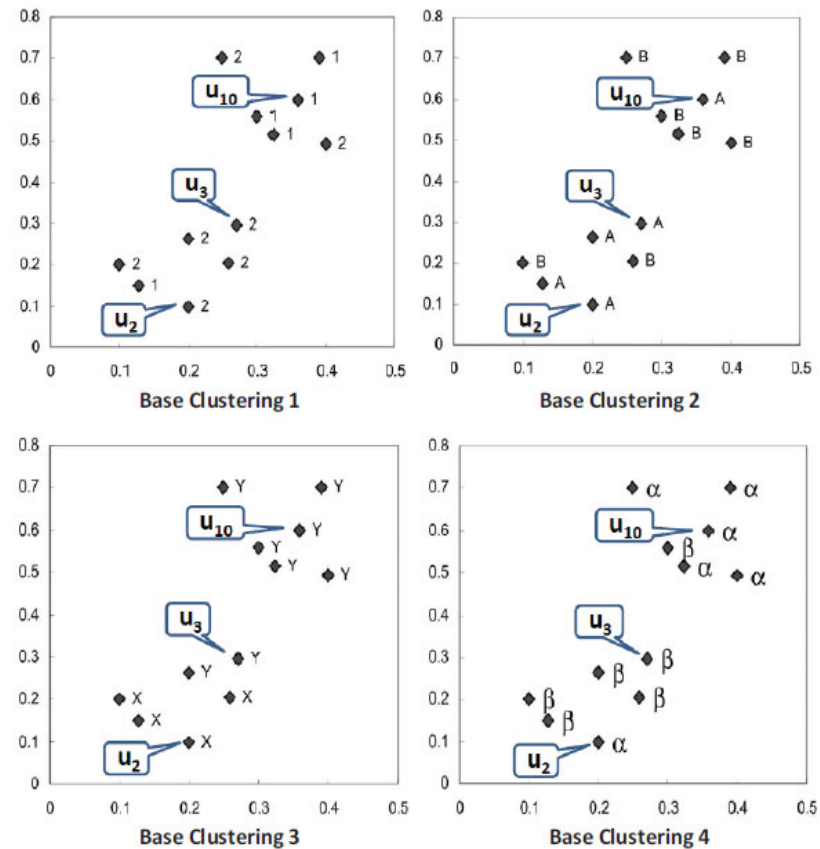
Possible cluster labels based on four base clusterings

- object u_2 : $\{2, A, X, \alpha\}$
- object u_3 : $\{2, A, Y, \beta\}$
- object u_{10} : $\{1, A, Y, \alpha\}$

same

distinct

By following traditional way,
we have $\text{Sim}(u_2, u_3) =$
 $\text{Sim}(u_2, u_{10}) = \text{Sim}(u_3, u_{10}) = 0.5$,
which is problematic.



Problems

- **The reason** is that the similarity defined here is too limited to reveal the complete hidden relationships among the data set from the initial results of base clustering.
- **A conventional way** is to randomly distribute them in either an identical cluster or different groups, which will inevitably affect the clustering performance.

Motivation

Identify some coupling relationships: between the base clusterings
and between the data objects

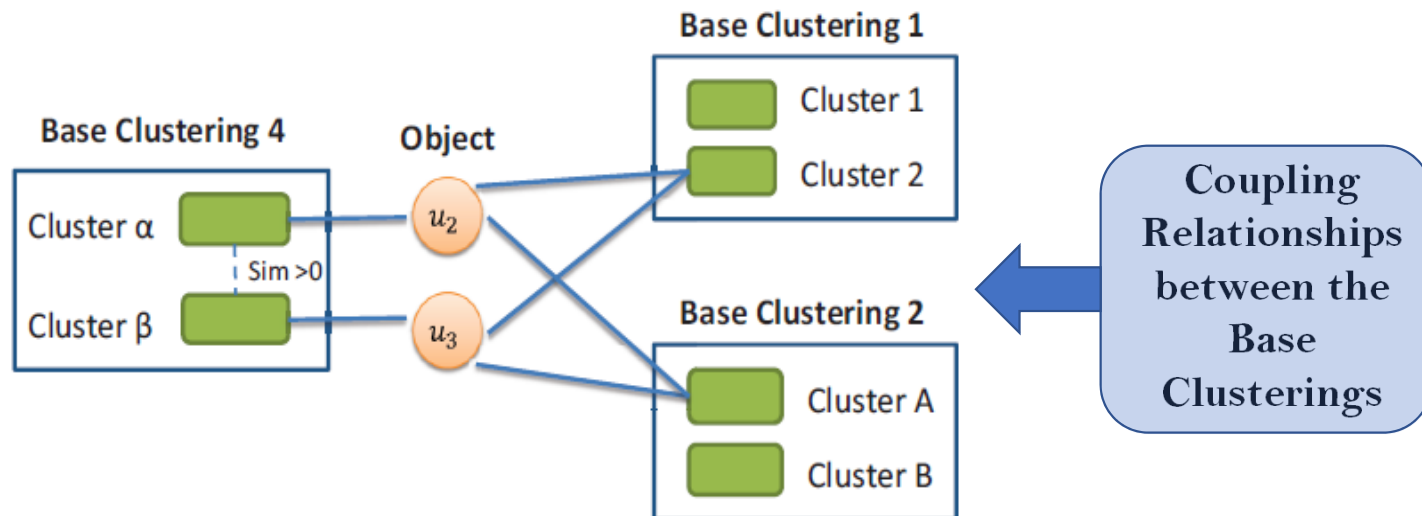


Fig. 2. A graphical representation of the coupled relationship between base clusterings, where each circle denotes an object, each rectangle represents an cluster, and an edge exists if an object belongs to a cluster.

Hierarchical Couplings

We then come up with three research questions in the following.

- ***Clustering Coupling:*** There is likely structural relationship between base clusterings since they are induced from the same data set. How to describe the coupling relationship between base clusterings?
- ***Object Coupling:*** There is context surrounding two objects which makes them dependent on each other. How to design the similarity or distance between objects to capture their relations with other data objects?
- ***Integrated Coupling:*** If there are interactions between both clusterings and objects, then how to integrate such couplings in clustering ensemble?

Framework of Coupled Clustering Ensembles

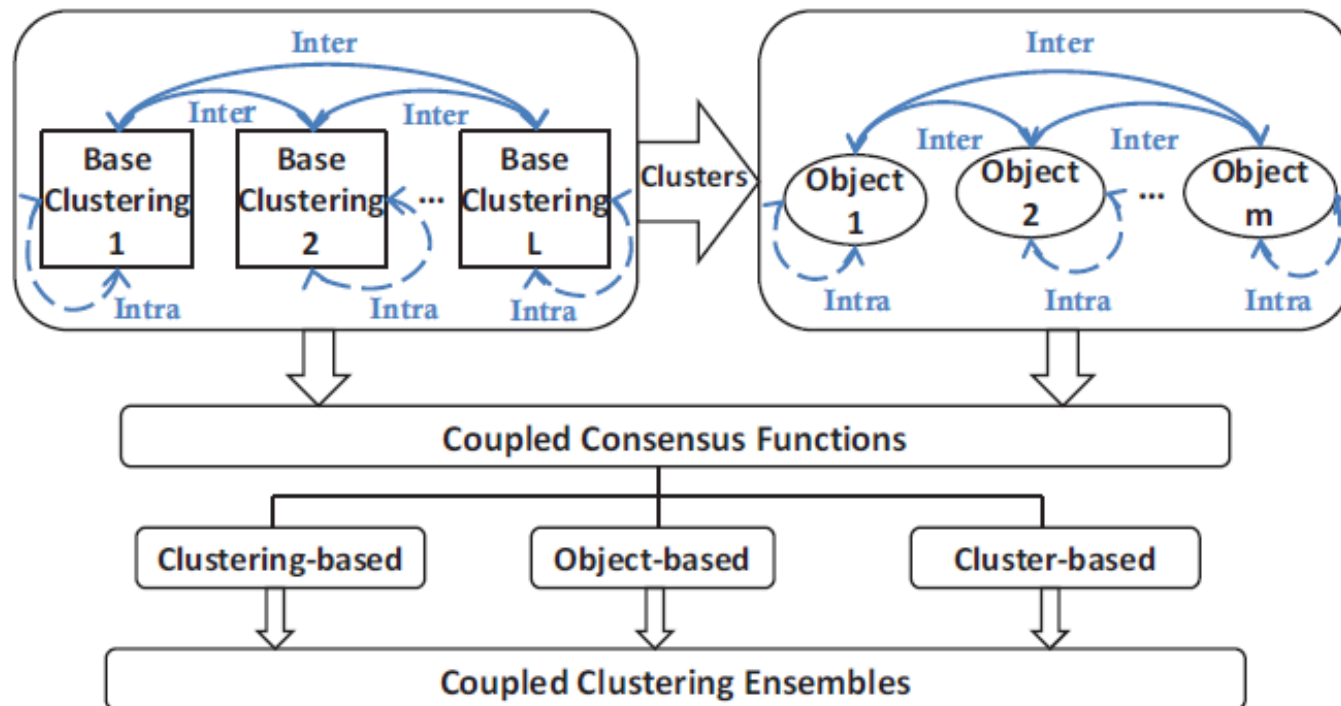
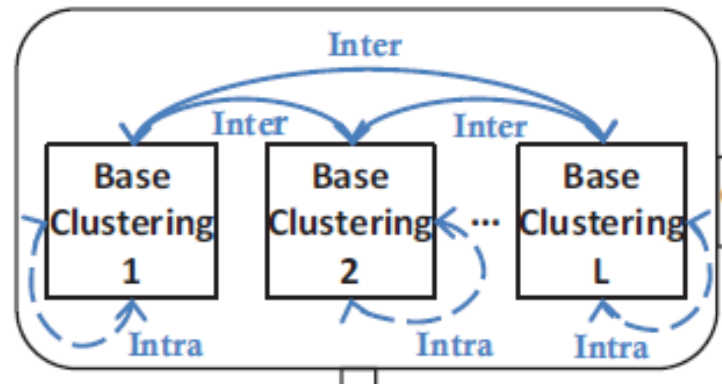


Fig. 3. A coupled framework of clustering ensembles (CCE), where $\leftarrow\cdots\rightarrow$ indicates the intra-coupling and \longleftrightarrow refers to the inter-coupling.

Clustering Couplings



Clustering Coupling: relationships within each base clustering and the interactions between distinct base clusterings are induced from the **coupled nominal similarity measure**

Intra-coupling of base clusterings: cluster label frequency distribution

Inter-coupling of base clusterings: cluster label co-occurrence dependency

Couplings in CCE

Coupling of Clusterings

- **Intra-coupling of base clusterings** indicates the involvement of cluster label occurrence frequency within one base clustering

Definition 5.1: (IaCSC) The **Intra-coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j is:

$$\delta_j^{IaC}(v_j^x, v_j^y) = \frac{|g_j(v_j^x)| \cdot |g_j(v_j^y)|}{|g_j(v_j^x)| + |g_j(v_j^y)| + |g_j(v_j^x)| \cdot |g_j(v_j^y)|}, \quad (V.1)$$

the set of objects whose cluster labels is v_j^y in base clustering bc_j

where $g_j(v_j^x)$ and $g_j(v_j^y)$ are the set information functions.

Greater similarity is assigned to labels with approximately equal frequencies. The larger these frequencies, the closer two labels.

Couplings in CCE

Coupling of Clusterings

- **Inter-coupling of base clusterings** means the interaction of other base clusterings with this base clustering

Definition 5.2: (IeRSC) The **Inter-coupled Relative Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j based on another base clustering bc_k is:

$$\delta_{j|k}(v_j^x, v_j^y | V_k) = \sum_{v_k \in \cap} \min\{P_{k|j}(v_k | v_j^x), P_{k|j}(v_k | v_j^y)\}, \quad (\text{V.2})$$

where $v_k \in \cap$ denotes $v_k \in \varphi_{j \rightarrow k}(v_j^x) \cap \varphi_{j \rightarrow k}(v_j^y)$, $\varphi_{j \rightarrow k}$ is the inter-information function, and $P_{k|j}$ is the information conditional probability formalized in Equation (III.1).

Couplings in CCE

Coupling of Clusterings

- **Inter-coupling of base clusterings** means the interaction of other base clusterings with this base clustering

Definition 5.3: (IeCSC) The **Inter-coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of base clustering bc_j is:

$$\delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k \neq j}) = \sum_{k=1, k \neq j}^L \lambda_k \delta_{j|k}(v_j^x, v_j^y | V_k), \quad (V.3)$$

where λ_k is the weight for base clustering bc_k , $\sum_{k=1, k \neq j}^L \lambda_k = 1$, $\lambda_k \in [0, 1]$, $V_k (k \neq j)$ is a cluster label set of base clustering bc_k different from bc_j to enable the inter-coupled interaction, and $\delta_{j|k}(v_j^x, v_j^y | V_k)$ is *IeRSC*.

Couplings in CCE

Coupling of Clusterings

IaCSC captures the **base clustering frequency distribution** by calculating occurrence times of cluster labels within one base clustering, and **IeCSC** characterizes the **base clustering dependency aggregation** by comparing co-occurrence of the cluster labels in objects among different base clusterings. Finally, there is an eligible way to **incorporate these two couplings together**, specifically:

Definition 5.4: (CCSC) The **Coupled Clustering Similarity for Clusters** between cluster labels v_j^x and v_j^y of clustering bc_j is:

how often the cluster label occurs

$$\delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L) = \delta_j^{IaC}(v_j^x, v_j^y) \cdot \delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k \neq j}),$$

the extent of the cluster difference

where δ_j^{IaC} and δ_j^{IeC} are *IaCSC* and *IeCSC*, respectively.

Couplings in CCE

Coupling of Clusterings

TABLE I
AN EXAMPLE OF BASE CLUSTERINGS

$U \backslash C$	bc_1	bc_2	bc_3	bc_4
u_1	2	B	X	β
u_2	2	A	X	α
u_3	2	A	Y	β
u_4	2	B	X	β
u_5	1	A	X	β
u_6	2	A	Y	β
u_7	2	B	Y	α
u_8	1	B	Y	α
u_9	1	B	Y	β
u_{10}	1	A	Y	α
u_{11}	2	B	Y	α
u_{12}	1	B	Y	α

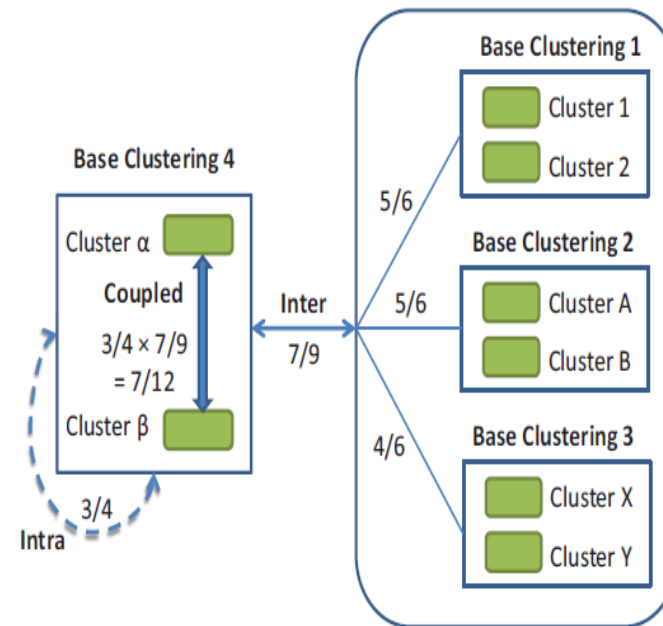
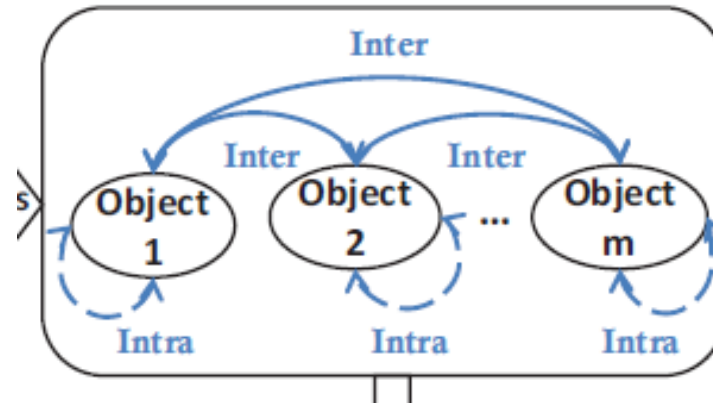


Fig. 4. An example of the coupled similarity for cluster labels α and β , where \longleftrightarrow indicates the intra-coupling and \longleftrightarrow refers to the inter-coupling, the value along each line is the corresponding similarity.

Object Couplings



Object Coupling: also focuses on the intra and inter-coupling and leads to a **more accurate similarity** ($\in [0, 1]$) between data objects.

Intra-coupling of objects: all the results of base clusterings for data objects

Inter-coupling of objects: the neighborhood relationship among data objects

Couplings in CCE

Coupling of Objects

In terms of the **intra-perspective**, the objects u_x coupled with u_y by involving the cluster labels of all the base clusterings for them.

Definition 5.5: (IaOSO) The **Intra-coupled Object Similarity for Objects** between objects u_x and u_y with respect to all the base clustering results of these two objects is:

$$\delta^{IaO}(u_x, u_y) = \frac{1}{L} \cdot \sum_{j=1}^L \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L), \quad (V.5)$$

where $\delta_j^C(v_j^x, v_j^y, \{V_k\}_{k=1}^L)$ refers to CCSC between cluster labels v_j^x and v_j^y of base clustering bc_j .

Couplings in CCE

Coupling of Objects

Further, we can embody the **inter-coupled** interaction between different objects by exploring the relationship between their neighborhood.

Definition 5.6: A pair of objects u_x and u_y are defined to be **neighbors** if the following holds:

$$\delta^{Sim}(u_x, u_y) \geq \theta, \quad (V.6)$$

where δ^{Sim} denotes any similarity measure for objects, $\theta \in [0, 1]$ is a given threshold.

The neighbor set of object u_x : $N_{u_x} = \{u_z | \delta^{Sim}(u_x, u_z) \geq \theta\}$

Couplings in CCE

Coupling of Objects

Intuitively, objects u_x and u_y more likely belong to the same cluster if they have a larger overlapping in their neighbor sets N_{u_x} and N_{u_y} . Accordingly, below we use the common neighbors to define the **inter-coupled similarity** for objects.

Definition 5.7: (IeOSO) The **Inter-coupled Object Similarity for Objects** between objects u_x and u_y in terms of other objects u_z is defined as the ratio of common neighbors of u_x and u_y upon all the objects in U .

$$\delta^{IeO}(u_x, u_y|U) = \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{Sim} \cap N_{u_y}^{Sim}\}|, \quad (V.8)$$

where $N_{u_x}^{Sim}$ and $N_{u_y}^{Sim}$ are the neighbor sets of objects u_x and u_y based on δ^{Sim} , respectively.

Couplings in CCE

Coupling of Objects

Finally, the **intra-coupled** and **inter-coupled** interactions could be considered together to induce the following **coupled similarity** for objects by exactly specializing the similarity measure δ^{Sim} in (V.7) to be IaOSO δ^{IaO} in Equation (V.5).

Definition 5.8: (CCOSO) The **Coupled Clustering and Object Similarity for Objects** between objects u_x and u_y is defined when δ^{Sim} is in particular regarded as δ^{IaO} . Specifically:

$$\delta^{CO}(u_x, u_y|U) = \frac{1}{m} \cdot |\{u_z \in U | u_z \in N_{u_x}^{IaO} \cap N_{u_y}^{IaO}\}|, \quad (V.9)$$

where sets of objects $N_{u_x}^{IaO} = \{u_z | \delta^{IaO}(u_x, u_z) \geq \theta\}$ and $N_{u_y}^{IaO} = \{u_z | \delta^{IaO}(u_y, u_z) \geq \theta\}$.

Couplings in CCE

Coupling of Objects

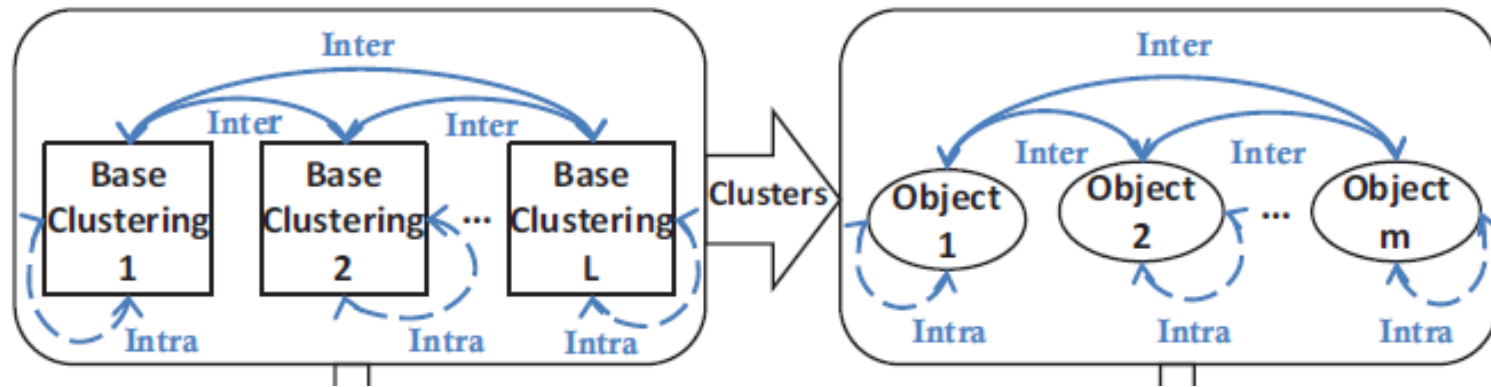
TABLE II
AN EXAMPLE OF NEIGHBORHOOD DOMAIN FOR OBJECT

Object	Neighborhood Domain
u_2	$\{u_1, u_3, u_4, u_5, u_6, u_7, u_8, u_{10}, u_{11}, u_{12}\}$
u_3	$\{u_1, u_2, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}\}$
u_{10}	$\{u_2, u_3, u_6, u_7, u_8, u_9, u_{11}, u_{12}\}$
Object Pair	Common Neighbors
u_2, u_3	$\{u_1, u_4, u_5, u_6, u_7, u_8, u_{10}, u_{11}, u_{12}\}$
u_2, u_{10}	$\{u_3, u_6, u_7, u_8, u_{11}, u_{12}\}$

$$\delta^{CO}(u_2, u_3|U) = 0.75 \text{ and } \delta^{CO}(u_2, u_{10}|U) = 0.5$$

It means that the similarity between objects u_2 and u_3 is larger than that between u_2 and u_{10}

Integrated Couplings



The data objects and base clusterings are associated through the corresponding clusters, i.e., the position of an object in a clustering is determined by which cluster the object belongs to

Integrated Coupling: treating each cluster label as an attribute value, and then defining the similarity between objects on the **similarity between cluster labels** over all base clusterings.

Clustering-based Coupling

The usual way:

V_j^x indicates the label of a cluster to which the object u_x belongs in the j th base clustering bc_j

$$BC_j^N(x, y) = \delta^N(v_j^x, v_j^y) = \begin{cases} 1 & \text{if } v_j^x = v_j^y \\ 0 & \text{otherwise} \end{cases}$$

Our proposed way CgC:

$$BC_j^C(x, y) = \delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^n)$$
$$S_{Cg}^C(bc_{j_1}, bc_{j_2}) = \sum_{1 \leq x, y \leq m} [BC_{j_1}(x, y) - BC_{j_2}(x, y)]^2$$

Coupled Clustering Similarity for Clusters:

$$\delta_j^C(v_j^x, v_j^y | \{V_k\}_{k=1}^L) = \delta_j^{IaC}(v_j^x, v_j^y) \cdot \delta_j^{IeC}(v_j^x, v_j^y | \{V_k\}_{k \neq j})$$

Experiments

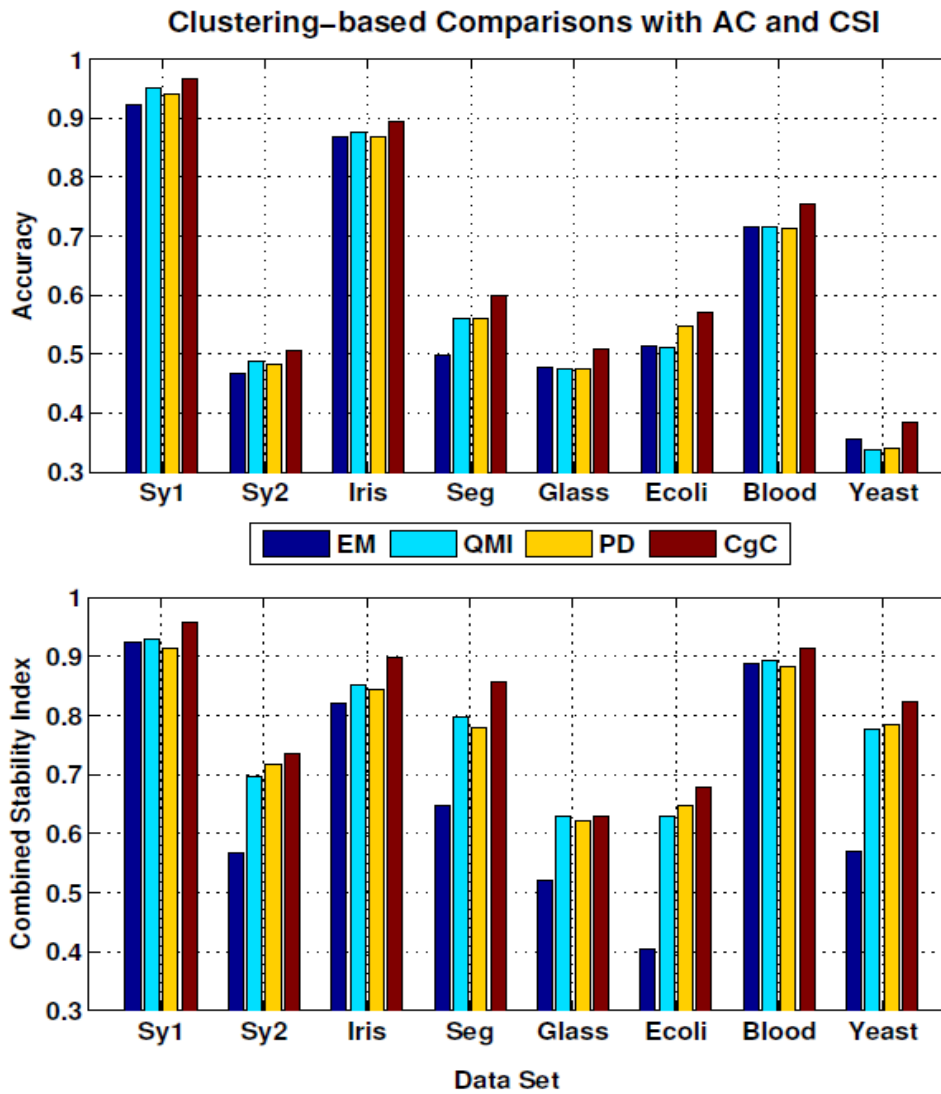


Fig. 5. Clustering-based comparisons.

Experiments

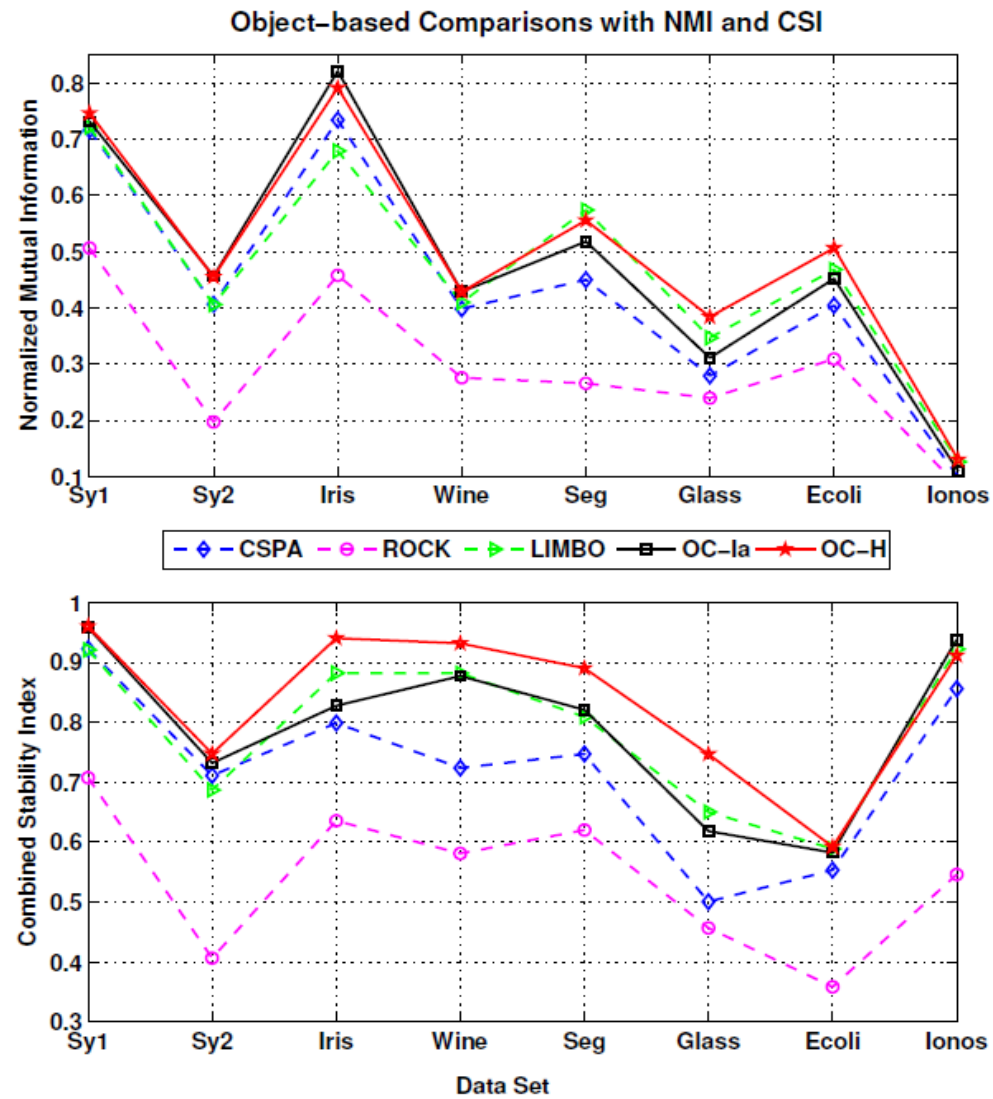


Fig. 6. Object-based comparisons.

TABLE V
CLUSTER-BASED COMPARISONS ON AC, NMI AND CSI

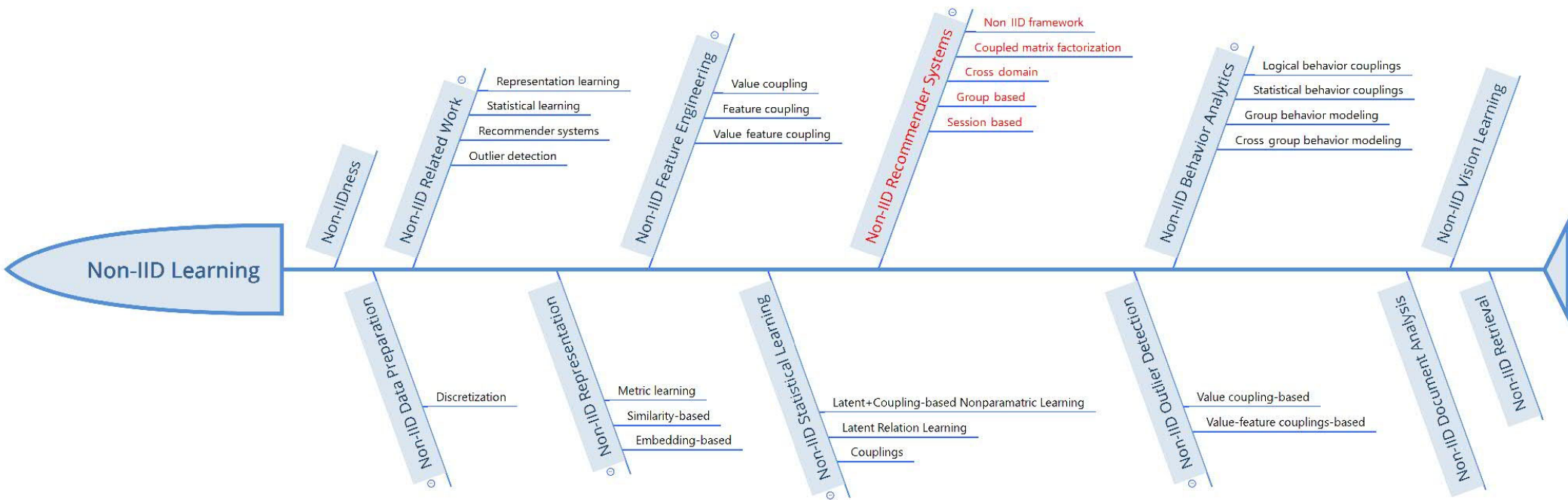
Data Set		Sy1	Sy2	Iris	Wine	Seg	Glass	Ecoli	Ionos	Blood	Vowel	Yeast	Avg
AC	<i>MCLA</i>	0.945	0.501	0.875	0.702	0.560	0.472	0.528	0.711	0.680	0.365	0.341	0.607
	<i>HBGF</i>	0.949	0.503	0.877	0.690	0.532	0.445	0.468	0.684	0.528	0.379	0.301	0.578
	<i>LB-P</i>	0.952	0.504	0.878	0.703	0.582	0.459	0.530	0.711	0.719	0.330	0.328	0.609
	<i>LB-S</i>	0.951	0.486	0.844	0.690	0.560	0.483	0.539	0.711	0.713	0.364	0.332	0.607
	<i>CrC-Ia</i>	0.954	0.513	0.893	0.731	0.579	0.482	0.539	0.721	0.713	0.394	0.379	0.627
	<i>CrC-C</i>	0.969	0.518	0.902	0.764	0.579	0.511	0.587	0.742	0.723	0.430	0.378	0.646
NMI	<i>MCLA</i>	0.725	0.406	0.744	0.429	0.526	0.318	0.510	0.129	0.015	0.411	0.223	0.403
	<i>HBGF</i>	0.710	0.389	0.706	0.355	0.486	0.316	0.444	0.109	0.007	0.414	0.206	0.377
	<i>LB-P</i>	0.723	0.406	0.745	0.429	0.548	0.318	0.511	0.130	0.016	0.420	0.221	0.406
	<i>LB-S</i>	0.724	0.363	0.687	0.412	0.531	0.335	0.502	0.130	0.015	0.394	0.210	0.391
	<i>CrC-Ia</i>	0.734	0.436	0.752	0.556	0.543	0.323	0.511	0.164	0.018	0.445	0.226	0.428
	<i>CrC-C</i>	0.764	0.456	0.753	0.580	0.540	0.337	0.539	0.171	0.019	0.477	0.228	0.442
CSI	<i>MCLA</i>	0.950	0.710	0.876	0.828	0.775	0.554	0.640	0.937	0.897	0.783	0.774	0.793
	<i>HBGF</i>	0.953	0.703	0.761	0.712	0.716	0.594	0.528	0.839	0.642	0.736	0.742	0.721
	<i>LB-P</i>	0.954	0.713	0.860	0.829	0.840	0.601	0.673	0.943	0.893	0.774	0.786	0.806
	<i>LB-S</i>	0.943	0.662	0.787	0.846	0.767	0.601	0.594	0.926	0.892	0.757	0.727	0.773
	<i>CrC-Ia</i>	0.967	0.736	0.892	0.868	0.878	0.621	0.649	0.955	0.897	0.808	0.817	0.826
	<i>CrC-C</i>	0.963	0.752	0.910	0.880	0.880	0.639	0.679	0.957	0.940	0.872	0.822	0.845

Conclusions

We draw the following three conclusions to address the research questions :

- **Base clusterings** are indeed coupled with each other, and the consideration of such couplings can result in better clustering quality
- The inclusion of **coupling between objects** further improves the clustering accuracy and stability
- The improvement level brought by the coupling of base clusterings is associated with the **accuracy of base clusterings**, while the improvement degree caused by the inter-coupling of objects is dependent on the **consistency of base clustering results**

Non-IID 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.

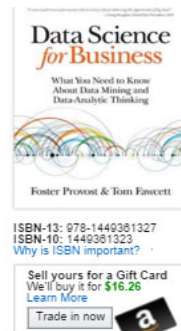
Longbing Cao, Philip Yu. [Non-IID Recommendation Theories and Systems](#). IEEE Intelligent Systems, 31(2), 81-84, 2016.

Challenges

Amazon

Recommendation problems:

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



ISBN-13: 978-1449361327
ISBN-10: 1449361323
[Why is ISBN important?](#)

Sell yours for a Gift Card
We'll buy it for **\$16.26**
[Learn More](#)

Trade in now

Have one to sell?

[Add to List](#) [Sell on Amazon](#)

Share

Rent \$15.20

Buy new **\$37.99**

List Price: ~~\$66.99~~ Save: \$2.00
43 New from \$23.22

In Stock. Ships from and sold by Amazon.com. Gift-wrap available.

Want it Friday, Dec. 18? Order within **8 hrs 37 mins** and choose Two-Day Shipping at checkout. [Details](#)

FREE Shipping

Qty: 1

Add to Cart

Turn on 1-Click ordering

Ship to: Select a shipping address:

More Buying Choices

43 New from \$23.22 26 Used from \$23.23

See All Buying Options

amazonstudent FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS [Learn more](#)

Written by renowned data science experts Foster Provost and Tom Fawcett, *Data Science for Business* introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining

[Read more](#)

Get Up to 80% Back
For Your Textbooks
[Learn More](#)

Frequently Bought Together



Total price: **\$81.20**
[Add all three to Cart](#)
[Add all three to List](#)

- ✓ This item: *Data Science for Business: What you need to know about data mining and data-analytic thinking* by Foster Provost Paperback **\$37.99**
- ✓ *Data Smart: Using Data Science to Transform Information into Insight* by John W. Foreman Paperback **\$27.48**
- ✓ *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* by Eric Siegel Hardcover **\$15.73**

Customers Who Bought This Item Also Bought



Data Smart: Using Data Science to Transform Information into Insight
John W. Foreman
★★★★★ 84
Paperback
\$27.48 [Prime](#)



Data Science from Scratch: First Principles with Python
Joel Grus
★★★★★ 43
#1 Best Seller in Computer Programming...
Paperback
\$33.99 [Prime](#)



Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die
Eric Siegel
★★★★★ 259
Hardcover
\$15.73 [Prime](#)



Storytelling with Data: A Data Visualization Guide for Business Professionals
Cole Nussbaumer...
★★★★★ 12
#1 Best Seller in Information Management
Paperback
\$11.34 [Prime](#)



Naked Statistics: Stripping the Dread from the Data
Charles Wheelan
★★★★★ 308
Paperback
\$11.34 [Prime](#)



Practical Data Science with R
Nina Zumel
★★★★★ 28
Paperback
\$40.42 [Prime](#)



Big Data: A Revolution That Will Transform...
Viktor...
★★★★★ 355
Paperback
\$8.96 [Prime](#)



Big Data: Principles and best practices of scalable...
Nathan Marz
★★★★★ 23
#1 Best Seller in User Generated Content
Paperback
\$36.58 [Prime](#)



Doing Data Science: Straight Talk from the Frontlines
Cathy O'Neil
★★★★★ 48
#1 Best Seller in Stochastic Modeling
Paperback
\$29.82 [Prime](#)



Data Analysis Using SQL and Excel
Gordon S. Linoff
★★★★★ 30
Paperback
\$31.59 [Prime](#)



Show Me the Numbers: Designing Tables and Visualizations to Highlight Information
Stephen Few
★★★★★ 38
#1 Best Seller in Graph Theory
Hardcover
\$26.52 [Prime](#)



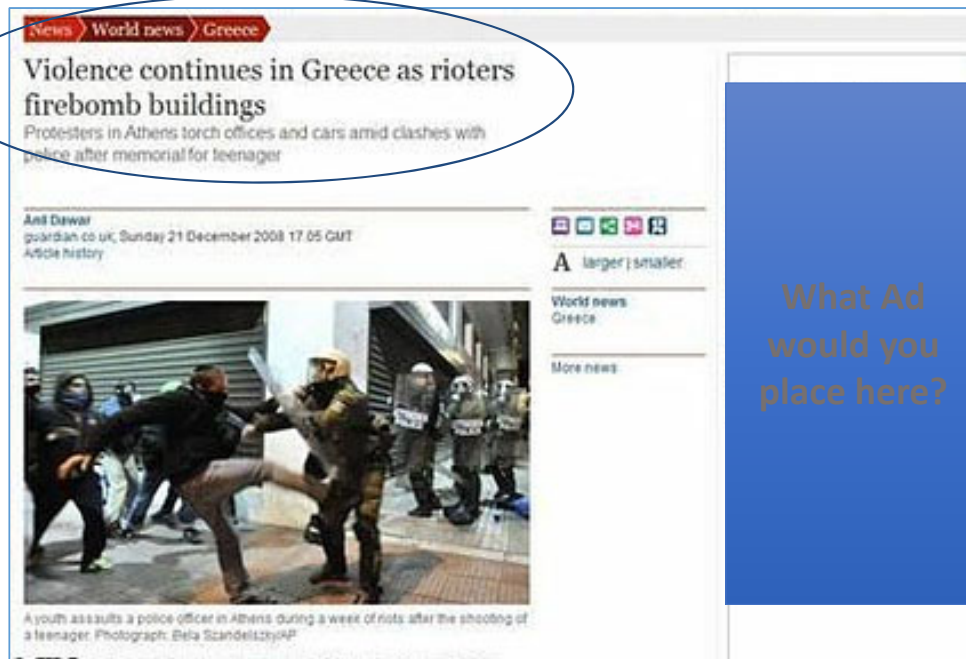
Analytics in a Big Data World: The Essential Guide...
Bart Baesens
★★★★★ 27
Hardcover
\$37.42 [Prime](#)



A PRACTITIONER'S GUIDE TO BUSINESS ANALYTICS
Randy Bartlett
★★★★★ 41
Hardcover
\$44.03 [Prime](#)

Big data challenges existing theories and systems

Irrelevant and
Damaging to Brand

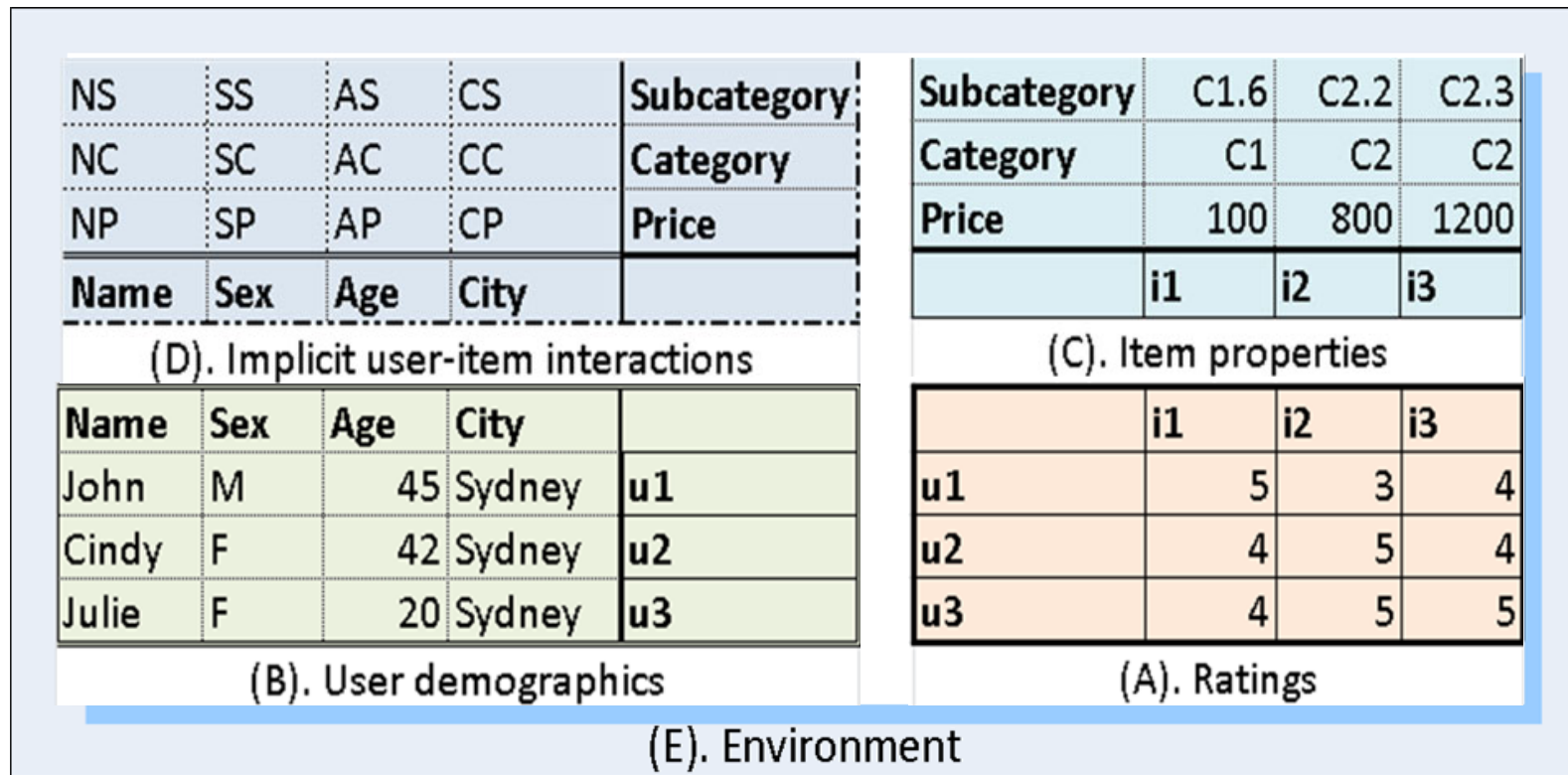


What Ad
would you
place here?

Why the prediction doesn't work?

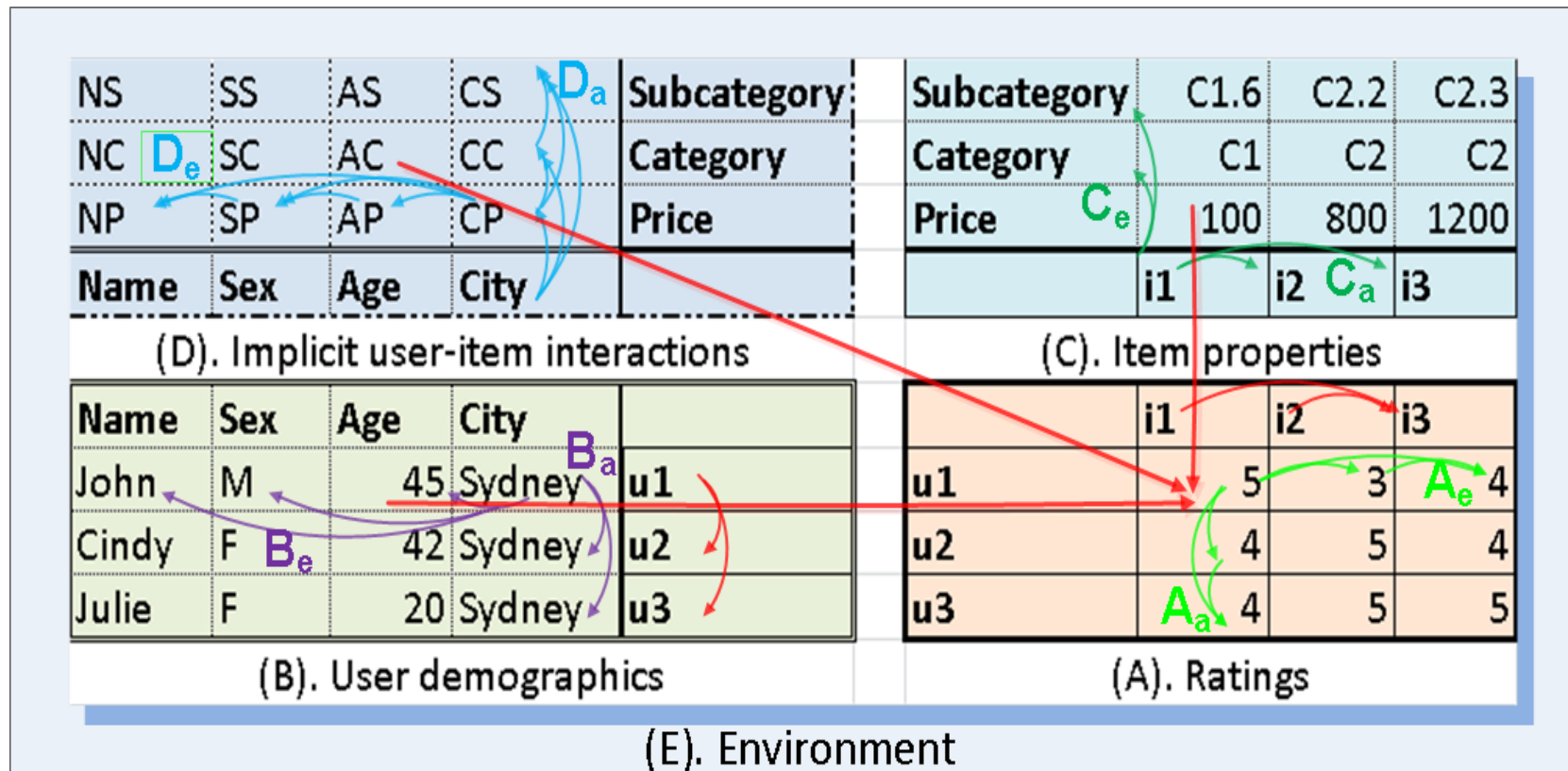
- There may be many reasons,
 - Content understanding
 - Understand the semantic hidden in contents
 - Analyze the relevance between news and ads from every possible aspect
 - Treat each piece of news differently
 - ...
- A fundamental assumption - IIDness
 - Weaken or overlook the data complexities
 - Relationships between objects, syntactically, semantically,
 - Heterogeneity between objects, sources, ...

A Systematic View of Recommendation

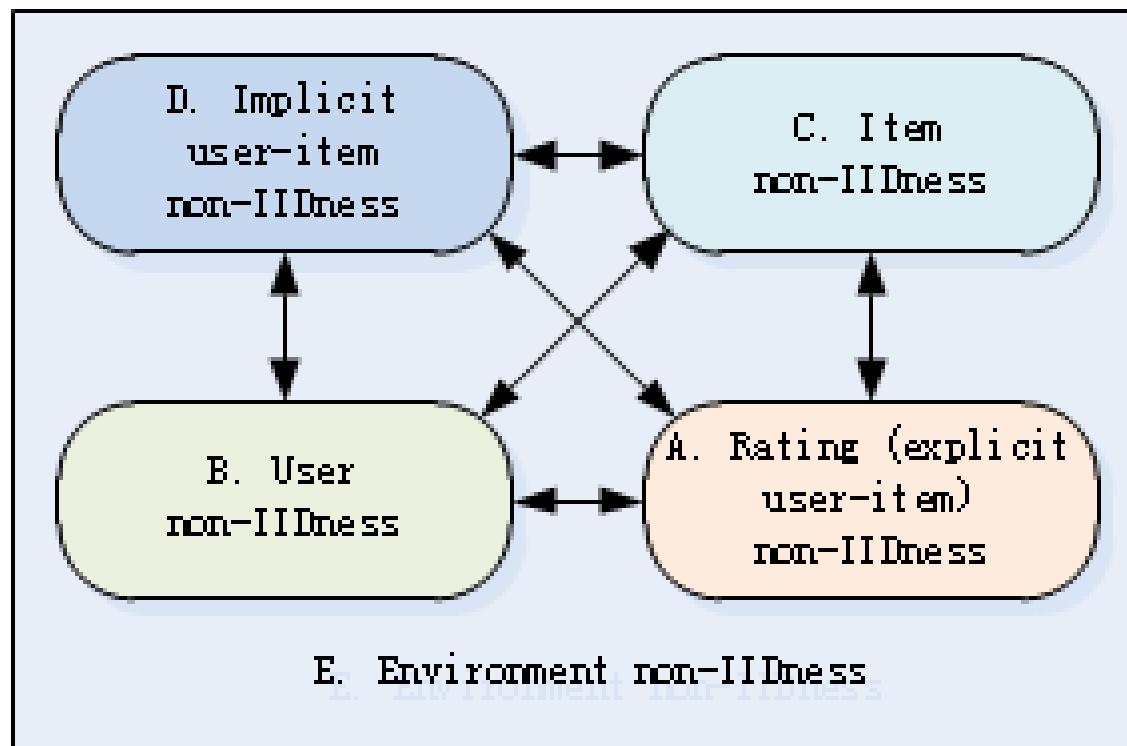


Longbing Cao. [Non-IID Recommender Systems: A Review and Framework of Recommendation Paradigm Shifting](#). Engineering, 2: 212-224, 2016.

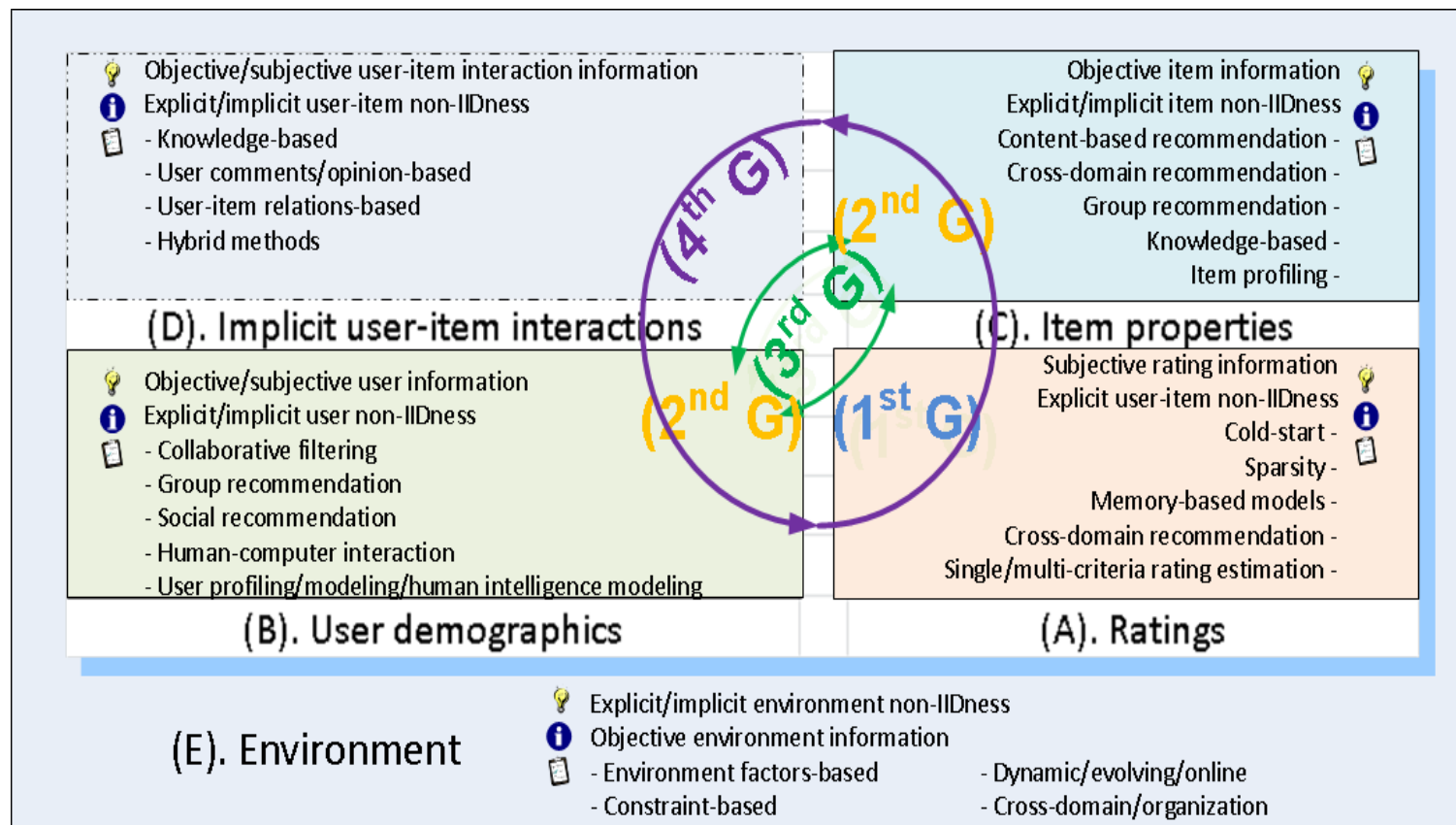
Non-IIDness in Recommendation



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:

$$R \approx \hat{R} = P^T Q$$
$$\hat{R}_{ij} = P_i^T \cdot Q_j$$

	v_1	v_2	...	v_m
u_1	1	2	?	3
u_2	2	?	?	4
\vdots				
u_n	4	1	?	?

R

	1	2	...	k
u_1
u_i
\vdots
u_n

P^T

	v_1	v_j	...	v_m
1
2
\vdots
k

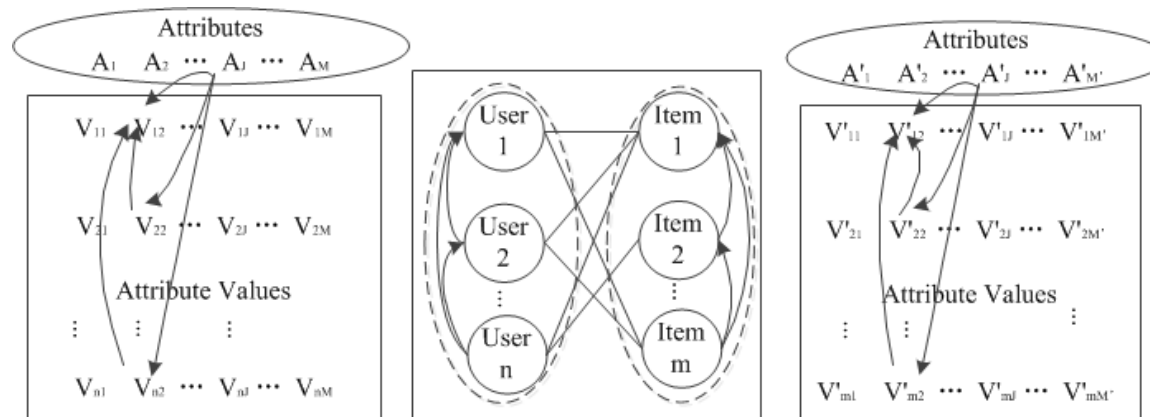
Q

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 (not independently and identically distributed) characteristics for items and users.

User/item Coupling Analysis

- Deep couplings within users and items contribute to the rating behavior.
 - Attribute values are coupled together and not independent,
 - Attributes are also coupled together 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 $S_U = \langle U, A, V, f \rangle$, the **Coupled User Similarity (CUS)** between two users u_i and u_j is defined as follows.

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

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 $S_O = \langle O, A', V', f' \rangle$, 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)

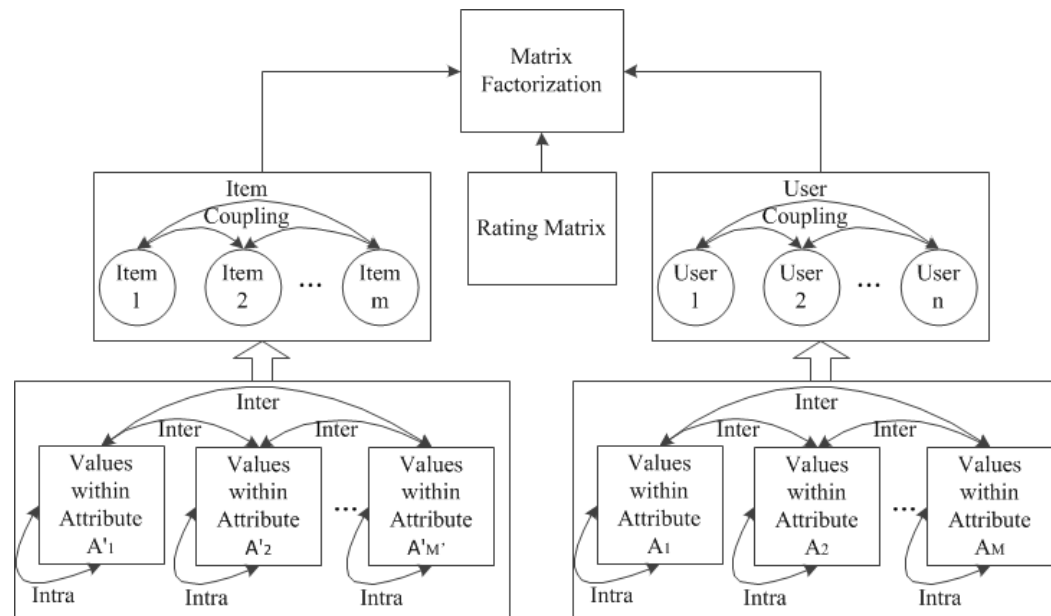
Matrix Factorization

- 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 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

$$L = \frac{1}{2} \sum_{(u,o_i) \in K} \left(R_{u,o_i} - \hat{R}_{u,o_i} \right)^2 + \frac{\lambda}{2} (\|Q_i\|^2 + \|P_u\|^2) + \frac{\alpha}{2} \sum_{all(u)} \left\| P_u - \sum_{v \in \mathbb{N}(u)} CUS(u,v) P_v \right\|^2 + \frac{\beta}{2} \sum_{all(o_i)} \left\| Q_i - \sum_{o_j \in \mathbb{N}(o_i)} CIS(o_i,o_j) Q_j \right\|^2$$

- Optimization

$$\frac{\partial L}{\partial P_u} = \sum_{o_i} I_{u,o_i} (r_m + P_u Q_i^T - R_{u,o_i}) Q_i + \lambda P_u + \alpha (P_u - \sum_{v \in \mathbb{N}(u)} CUS(u,v) P_v) - \alpha \sum_{v: u \in \mathbb{N}(v)} CUS(u,v) (P_v - \sum_{w \in \mathbb{N}(v)} CUS(v,w) P_w)$$

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

Baselines

- **PMF** is a probabilistic matrix factorization approach;
- **RSVD**: Singular value decomposition is a factorization method to decompose the rating matrix;
- **ISMF** is an unified model which incorporates implicit social relationships between users and between items computed by Pearson similarity.
- User-based CF (**UBCF**) computes users' similarity by Pearson Correlation on the rating matrix
- Item-based CF (**IBCF**) considers items' similarity by Pearson Correlation on the rating matrix
- Hybrid models **PSMF**, **CSMF** and **JSMF** respectively augment MF with Pearson Correlation Coefficient, Cosine and Jaccard similarity measures to compute the relationships between users and between items based on their attributes.

Data and Evaluation Metrics

- Movielens 1M:
 - 1,000,209 anonymous ratings; 3,900 movies; 6,040 users;
 - User information: “gender”, “age”, “occupation” and “zipcode”
 - Movie information: “genre” attribute.
- Book-Crossing
 - 278,858 users, 1,149,780 ratings on 271,379 books;
 - User information: “gender” and “age”
 - Book information: “book-author”, “year of publication” and “publisher”
- Evaluation Metrics

$$RMSE = \sqrt{\frac{\sum_{(u,i) \in R_{test}} (r_{u,i} - \hat{r}_{u,i})^2}{|R_{test}|}}$$

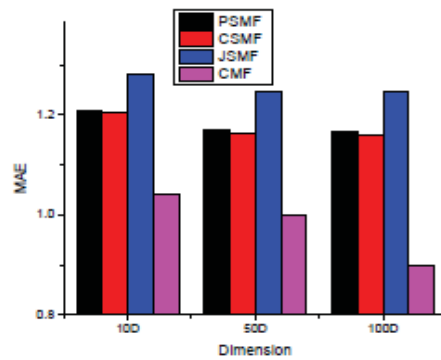
$$MAE = \frac{\sum_{(u,i) \in R_{test}} |r_{u,i} - \hat{r}_{u,i}|}{|R_{test}|}$$

Compared to MF and CF

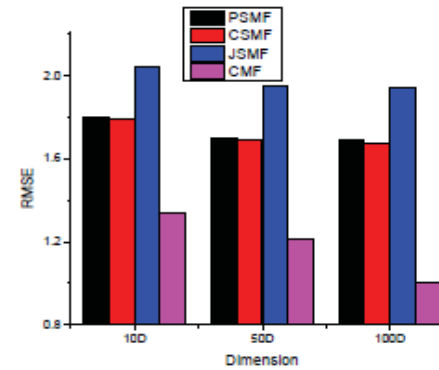
Data Set	Dim	Metrics	PMF (Improve)	ISMF (Improve)	RSVD (Improve)	CMF
Movielens	100D	MAE	1.1787(28.09%)	1.1125 (21.47%)	1.1076 (20.98%)	0.8978
		RMSE	1.7111 (71.07%)	1.5918 (59.14%)	1.5834 (58.30%)	1.0004
	50D	MAE	1.1852 (18.43%)	1.1188 (11.79%)	1.1088 (10.79%)	1.0009
		RMSE	1.8051 (58.98%)	1.6103 (39.50%)	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

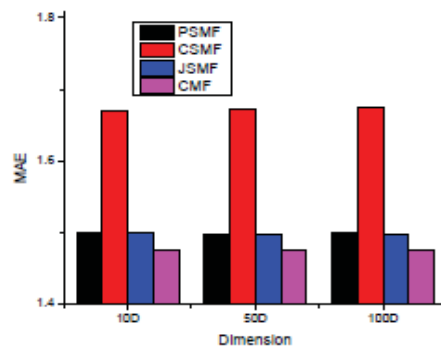
Compared to Hybrid Methods



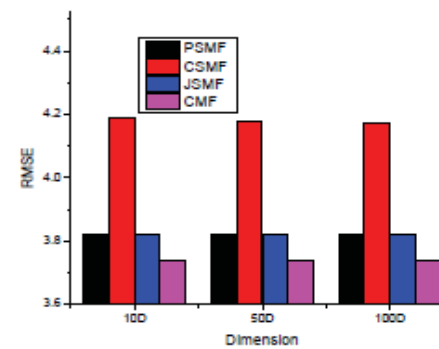
(a) MAE on Movielens



(b) RMSE on Movielens



(c) MAE on Bookcrossing



(d) RMSE on Bookcrossing

Summary of CMF

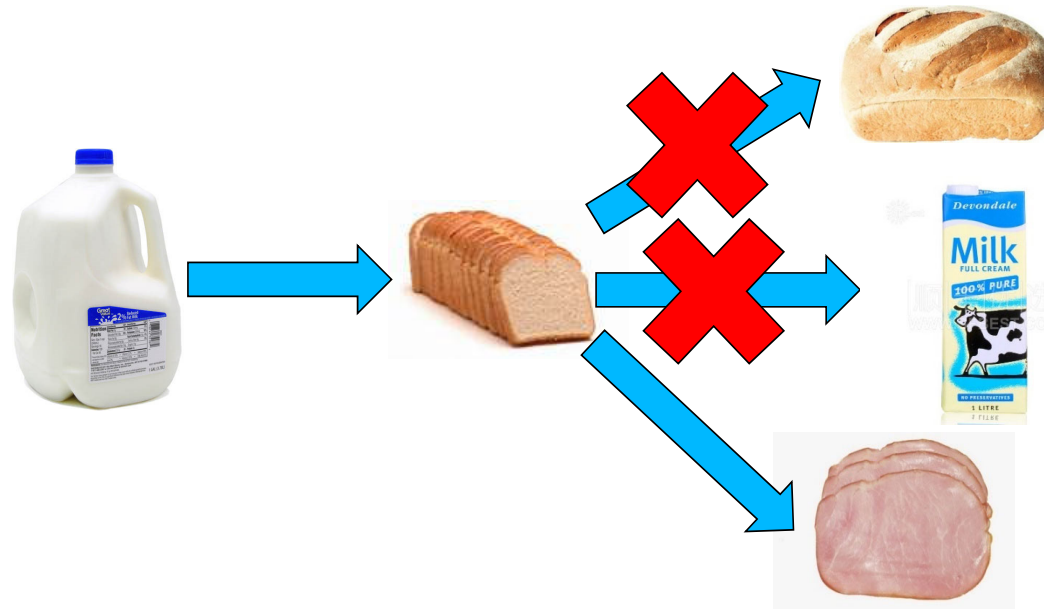
- Contributions
 - Applied a NonIID-based method to capture the couplings between users and items, based on their objective attribute information;
 - Integrated user coupling, item coupling and users' subjective rating preferences into matrix factorization learning model;
 - Evaluated the effectiveness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

Session-based Recommender Systems

Liang Hu, Longbing Cao, Shoujin Wang, Guandong Xu, Jian Cao, Zhiping Gu.
Diversifying Personalized Recommendation with User-session Context. In *IJCAI*.
2017

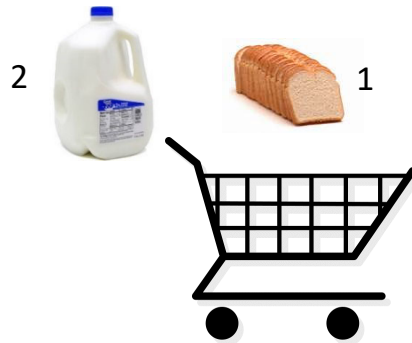
Deficiency of Current Recommender Systems

- Items are often **repeatedly recommended**.
- Users prefer **more diversified options** than those they have had.
 - It is unlikely that a consumer will purchase another a loaf of bread if they have purchased one, whereas butter or ham may be a more appealing recommendation.



Modeling Session

- Generally, choices are non-iid, which depend on previous choices in a session.
- A system makes more sensible and relevant recommendations if the **session context** was taken into consideration.
- The choices of items in a session may **not follow a rigidly ordered sequence**
 - For example, the order in which toast, milk and ham are put into a shopping cart makes no difference to the transaction.



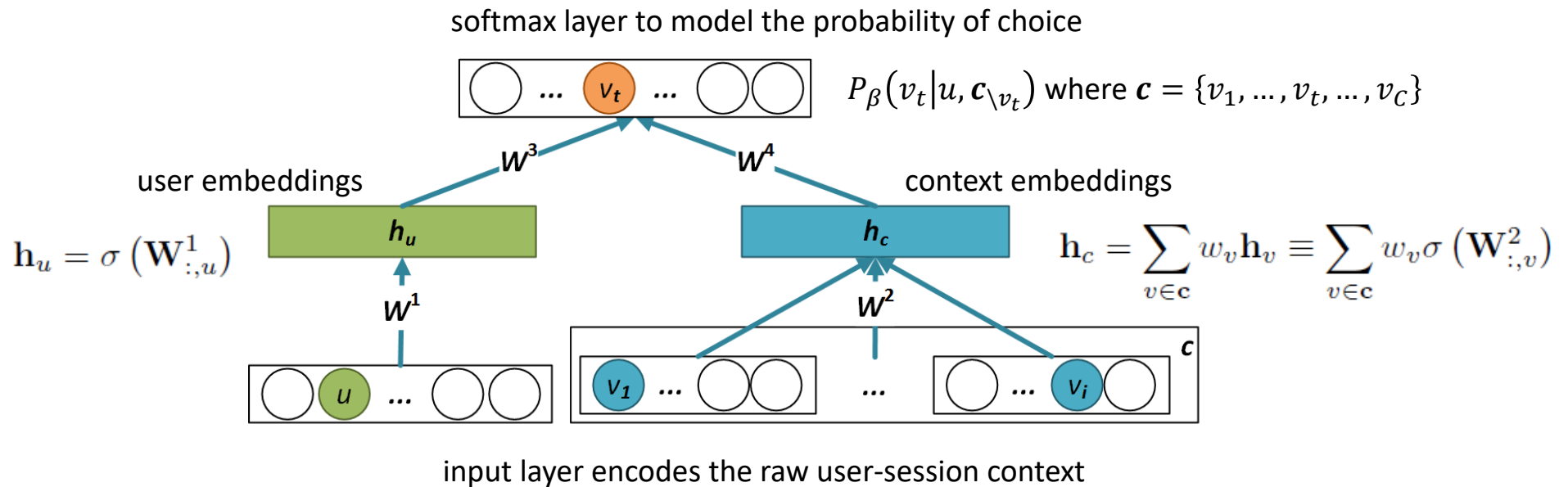
Inspiration by Language Model

- Language model is the probability distribution over sequences of words in natural language processing (NLP).
- $P(w_t | \mathbf{c})$ where $\mathbf{c} = \{w_1, \dots, w_K\}$ is context and $w_t \in V$
- If we think of **words as items**, predicting a relevant word based on context is equivalent to recommending a relevant item according to the current session.
- Both the number of items in RS and the size of vocabulary in language modeling are **large, usually $> 10^5$**

Wide-in-wide-out Shallow Networks

- SWIWO Architecture

- Three-layer shallow wide-in-wide-out networks



Maximum Log-likelihood Estimation

- Given session context \mathbf{c} and target item v_c , if we have N data samples:

$$L_{\Theta} = \sum_d \log P_{\Theta}(v_c | u_c, \mathbf{c}) = \sum_d S_{v_c}(u_c, \mathbf{c}) - \log \mathbf{Z}$$

where $d = \langle \mathbf{c}_u, v_c \rangle$ denotes one user session data, $\mathbf{c}_u = \langle u_c, \mathbf{c} \rangle$

$$S_{v_t}(u, \mathbf{c}) = \mathbf{W}_{t,:}^3 \mathbf{h}_u + \mathbf{W}_{t,:}^4 \mathbf{h}_c$$

- The challenge is the large size of item to compute normalizing constant
 - $\mathbf{Z} = \sum_V e^{S_v(\mathbf{c}_i, u)}$, normally $|\mathbf{V}| > 10^5$
 - For each data sample, it needs to compute $Z = \sum_V e^{S_v(\mathbf{c}_i, u)}$.
 - The total computation complexity $N|\mathbf{V}| > 10^{10}$ for each iteration, if $N > 10^5$

Softmax Approximation

- Noise-contrastive estimation (NCE)
 - Given a noise distribution $Q(w)$
 - Draw K noise samples $\{\tilde{w}_1, \dots, \tilde{w}_K\} \sim Q(w)$

- The probability comes from data distribution is

$$P_{\beta}(y = 1|w, c) = \frac{P_{\beta}(w|\mathbf{c})}{P_{\beta}(w|\mathbf{c}) + KQ(w)}$$
$$Q(w) = 1 - P_{\beta}(y = 1|w, c) = \frac{KQ(w)}{P_{\beta}(w|\mathbf{c}) + KQ(w)}$$

- Log-likelihood (LL)

$$\log P_{\beta}(y = 1|w, c) + \sum_{\tilde{w}_k} \log[1 - P_{\beta}(y = 1|w, c)]$$

Experiments

- IJCAI-15 Dataset
 - This real-world dataset was collected from Tmall.com which is the largest online B2C platform in China, and it contains anonymized users' shopping logs for the six months before and on the “Double 11” day (November 11th).

Training and Testing Data

- From the six-month shopping logs, we randomly held out 20% of the sessions from the last 30 days for testing, and the remaining data are used for training.
- We constructed two testing sets: ***LAST*** and ***LOO (Leave one out)***.

Statistic of IJCAI-15 dataset for evaluation
#users: 50K
#items: 52K
avg. session length: 2.99
#training sessions: & 0.20M
#training examples: & 0.59M
#testing cases (<i>LAST</i>): 4.5K
#testing cases (<i>LOO</i>): 11.9K

Comparison Methods

- **POP**: This recommender simply ranks items for recommendation according to occurrence frequency.
- **FPMC**: This recommender is a combination of MF and first-order MC, which uses personalized MC for sequential prediction.
- **PRME**: This recommender learns personalized transition probability in a MC model by applying a pairwise embedding metric method to handle data sparsity.
- **GRU4Rec**: This recommender is a deep RNN which consists of GRU units.
- **SWIWO**: This is the full model proposed in our paper.
- **SWIWO-I**: This a sub-model of SWIWO which only models item-session contexts without considering users.

Accuracy Evaluation

- The result of $REC@10$, $REC@20$ and MRR over the testing sets Last and LOO.

<i>LAST</i>			
Model	REC@10	REC@20	MRR
<i>POP</i>	0.0185	0.0317	0.0104
<i>FPMC</i>	0.0023	0.0068	0.0021
<i>PRME</i>	0.0670	0.0821	0.0363
<i>GRU4Rec</i>	0.2283	0.2464	0.1586
<i>SWIWO-I</i>	0.3223	0.3797	0.1918
<i>SWIWO</i>	0.3131	0.3689	0.1896
<i>LOO</i>			
Model	REC@10	REC@20	MRR
<i>POP</i>	0.0234	0.0420	0.0123
<i>FPMC</i>	0.0064	0.0117	0.0044
<i>PRME</i>	0.0757	0.0976	0.0431
<i>GRU4Rec</i>	0.2242	0.2425	0.1574
<i>SWIWO-I</i>	0.3177	0.3810	0.1903
<i>SWIWO</i>	0.3082	0.3703	0.1885

Diversity Evaluation

- We aim to diversify recommendation with session context.
- Now, let's consider the following metrics.

- **DIV@K**: This diversity measures the mean non-overlap ratio between each pair of recommendations $\langle \mathbf{R}_i, \mathbf{R}_j \rangle$ over all N top- K recommendations (note that the number of all possible pairs is $N(N - 1)/2$).

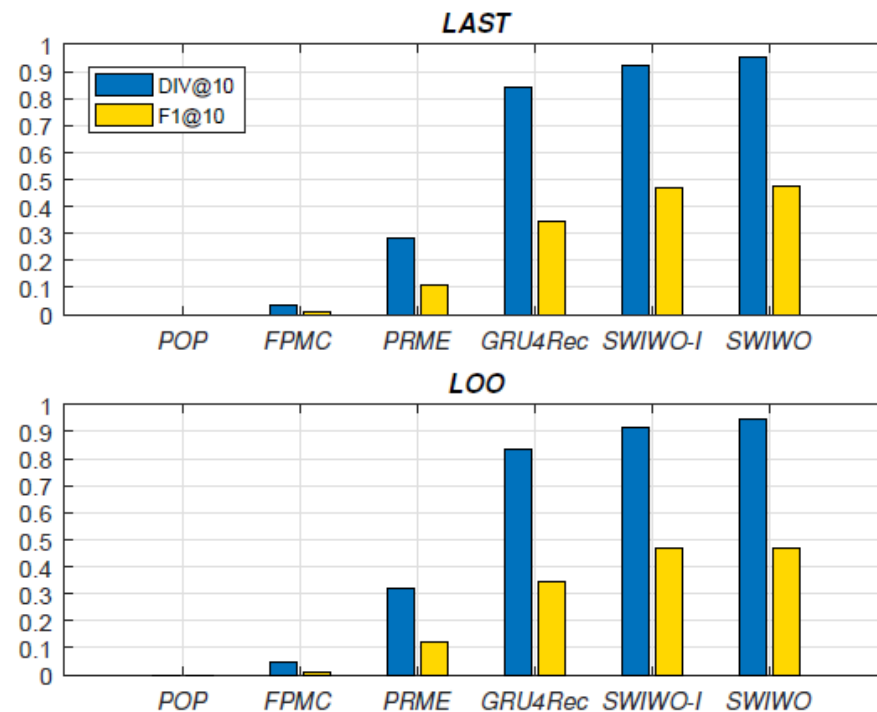
$$DIV@K = \frac{2}{N(N - 1)} \sum_{i \neq j} \left(1 - \frac{|\mathbf{R}_i \cap \mathbf{R}_j|}{|\mathbf{R}_i \cup \mathbf{R}_j|} \right)$$

- **F1@K**: The traditional F1 score is the harmonic mean of recall and precision. Here, we replace precision with diversity to jointly consider accuracy and diversity.

$$F1@K = \frac{2(REC@K \times DIV@K)}{REC@K + DIV@K}$$

Diversity Evaluation

- SWIWO considers the whole session context so they more easily provide diverse recommendation results.

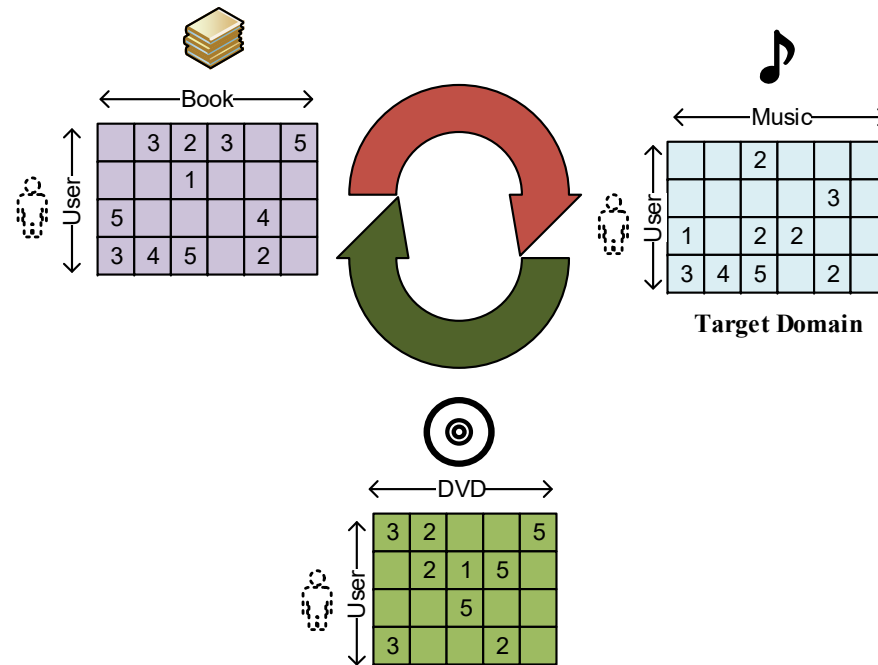


Cross-domain Recommender Systems

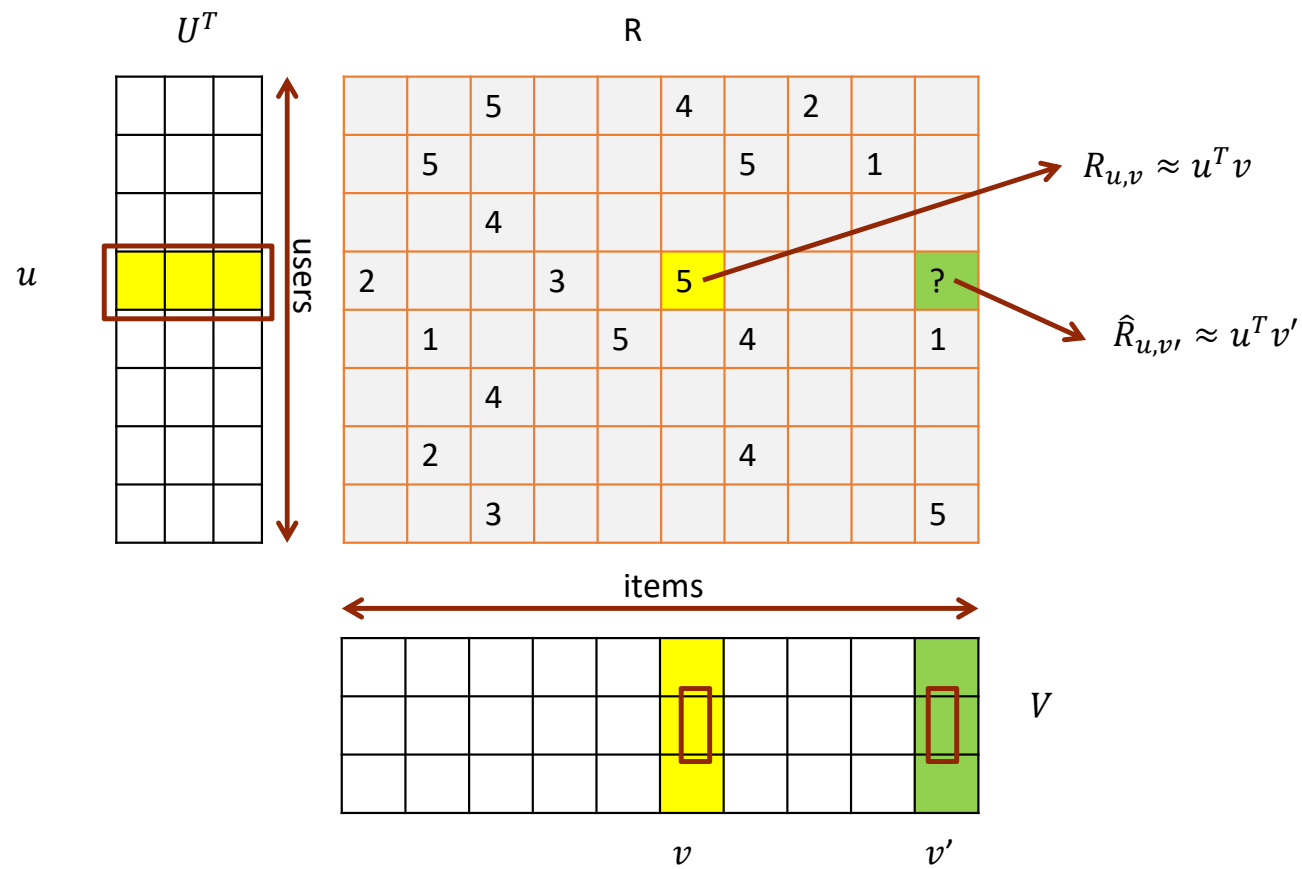
Hu, L., Cao, L., Cao, J., Gu, Z., Xu, G., & Yang, D. (2016). Learning Informative Priors from Heterogeneous Domains to Improve Recommendation in Cold-Start User Domains. *ACM Transactions on Information Systems (TOIS)*, 35(2), 13.

Cross-Domain Collaborative Filtering

- Leverage information from multiple related domains
 - The basic idea is based on the assumption of the existence of *multiple related domains* and the user preference from each domain is not independent



Matrix Factorization



MF for CDCF

- Concatenating the rating matrices for all domains

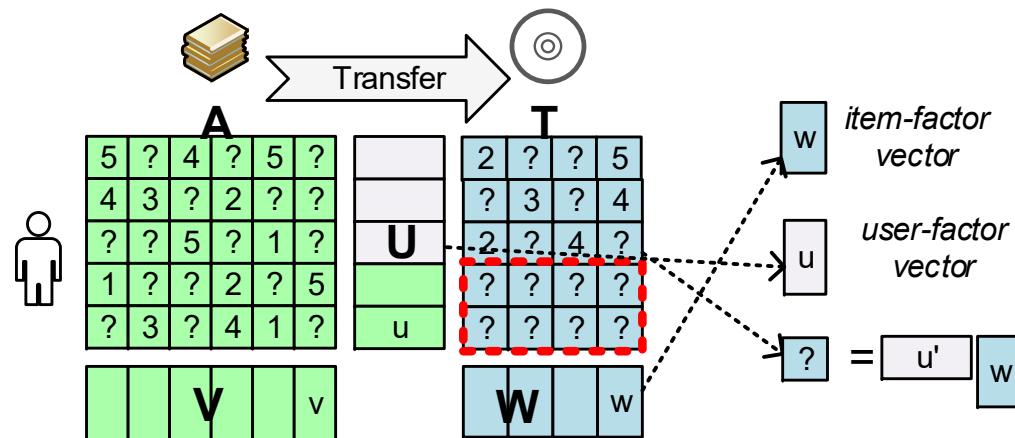
		3	2	3		5	3	3	2			5					5
									2	1	5				2		5
1			2				1						1		2	2	
3	4	5		2		5	4	5		2		3	4	5		2	
← Book →							← DVD →					← Music →					

Disadvantages

1. Each domain may be quite heterogeneous
 - E.g. the factor of *color* has big impact on the user preference in the domain of *cloth*
 - but hardly has impact on the user preference in domain of *book*
2. Above methods using the single domain model implicitly assume the homogeneity of items.
 - Obviously, such assumption may decrease the accuracy of prediction due to the *heterogeneities* of different domains.

MF-based Transfer Learning

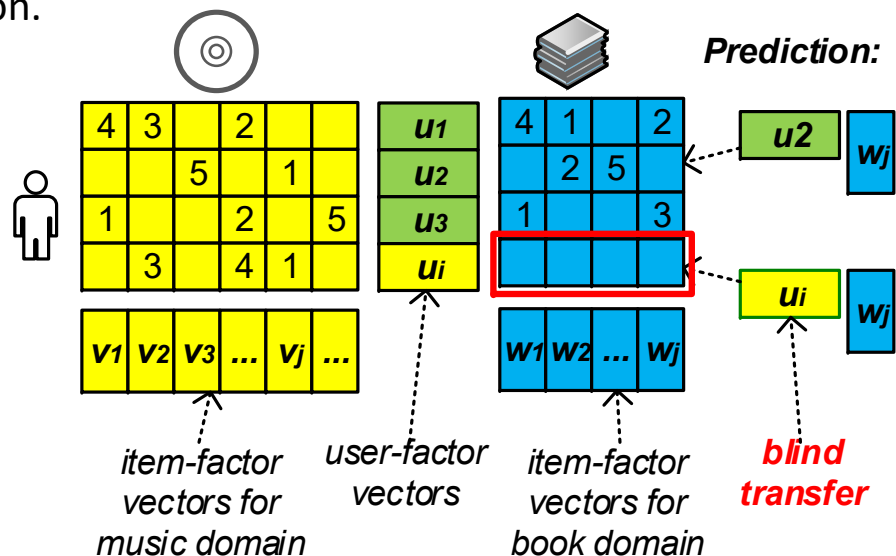
- Transfer the knowledge learned from the auxiliary domain to the target domain [Pan, *et al.* 2010] [Singh and Gordon, 2008].
- Assume dense user data in the auxiliary domain
- The user-factor vectors are *co-determined* by the feedback in auxiliary and target domains



Deficiency

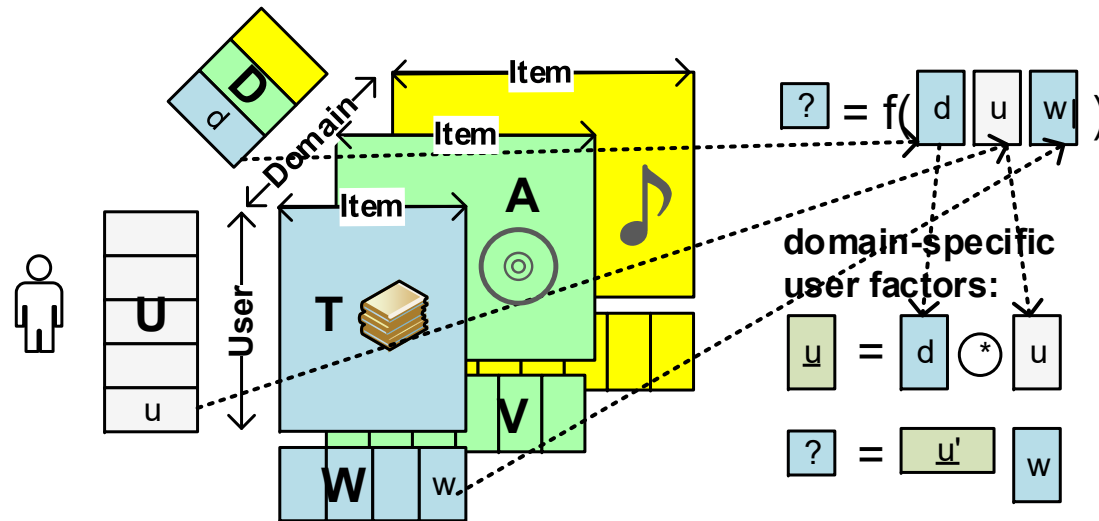
- Blind Transfer

- If no data is available for a user in the target domain (marked with a red box), the user-factor vector u_i is simply determined by the data in the **auxiliary domain**.
- If u_i is transferred to the target domain and interacts with heterogeneous item factors, it may yield a poor prediction.



Modeling Domain Heterogeneity

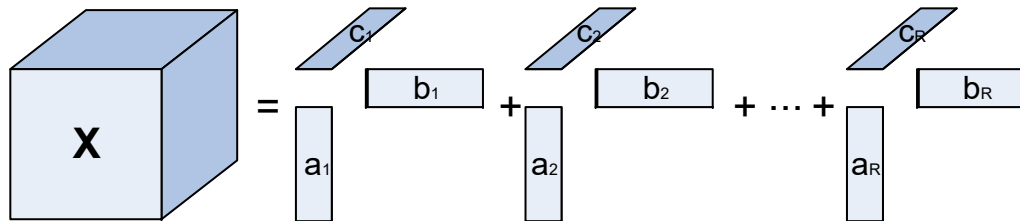
- Jointly leveraging the complementary data from multiple domains
- Domain factor** is an essential element in cross “domain” problem to model domain heterogeneity
- Triadic relation **user-item-domain** to reveal the domain-specific user preference



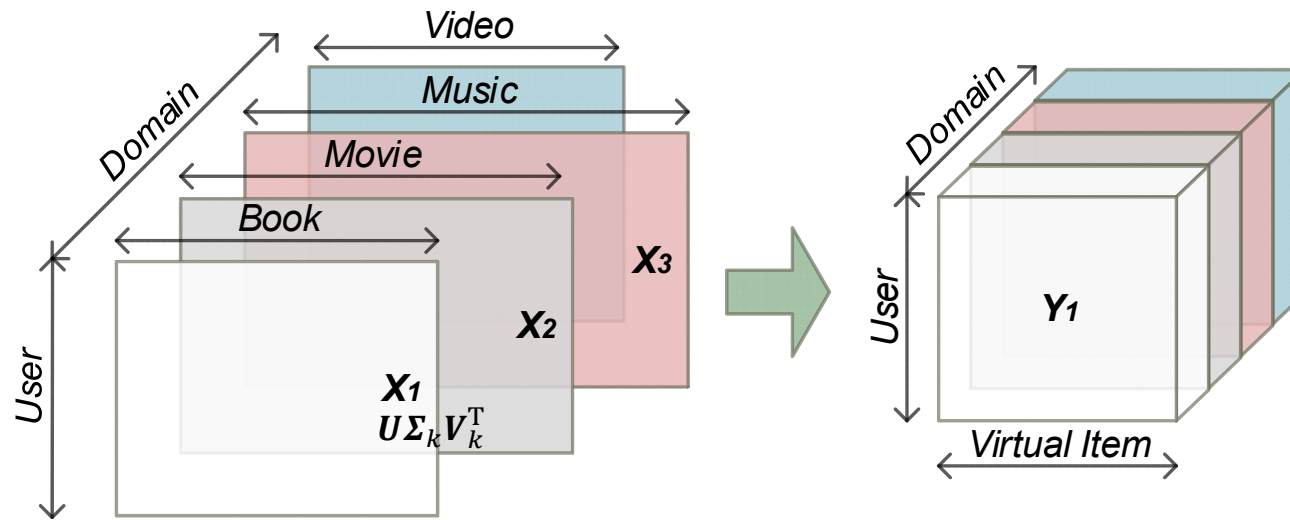
Canonical Decomposition/Parallel Factor Analysis

- Decompose a tensor into a sum of rank-one components
 - E.g. 3D Tensor:

$$\mathcal{X} = \llbracket \mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket = \sum_{r=1}^R \mathbf{A}_{:,r} \circ \mathbf{B}_{:,r} \circ \mathbf{C}_{:,r}$$



Irregular Tensor Factorization



- Sum loss over all domains:

$$\underset{\mathbf{U}, \mathbf{V}, \mathbf{C}}{\operatorname{argmin}} \frac{1}{2} \sum_{k=1}^K \|\mathbf{W}_k \odot (\mathbf{X}_k - \mathbf{U} \Sigma_k \mathbf{V}_k^T)\|_F^2 + \frac{\lambda_U}{2} \|\mathbf{U}\|^2 + \frac{\lambda_V}{2} \|\mathbf{V}\|^2 + \frac{\lambda_C}{2} \|\mathbf{C}\|^2$$

- With orthonormal constraints, we can obtain equivalent loss:

$$\underset{\mathbf{U}, \mathbf{V}, \mathbf{C}}{\operatorname{argmin}} \frac{1}{2} \left[\underbrace{(\|\mathbf{y} - [\mathbf{U}, \mathbf{V}, \mathbf{C}]\|^2 + \lambda_U \|\mathbf{U}\|_F^2 + \lambda_V \|\mathbf{V}\|_F^2 + \lambda_C \|\mathbf{C}\|_F^2)}_{1: \text{Regularized TF Model}} + \underbrace{\sum_k \|\hat{\mathbf{X}}_k \odot \mathbf{H}_k\|_F^2}_{2: \text{Loss Compensation}} \right]$$

Weight Matrix Configuration

- *Rating Data*

- $$w_{k,i,j} = \begin{cases} 1 & (k,i,j) \text{ is an observation} \\ a & (k,i,j) \text{ is a noisy example} \\ 0 & \text{else} \end{cases}$$

- *Noisy data act as regularization*

One-class Data

- One-class feedback
 - E.g. purchase record matrix marks entries with 1 to indicate the buy and the rest of data are unknown
 - It does not have observed negative examples so one-class data is purely indiscriminate
- Implicit feedbacks can indirectly reflect opinions through user behavior
 - Users may deliberately choose to access which items [Marlin *et al*, 2007]

Confidence Modeling

- Confidence level
 - Observed chosen items imply more confidence of like over unchosen ones
 - Low confidence level to model users' dislike over unrated items since we have no evidence to prove the explicit dislike

- Weight Matrix (Confidence Matrix)

$$w_{k,i,j} = \begin{cases} c_{k,i,j} + 1 & (k,i,j) \text{ is observed} \\ 1 & \text{else} \end{cases}$$

Learning Algorithm

ALGORITHM 1. Weighted Irregular Tensor Factorization
$[U, V, C, \{P_k\}] = \text{WITF}(\{X_k\}, \{\omega_k\}, \{w_{k,i,j}\}, \lambda_U, \lambda_V, \lambda_C)$
Input: X_k is the data matrix for each domain ω_k is the influence weight for each domain $w_{k,i,j}$ is the weight on each entry $\lambda_U, \lambda_V, \lambda_C$ are the regularization parameters
Output: U is the factor matrix for users C is the factor matrix for domains $V, \{P_k\}$ are the factor matrices for items
Begin: Initialization: $\tilde{W}_{k,i,j} \leftarrow \omega_k w_{k,i,j}, V \leftarrow I$ Randomly initialize U, C $P_k \leftarrow A_R B_R^T$, with the SVD: $X_k^T U \Sigma_k V^T \approx A_R \Sigma_R B_R^T$ Iteration: Add neighbor noisy examples (optional): Randomly select S blank entries for each user i Fill neighbor noisy examples in the selected entries Generate tensor \mathcal{Y} with the slice for each domain k : $Y_k \leftarrow (\tilde{W}_k \otimes X_k) P_k$ Sub-iteration for $\{U, V, C\}$: Update $U_{i,:}$ in parallel for each user i using Eq. (23) Update $C_{k,:}$ in parallel for each domain k using Eq. (24) Update V using Eq. (25) Repeat 7-9 with m iterations Sub-iteration for $\{P_k\}$: Update P_k in parallel for each domain k using Eq. (22) Repeat 11 with n iterations Repeat 4 –12 until convergence Return $U, V, C, \{P_k\}$ End

Statistics of Epinions Dataset

- Covering 5 domains

Domain	# Items	# Ratings / # Users	# Ratings / # Items	Sparsity
Kids & Family*	3,769	4.9309	9.9077	0.0013
Hotels & Travel*	2,545	3.9210	11.6676	0.0015
Restaurants & Gourmet	2,543	3.3394	9.9446	0.0013
Wellness & Beauty	3,852	3.5481	6.9756	0.0009
Home and Garden	2,785	2.6003	7.0707	0.0009

Comparison Methods

- *kNN*: This is a baseline method to recommend movies watched by the top-k most similar groups.
- *MF-GPA*: This method performs matrix factorization (Salakhutdinov and Mnih 2008) on the group ratings that are aggregated from individual ratings through a specified strategy.
- *MF-IPA*: This method performs matrix factorization on individual ratings, and then aggregates the predicted ratings as the group ratings, using a specified strategy.
- *OCMF*: This method performs one-class MF (Hu et al. 2008) on the binary group ratings where the weights are set according to a specified strategy.
- *DLGR*: This is our deep learning approach, where the variance parameters of the DW-RBM (cf. the previous section) are set according to a specified strategy.
- *OCRBM*: This simply uses an RBM over the group choices without a connection to collective features. The variance parameters are set the same as the DW-RBM.

Rating Prediction on Epinions.com

- RMSE of comparative methods (the smaller the better)

Method	Target Domain	Kids & Family		Hotels & Travel	
		TR-80%	TR-50%	TR-80%	TR-50%
<i>kNN-CDCF</i>		1.2562	1.3016	1.1605	1.3338
<i>PMF-CDCF</i>		1.1719 [^]	1.3547 [^]	1.1260 [^]	1.2925 [^]
<i>CMF</i>		1.1312 [*]	1.2908 [*]	1.0805 [*]	1.2457 [*]
<i>PARAFAC2</i>		1.1102 [*]	1.1458 [*]	1.0647 [*]	1.0891 [*]
<i>CDTF</i>		1.0968 [*]	1.1219 [*]	1.0351 [*]	1.0585 [*]
<i>WITF</i>		1.1043 [*]	1.1293 [*]	1.0375 [*]	1.0619 [*]
<i>WITF+WRMF</i>		1.0563 ^{**}	1.0835 ^{**}	0.9983 ^{**}	1.0284 ^{**}

RMSEs of Comparison CDCF Methods on Epinions Dataset

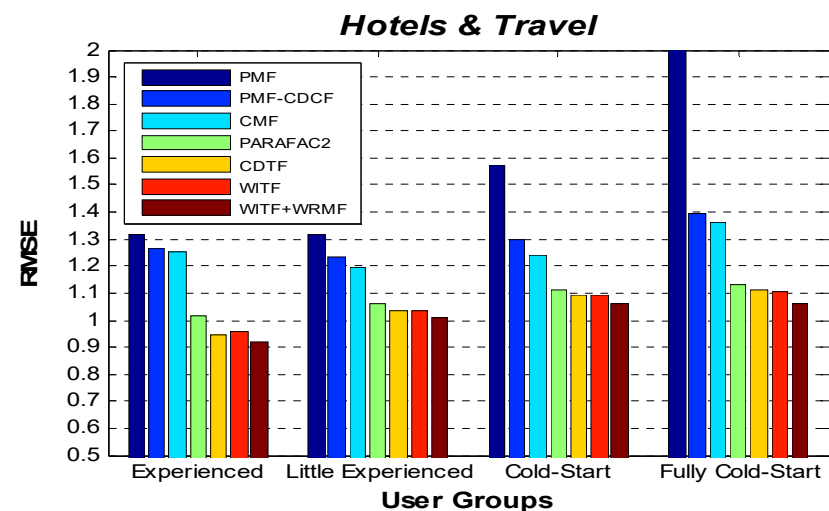
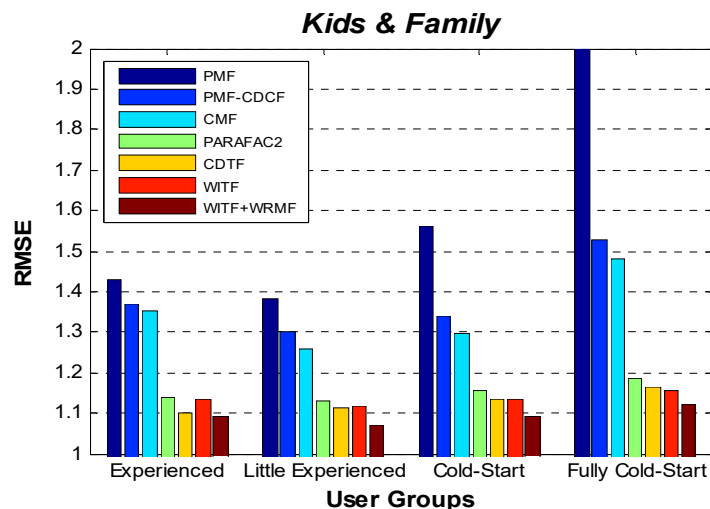
[^] baseline, ^{*} $p < 0.01$, ^{**} smallest p

Statistics of Testing Users Grouped by the Number of Ratings

User Group	# Ratings	Kids & Family	Hotels & Travel
		# testing users in TS-50%	# testing users in TS-50%
<i>Experienced</i>	> 20	120	55
<i>Little Experienced</i>	6 ~ 20	816	517
<i>Cold-Start</i>	1 ~ 5	2,260	2,807
<i>Fully Cold-Start</i>	0	695	1,072

The Prediction Performance over Different Numbers of Training Ratings

- RMSE of comparative methods (the smaller the better)



Click Statistics on Tmall.com Dataset

- One-class problem

Domain	# Items	# Clicks / # Users	# Clicks / # Items	Sparsity
<i>D1*</i>	8,179	23.2003	19.7170	0.0028
<i>D2*</i>	6,940	18.5455	18.5749	0.0027
<i>D3</i>	5,561	22.5005	28.1246	0.0040
<i>D4</i>	6,145	16.0606	18.1671	0.0026

The Mean AP@5,10 and nDCG@5,10

Target Domain Method	D1							
	TR-80%				TR-50%			
	AP@5	AP@20	nDCG@5	nDCG@20	AP@5	AP@20	nDCG@5	nDCG@20
Most-Pop	0.0161^	0.0175^	0.0269^	0.0382^	0.0322^	0.0223^	0.0567^	0.0577^
N-CDCF	0.0252*	0.0240*	0.0441*	0.0465*	0.0352*	0.0210	0.0604*	0.0534
MF-IF	0.0263*	0.0293*	0.0432*	0.0631*	0.0455*	0.0324	0.0813*	0.0854*
MF-IF-CDCF	0.0242*	0.0258*	0.0399*	0.0552*	0.0431*	0.0296	0.0763*	0.0775*
PARAFAC2	0.0213*	0.0226*	0.0350*	0.0476*	0.0395*	0.0267	0.0691*	0.0687*
CDTF-IF	0.0258*	0.0276*	0.0425*	0.0587*	0.0423*	0.0294	0.0758*	0.0767*
WITF	0.0267*	0.0285*	0.0451*	0.0623*	0.0484*	0.0340	0.0849*	0.0872*
WITF+WRMF	0.0271**	0.0290**	0.0462**	0.0643**	0.0486**	0.0343**	0.0851**	0.0879**

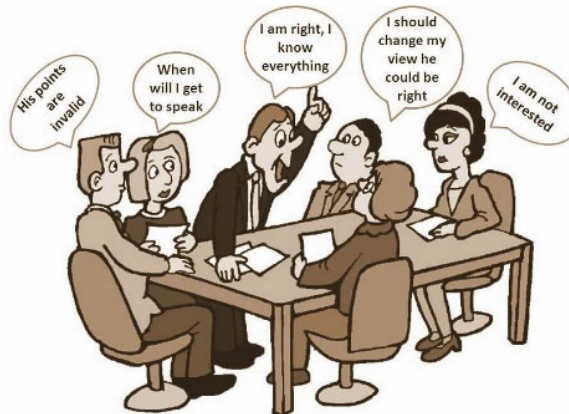
Target Domain Method	D2							
	TR-80%				TR-50%			
	AP@5	AP@20	nDCG@5	nDCG@20	AP@5	AP@20	nDCG@5	nDCG@20
Most-Pop	0.0175^	0.0194^	0.0288^	0.0424^	0.0297^	0.0231^	0.0530^	0.0591^
N-CDCF	0.0281*	0.0261*	0.0435*	0.0520*	0.0228	0.0243*	0.0380	0.0357
MF-IF	0.0320*	0.0354*	0.0528*	0.0747*	0.0501*	0.0370*	0.0872**	0.0924**
MF-IF-CDCF	0.0240*	0.0262*	0.0397*	0.0563*	0.0380*	0.0285*	0.0675	0.0724*
PARAFAC2	0.0215*	0.0234*	0.0356*	0.0506*	0.0327*	0.0251*	0.0589*	0.0638*
CDTF-IF	0.0326*	0.0337*	0.0526*	0.0662*	0.0454*	0.0316*	0.0761*	0.0750*
WITF	0.0338*	0.0363*	0.0552*	0.0753*	0.0538*	0.0383*	0.0905*	0.0909*
WITF+WRMF	0.0343**	0.0369**	0.0556**	0.0758**	0.0542**	0.0386**	0.0907**	0.0915*

Group-based Recommender Systems

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).

Group Choices Are Joint Decision

- Human beings are of a social nature, so various kinds of group activities are observed throughout life
 - Seeing a family movie, Planning family travel
- Each member of a group may have **different opinions** on the same items, so the main challenge in GRSs is to satisfy most group members with **diverse preferences**.
- This is not achieved through an individual-based recommendation method.



Profile Aggregation

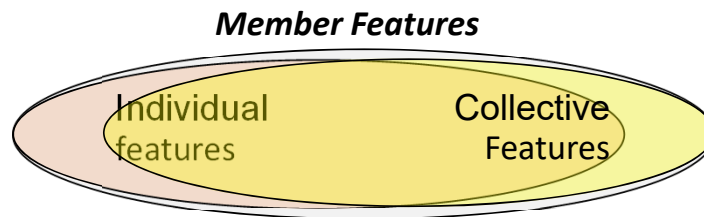
- Group Preference Aggregation (GPA)
 - GPA aggregates all members' ratings into a group profile, and then any individual-based CF approach can be used if it regards **groups as virtual individual users**.
- Individual Preference Aggregation (IPA)
 - IPA predicts the individual ratings over candidate items, and then aggregates the predicted ratings of members within a group via predefined strategies to represent group ratings.

Aggregation Strategies

- *Average* and *Least Misery* are the two most prevalent strategies (Masthoff 2011)
 - *Average strategy* recommends items with the highest average ratings over all members.
 - *Least misery strategy* assumes a group tends to be as happy as its **least happy member**.

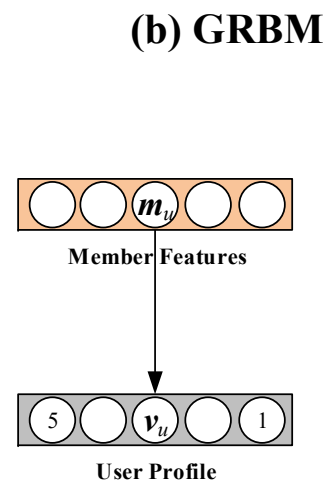
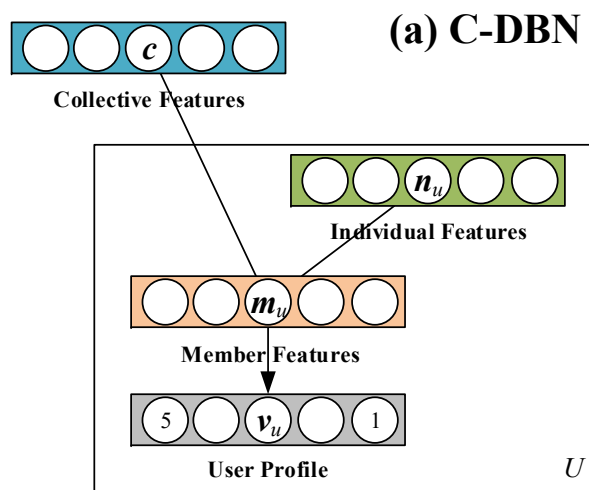
Modeling Features in Group-based Decision

- **Member Features:** these model the individual preference of a user when she/he makes choices as a group member, which can be regarded as **a mixture of Collective Features and Individual Features**.
- **Collective Features:** these represent compromised preferences of a group, which are **shared among all members** and can be disentangled from the *Member Features*.
- **Individual Features:** these represent **independent individual-specific preference**, which can be disentangled from the *Member Features* w.r.t. this user.

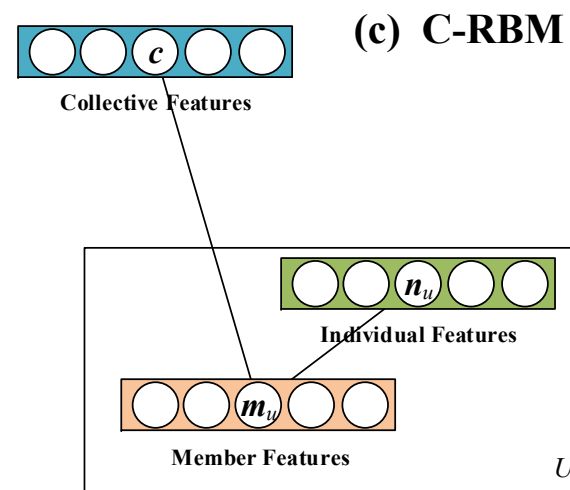


Disentangling Collective and Individual Features

- Each group choice can be regarded as a **joint decision** by all members



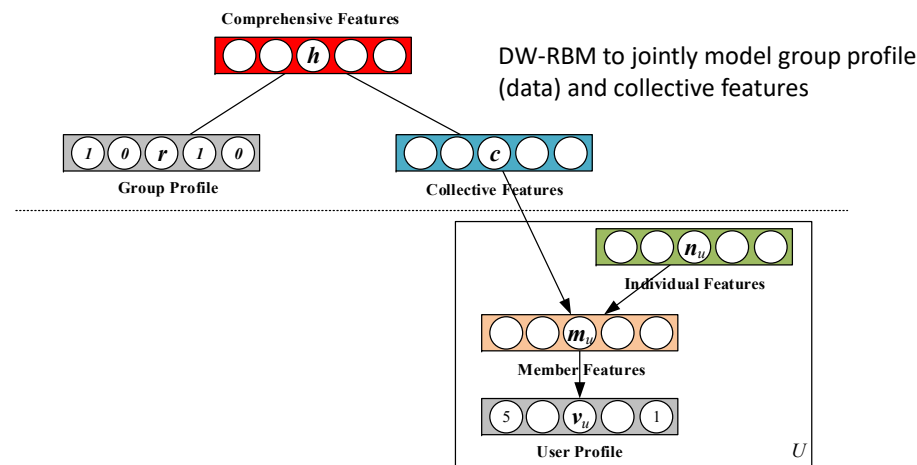
GRBM to learn
member features



C-RBM disentangles
collective features and
individual features from
member features

Comprehensive Representation of Group Preferences

- A dual-wing RBM is placed on the top of DBN, which jointly models the group choices and collective features to learn the **comprehensive** features of group preference



CAMRa2011 Dataset

- CAMRa2011 dataset containing the movie watching records of households and the ratings on each watched movie given by some group members.
- The dataset for track 1 of CAMRa2011 has 290 households with a total of 602 users who gave ratings (on a scale 1~100) over 7,740 movies.

Training and Testing Data

- Statistics of the evaluation data

Data	#Users/#Groups	#Ratings	Density
Train_{user}	602	145,069	0.0313
Train_{group}	290	114,783	0.0510
Eval_{group}	286	2,139	/

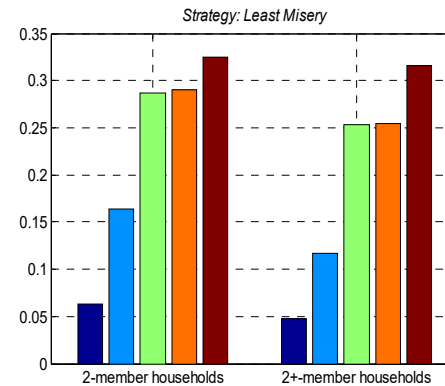
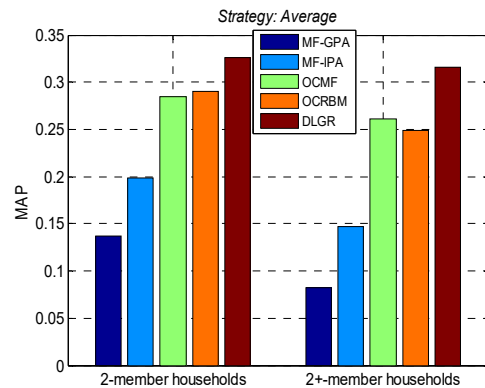
Results

MAP and mean AUC of all comparative models with different strategies

	MAP			AUC		
Model/Strategy	No Strategy	Average	Least Misery	No Strategy	Average	Least Misery
<i>kNN (k=5)</i>	0.1595	N/A	N/A	0.9367	N/A	N/A
<i>MF-GPA</i>	N/A	0.1341	0.0628	N/A	0.9535	0.9297
<i>MF-IPA</i>	N/A	0.1952	0.1617	N/A	0.9635	0.9503
<i>OCMF</i>	0.2811	0.2858	0.2801	0.9811	0.9813	0.9810
<i>OCRBM</i>	0.2823	0.2922	0.2951	0.9761	0.9778	0.9782
<i>DLGR</i>	0.3236	0.3252	0.3258	0.9880	0.9892	0.9897

Group with Different Number of Members

- A group with **more members implies more different preferences**, so it is harder to find recommendations satisfying all members.
- Each household may contain 2~4 members in this dataset. We additionally evaluated the MAP w.r.t. 2-member households and the 2⁺-member (>2) households under *Average* and *Least Misery* strategies.

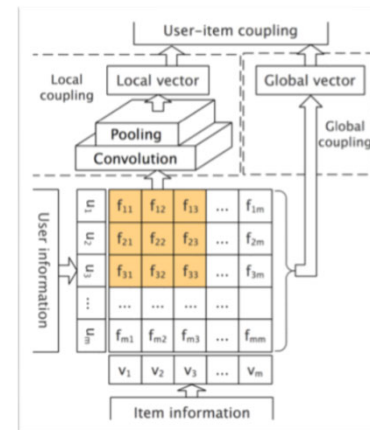


More Recent Work on non-IID recommender systems

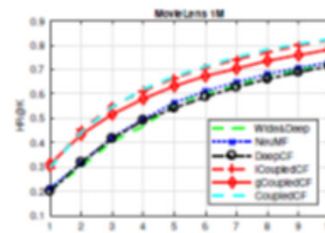
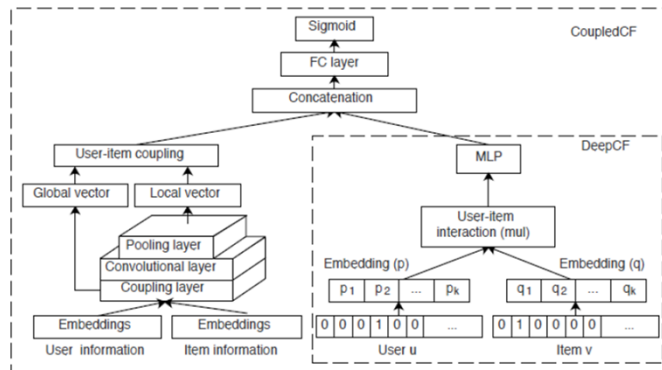
- *Trong Dinh Thac Do and Longbing Cao. Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence, NIPS2018*
- *CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering, IJCAI2018*
- *Interpretable Recommendation via Attraction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

Deep Representation with Explicit and Implicit Feature Couplings

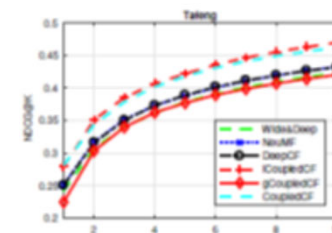
- Learn explicit user-product couplings by metadata-enabled CNN
- Build a deep collaborative filter model to learn the latent user-product relations
- Integrate both local and global user-product interactions components



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



(a) HR@K on MovieLens

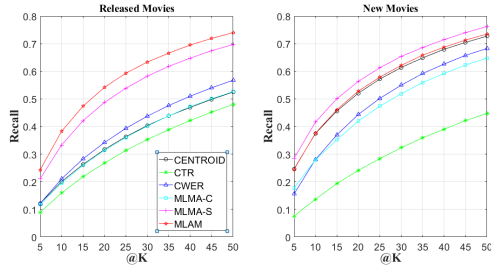


(b) NDCG@K on Tafeng

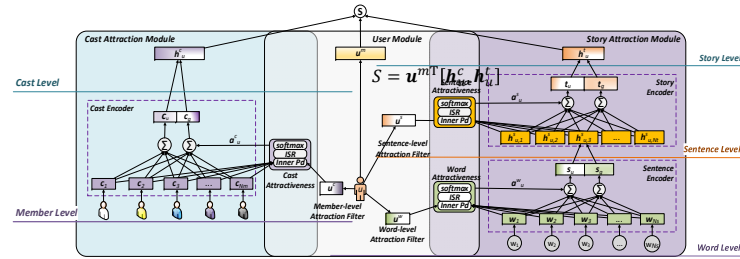
- *CoupledCF: Learning Explicit and Implicit User-item Couplings in Recommendation for Deep Collaborative Filtering, IJCAI2018*

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



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



$$a_u^{c_i} = \text{softmax}(\text{isr}(\mathbf{u}^c \mathbf{c}_i)) \quad \mathbf{c}_u = \sum a_u^{c_i} \mathbf{c}_i \quad a_u^{w_i} = \text{softmax}(\text{isr}(\mathbf{u}^w \mathbf{w}_i)) \quad \mathbf{s}_u = \sum a_u^{w_i} \mathbf{w}_i$$

$$a_u^{s_i} = \text{softmax}(\text{isr}(\mathbf{u}^s \mathbf{h}_i^s)) \quad \mathbf{t}_u = \sum a_u^{s_i} \mathbf{h}_i^s$$

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

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 granted 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 granted 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

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.

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

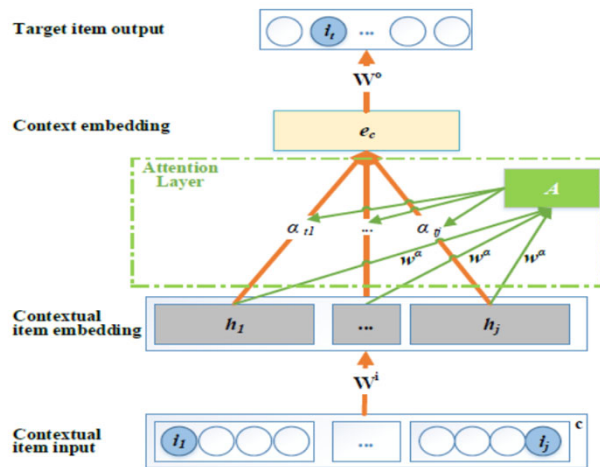


Figure 1: The ATEM architecture, which first learns item embeddings and then integrates them into the context embedding for target item prediction, where ‘A’ represents the attention model.

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

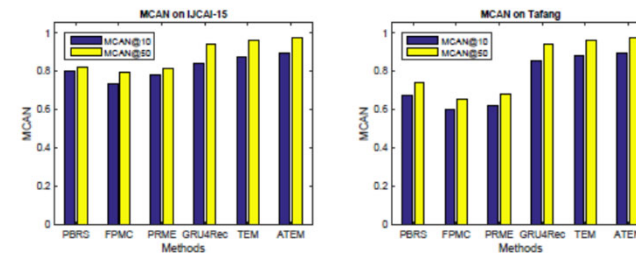
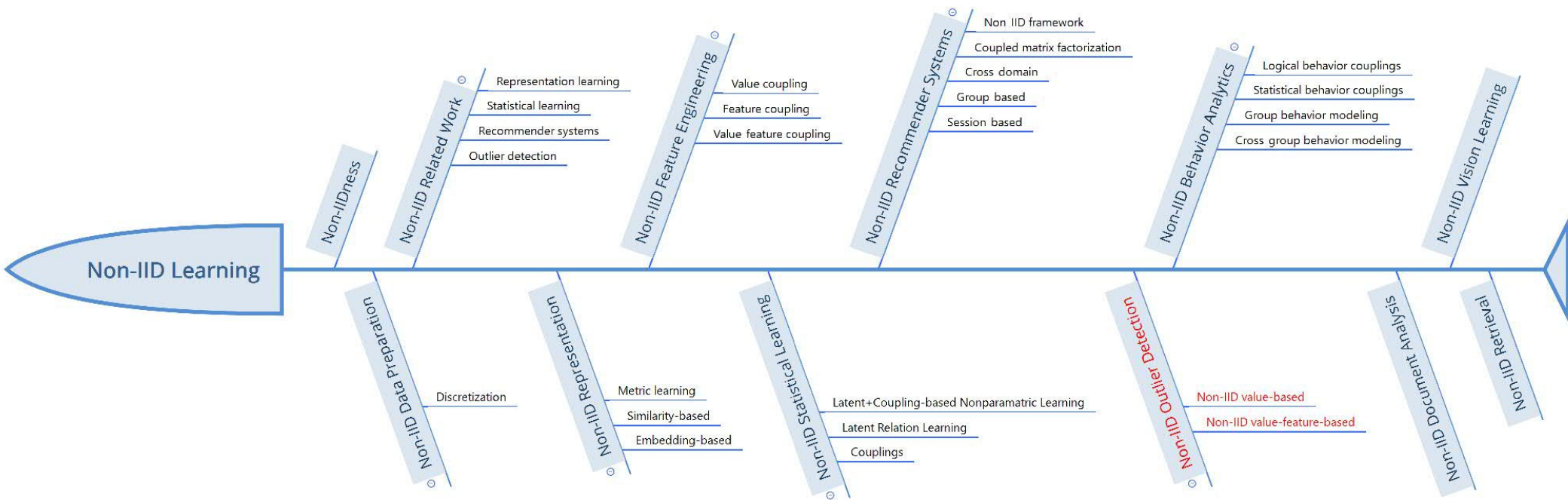


Figure 3: ATEM achieves higher novelty than the other approaches.

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



Non-IID Outlier Detection



Background and Non-IID Outliers

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

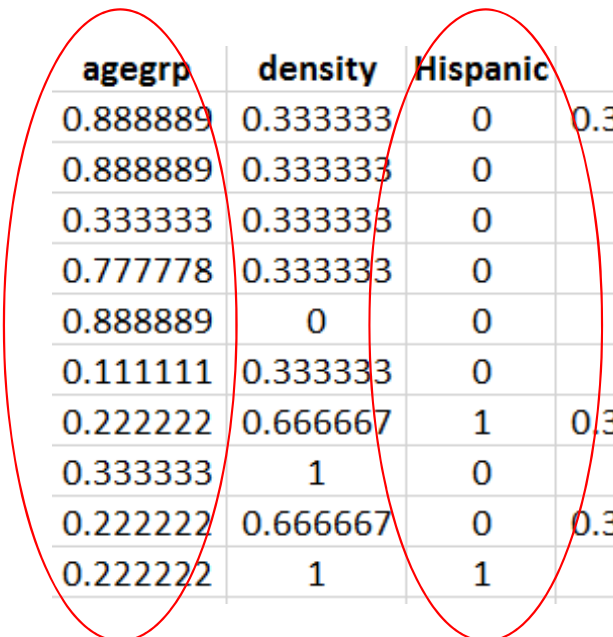
- Statistical/probabilistic-based approach
 - Statistical test-based → *deviation from distribution*
 - Depth-based → *data depth*
 - Deviation-based → *sensitivity or uncertainty*
- Proximity-based approach
 - Distance-based → *nearest neighbor distances*
 - Density-based → *local density*
 - Clustering-based → *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

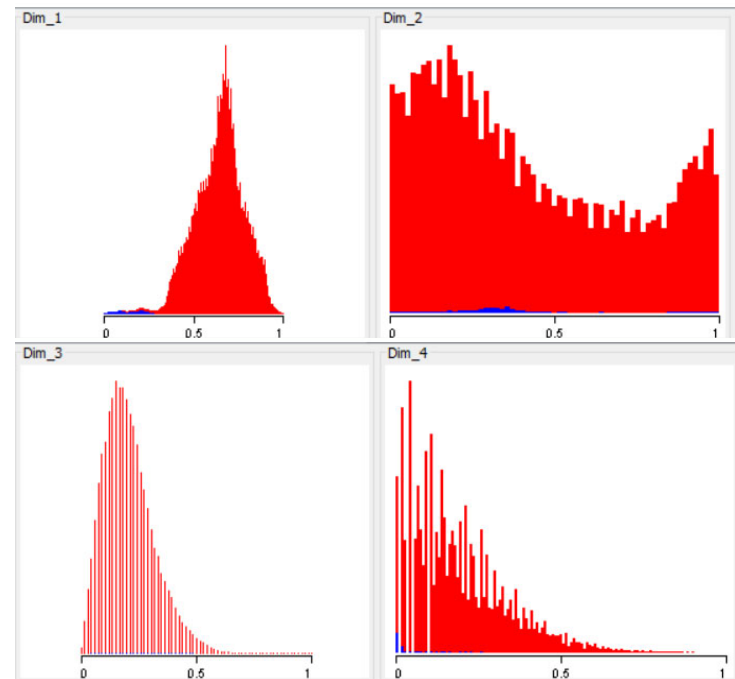
Non-IID Real-life Data

Couplings



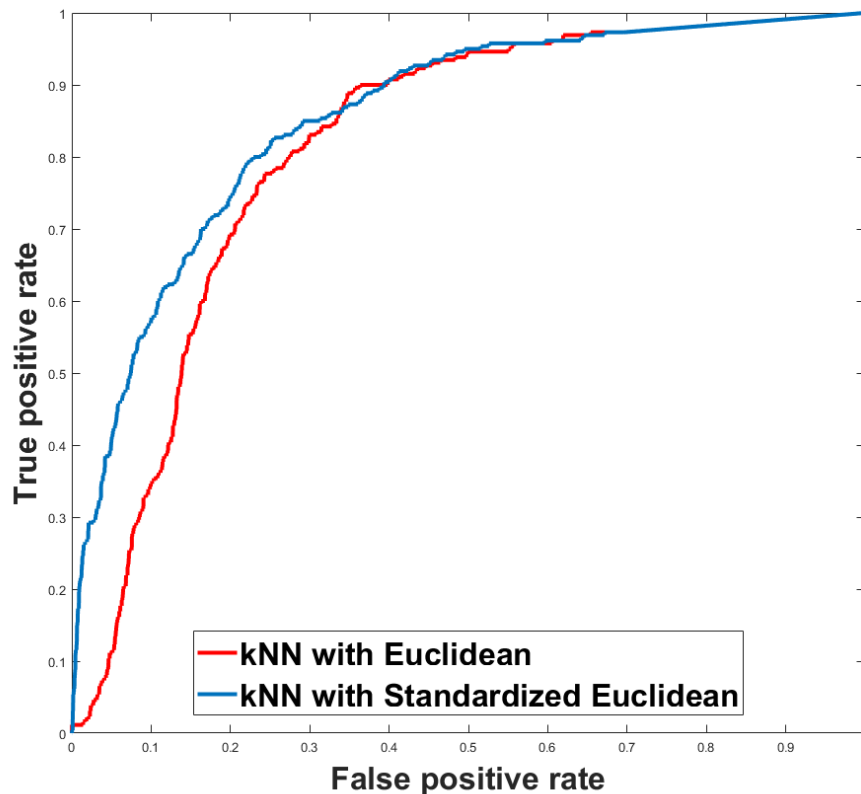
Source: <http://www.diabeticrockstar.com>

Heterogeneity



Four features from the *CoverType* data set

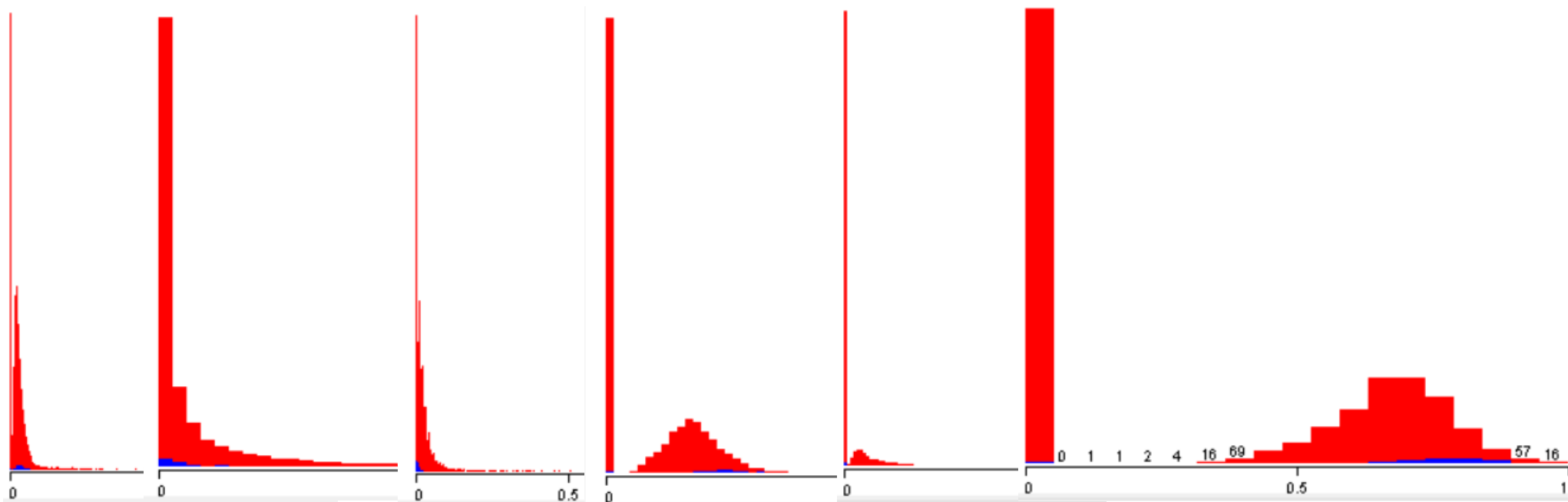
IID vs. Non-IID Outlier Detection – example



- ***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. Identifying Outliers in Complex Categorical Data by Modeling Feature Value Couplings. IJCAI16.

Motivation

- **Value heterogeneity**

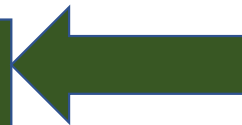
- Semantic differs in different contexts



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**

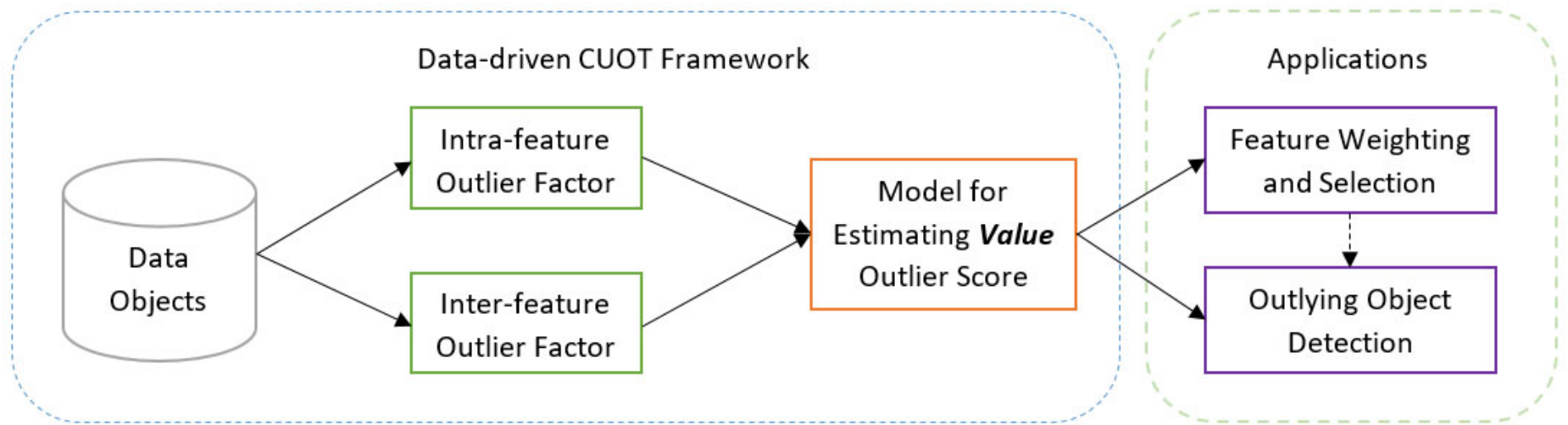
- “A man is known by the company he keeps”
 - Homophily couplings in outlying behaviors (values)

- **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 $v \in \text{dom}(f)$

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

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

CBRW: Inter-feature Outlier Factor

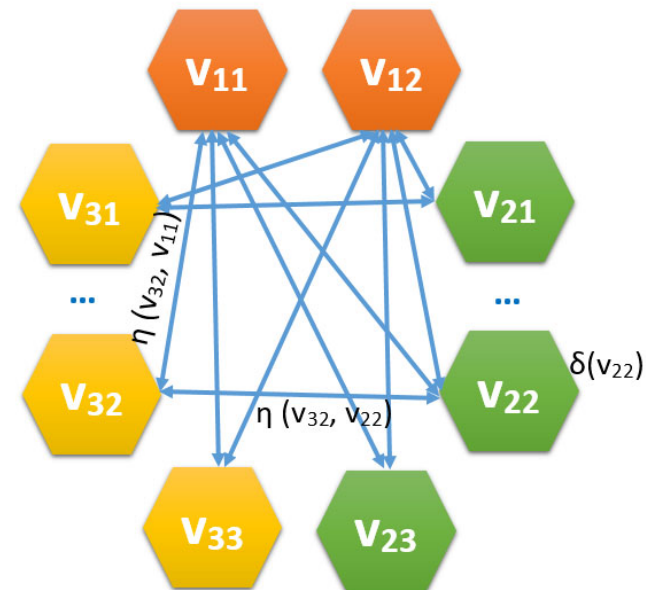
- **Inter-feature** outlier factor capturing the homophily value couplings
 - **Concurrent rare** values have high mutual conditional probabilities

$$\mathbf{q}_v = [\eta(u, v), \dots, \eta(w, v)]^\top = \left[\frac{\text{freq}(u, v)}{\text{freq}(v)}, \dots, \frac{\text{freq}(w, v)}{\text{freq}(v)} \right]^\top, \forall u, w \in V \setminus v$$

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



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

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

Analysis of CBRW

- Convergence guaranteed

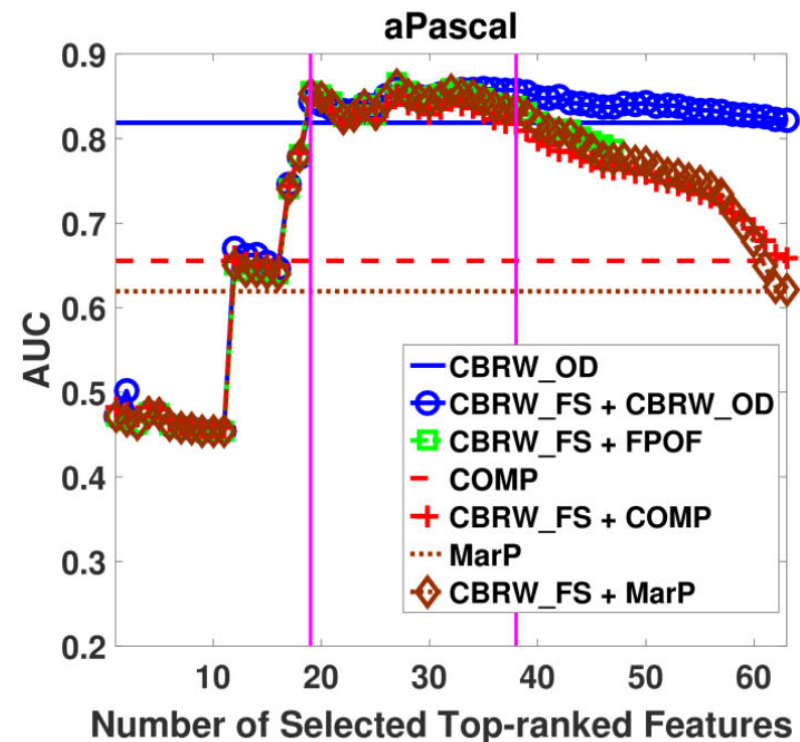
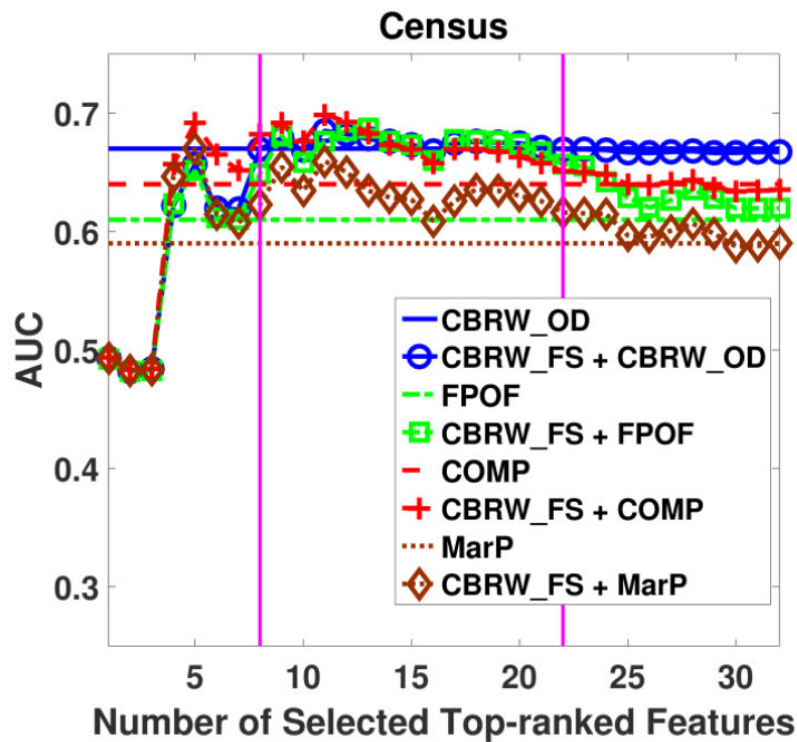
$$\pi_{t+1} = (1 - \alpha) \frac{1}{|\mathcal{V}|} \mathbf{1} + \alpha \mathbf{W}^b \pi_t$$

- Fast convergence rate
 - Small graph diameter, e.g., 2
 - Large graph density or Cheeger constant

Performance Evaluation I: 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
p-value	CBRW vs.	0.7959	<u>0.0392</u>	<u>0.0012</u>	<u>0.0008</u>	<u>0.0115</u>	<u>0.0147</u>	<u>0.0040</u>
		CBRWie vs.	0.4225	0.0969	0.0592	0.4316	0.3167	<u>0.0446</u>
			CBRWia vs.	0.1460	0.1223	0.2886	0.8490	0.0979

Performance Evaluation II: Outlying Feature Selection Performance

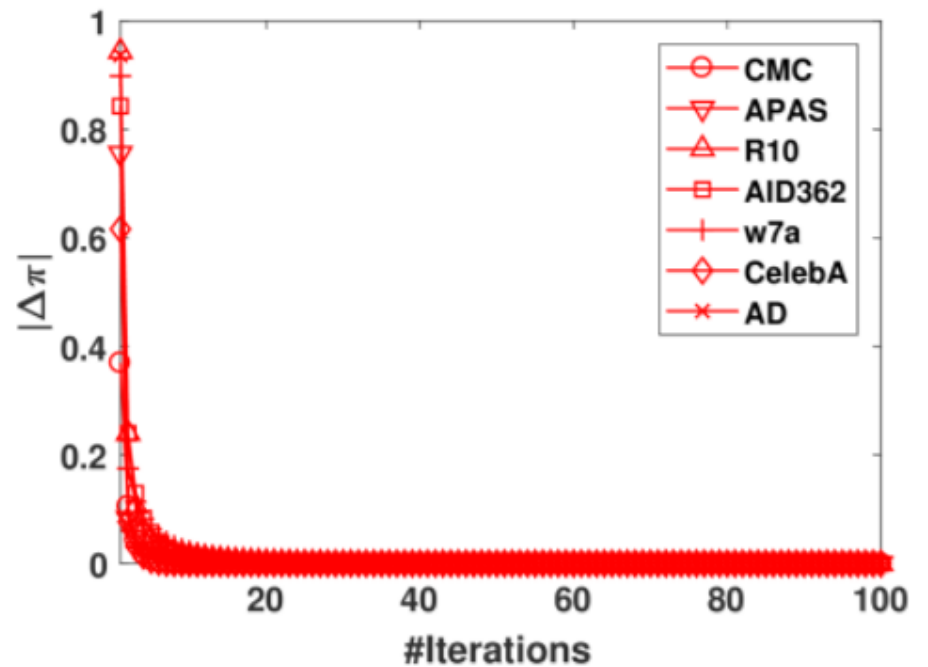
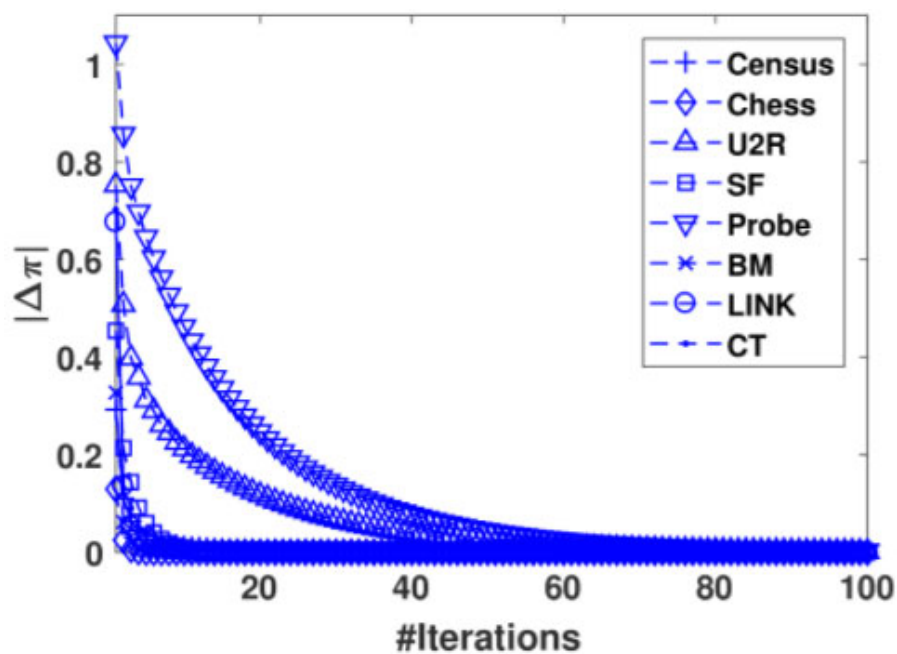


Performance Evaluation III: Convergence Analysis

- Characteristics of value graphs
 - Small graph diameter
 - Large graph density

Data	Diameter	Coefficient
Census	2	0.76
Chess	2	0.79
U2R	2	0.80
SF	2	0.81
Probe	2	0.82
BM	2	0.85
LINK	2	0.86
CT	2	0.87
CMC	2	0.89
APAS	2	0.90
R10	2	0.91
AID362	2	0.92
w7a	2	0.93
CelebA	2	0.99
AD	o	o

Performance Evaluation III: Convergence Analysis



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 I

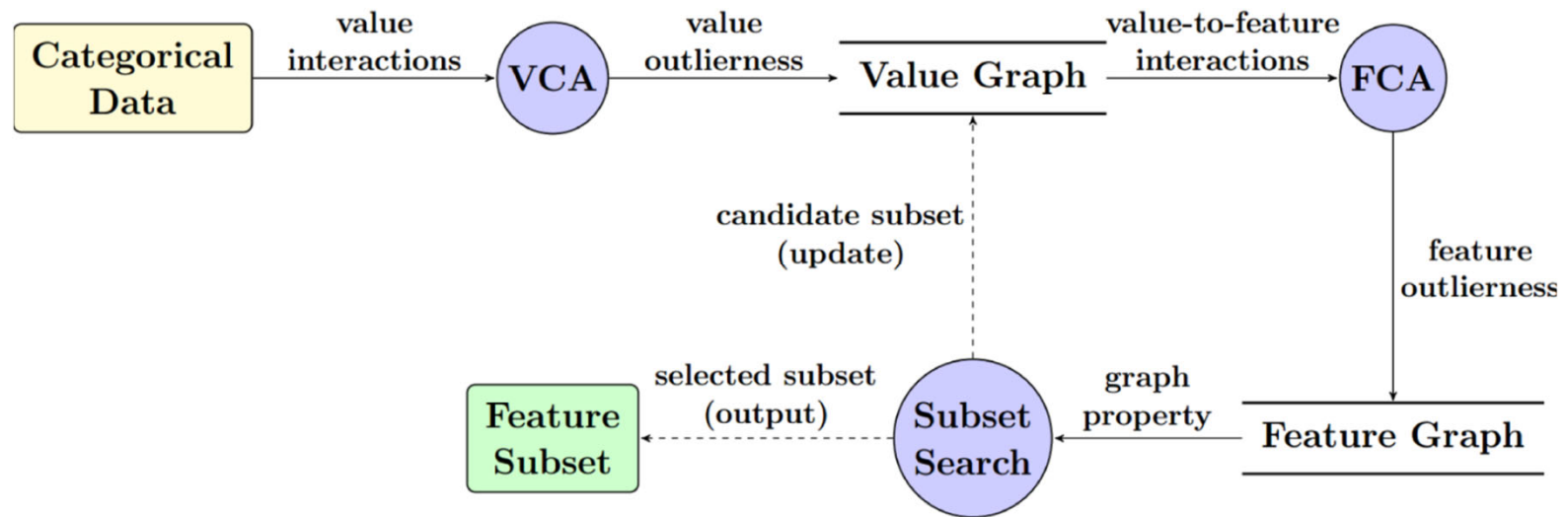
Guansong Pang, Longbing Cao, Ling Chen, Huan Liu. Unsupervised Feature Selection for Outlier Detection by Modelling Hierarchical Value-Feature Couplings. IEEE ICDM 2016, pp. 410-419.

Motivation

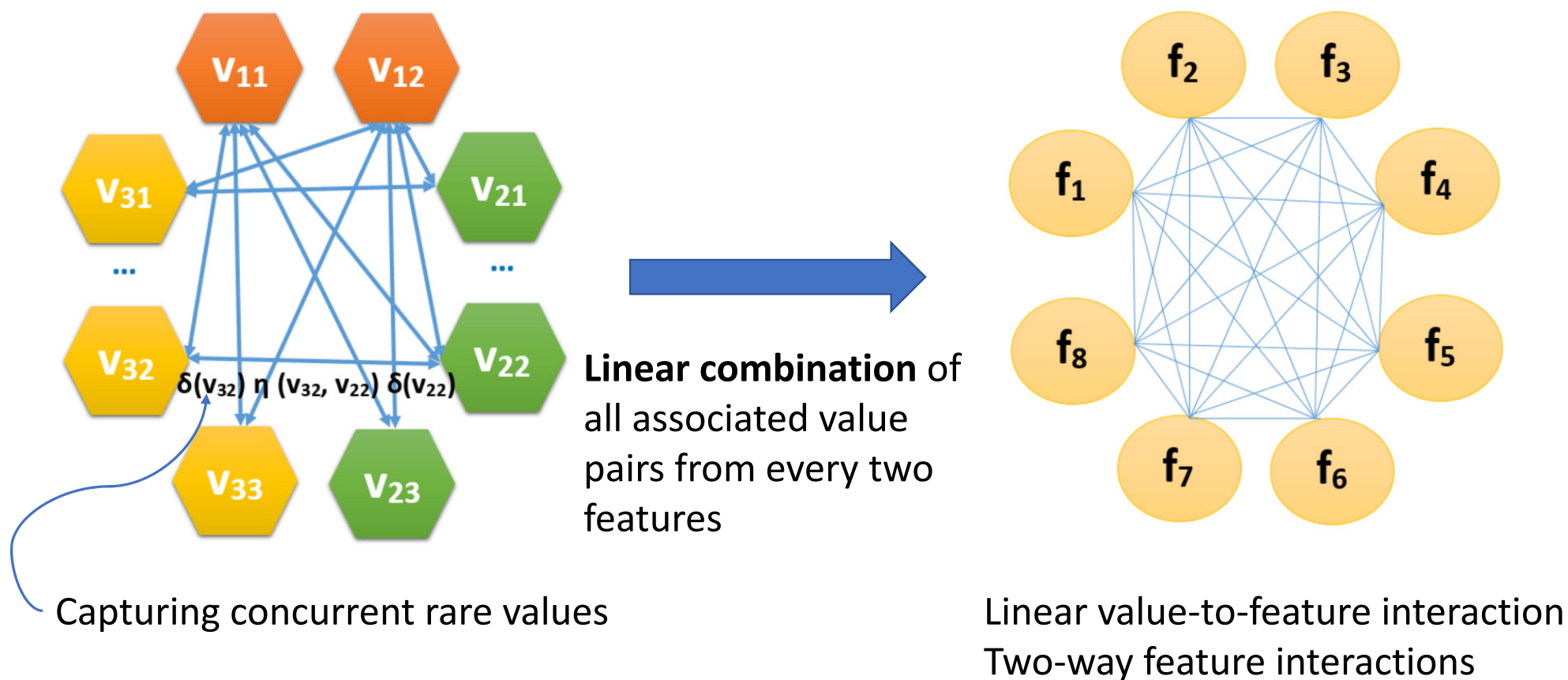
- Feature selection for outlier detection
 - Outliers are masked as normal objects by noisy features
 - Useless features downgrade detection efficiency
- Challenges
 - Unsupervised nature – no class labels
 - Complex feature interactions

Our Framework

- Two-way feature interactions
 - Estimate feature outlieriness by modeling value-to-feature couplings



DSFS: Value and Feature Graph Construction



DSFS: Objective Function

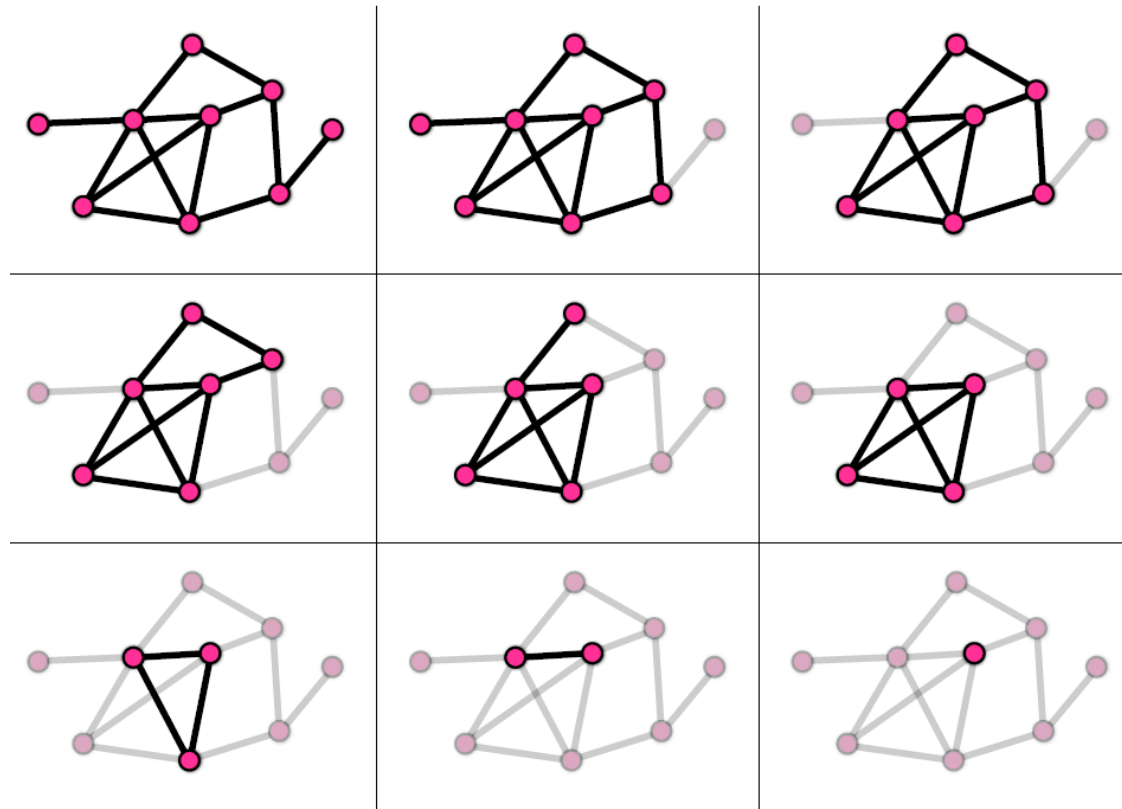
- Feature selection objective function

$$\max_{S \in \mathcal{F}} \frac{1}{|S|} \sum_{f \in S} \sum_{f' \in S} A^*(f, f')$$

where A^* is the weighted adjacent matrix of feature graph

- It is equivalent to finding the densest subgraph
- It can be solved by a **linear-time** greedy search method with a **2-approximation** guarantee

DSFS: Dense Feature Subgraph Search



From Gionis and Tsourakakis. *DSDTutorial at KDD15*

DSFS: The Algorithm

Input: \mathcal{X} - data objects

Output: \mathcal{S} - the feature subset selected

- 1: Initialise \mathbf{A} as a $|\mathcal{V}| \times |\mathcal{V}|$ matrix
- 2: $A(v, v') \leftarrow g(v, v'), \forall v, v' \in \mathcal{V}$
- 3: Initialise \mathbf{A}^* as a $|\mathcal{F}| \times |\mathcal{F}|$ matrix
- 4: $A^*(f, f') \leftarrow h(f, f'), \forall f, f' \in \mathcal{F}$
- 5: Set $\mathcal{S} \leftarrow \mathcal{F}$ and $s \leftarrow \text{den}(\mathbf{A}^*)$
- 6: **for** $i = 1$ to D **do**
- 7: Find f that has the smallest weighted degree in \mathbf{A}^*
- 8: $\mathcal{F} \leftarrow \mathcal{F} \setminus f$ and update \mathbf{A}^*
- 9: $\mathcal{S} \leftarrow \mathcal{F}$ and $s \leftarrow \text{den}(\mathbf{A}^*)$ if $s \leq \text{den}(\mathbf{A}^*)$
- 10: **end for**
- 11: **return** \mathcal{S}

Performance Evaluation I: Improving AUC Performance

Data Set	Acronym	κ_{nos}	κ_{rdn}	N	D	D'	RED
BankMarketing	BM	90%	0%	41188	10	4	60%
aPascal	-	81%	0%	12695	64	20	69%
Sylva	-	78%	0%	14395	87	66	24%
Census	-	58%	0%	299285	33	10	70%
CelebA	-	49%	4%	202599	39	34	13%
CMC	-	38%	4%	1473	8	5	38%
CoverType	CT	34%	22%	581012	44	5	89%
Chess	-	33%	0%	28056	6	4	33%
U2R	-	17%	7%	60821	6	3	50%
SolarFlare	SF	9%	0%	1066	11	8	27%
Optdigits	DIGIT	8%	26%	601	64	46	28%
Mushroom	MRM	5%	2%	4429	22	13	41%
Advertisements	AD	5%	78%	3279	1555	49	97%
Probe	-	0%	7%	64759	6	2	67%
Linkage	LINK	0%	0%	5749132	5	4	20%
Avg.		34%	10%	470986	131	18	48%

- 15 real-world data sets are used
- Remove 13%-97% features
- On average, 48% features are eliminated

Performance Evaluation I: Improving AUC Performance

AUC Performance									
	MarP	MarP*	IMP	COMP	COMP*	IMP	FPOF	FPOF*	IMP
BM	0.56	0.59	5%	0.63	0.62	-2%	0.55	0.58	5%
aPascal	0.62	0.88	42%	0.66	0.88	33%	○	0.88	○
Sylva	0.96	0.96	0%	0.95	0.96	1%	○	○	○
Census	0.59	0.69	17%	0.64	0.71	11%	0.61	0.72	18%
CelebA	0.74	0.74	0%	0.76	0.76	0%	0.74	0.75	1%
CMC	0.54	0.66	22%	0.57	0.66	16%	0.56	0.65	16%
CT	0.98	0.97	-1%	0.98	0.97	-1%	0.98	0.97	-1%
Chess	0.64	0.64	0%	0.64	0.63	-2%	0.62	0.61	-2%
U2R	0.88	0.92	5%	0.99	0.99	0%	0.92	0.97	5%
SF	0.84	0.85	1%	0.85	0.86	1%	0.86	0.86	0%
DIGIT	0.95	0.95	0%	0.97	0.97	0%	0.96	0.94	-2%
MRM	0.89	0.89	0%	0.93	0.94	1%	0.91	0.91	0%
AD	0.70	0.74	6%	●	0.75	●	○	0.74	○
Probe	0.98	0.98	0%	0.98	0.98	0%	0.99	0.98	-1%
LINK	1.00	1.00	0%	1.00	1.00	0%	1.00	1.00	0%
Avg.			6%			4%			3%

3%-6% improvement to three different types of outlier detectors

Performance Evaluation II: Reducing Runtime

Runtime (s)									
	MarP	MarP*	SU	COMP	COMP*	SU	FPOF	FPOF*	SU
BM	0.17	0.15	1	212.46	170.43	1	0.85	0.57	1
aPascal	0.31	0.12	3	451.36	41.00	11	○	53.29	○
Sylva	0.21	0.20	1	1137.07	498.59	2	○	○	○
Census	1.62	0.51	3	18174.49	12878.14	1	30790.78	75.23	409
CelebA	0.89	0.82	1	1647.47	1169.27	1	159377.51	50188.65	3
CMC	0.14	0.01	11	5.14	2.42	2	0.10	0.06	2
CT	3.14	0.36	9	3914.33	341.98	11	410016.55	1.09	377547
Chess	0.12	0.08	1	95.35	49.30	2	0.42	0.18	2
U2R	0.28	0.13	2	318.95	255.28	1	0.39	0.22	2
SF	0.02	0.01	1	6.33	4.40	1	0.39	0.09	4
DIGIT	0.04	0.03	1	217.10	111.51	2	10196.85	31.99	319
MRM	0.07	0.07	1	48.72	32.18	2	19.32	2.70	7
AD	0.85	0.10	9	●	126.35	●	○	54088.52	○
Probe	0.28	0.11	3	576.08	456.00	1	0.47	0.20	2
LINK	2.74	2.27	1	6365.26	5203.67	1	23.56	17.93	1
Avg.			3			3			31525

May gain up to six orders of magnitude faster

Conclusions

- A novel and flexible framework is introduced for outlying feature selection
- The instance DSFS is **parameter-free** and retains **2-approximation** to the optimum
 - Remove about 50% features while achieve 3-6% AUC improvements
 - Perform comparably well even when filtering out about 90% features
 - Two to six orders of magnitude speedup
 - Good scalability: linear w.r.t. data size and quadratic w.r.t. dimensionality

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)

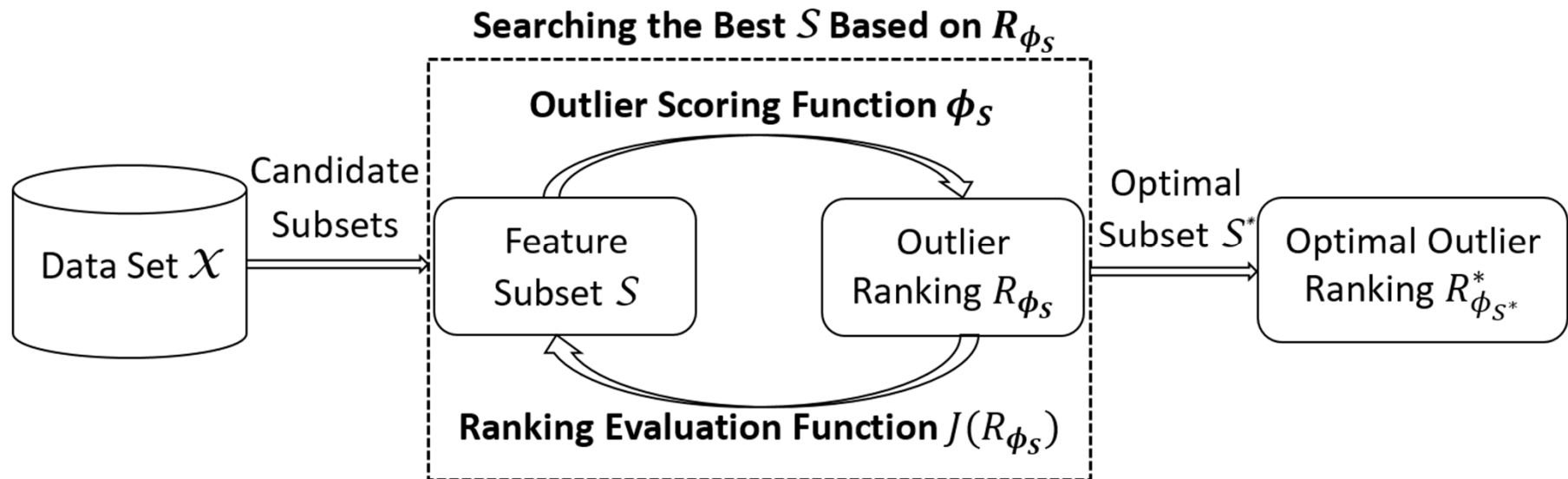
- Existing solutions: subspace/feature selection + OD
- Subspace/feature selection is independent from OD
 - Noisy features bias the subspace/feature search
 - Not optimal w.r.t. subsequent OD method
- Our solution: Simultaneous feature selection and outlier detection
 - **Wrapper approach** for this joint optimization



Filter
approach

Our WrapperOD Framework

Wrapper approach for joint optimization of feature selection and OD

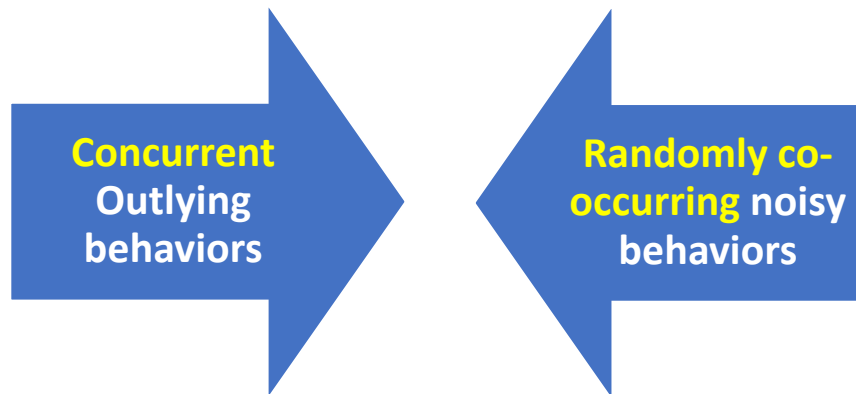


Challenge 1: how to ensure the outlier scoring efficacy

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

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

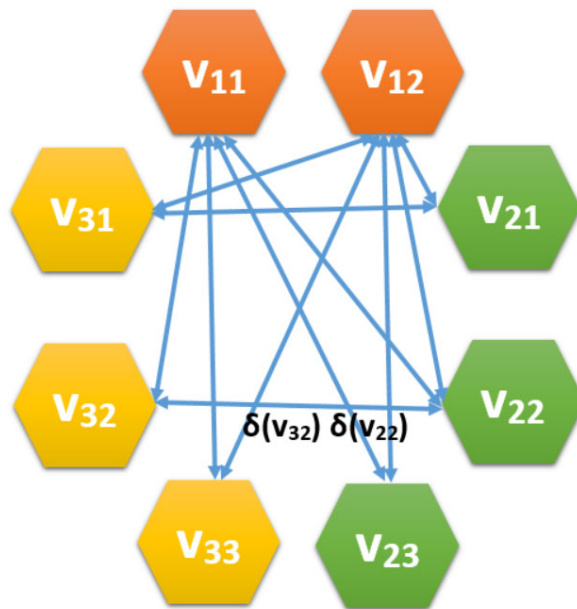
- The scoring function should at least be
 - **Sufficiently resilient to noisy features**
 - **Very efficient**
- Homophily couplings between outlying values



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

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

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

The WrapperOD Instance: HOUR

Algorithm 1 $HOUR(\mathcal{X}, k)$

Input: \mathcal{X} - data objects, k - the number of targeted outliers

Output: R - an outlier ranking of objects, \mathcal{S} - a feature subset

```
1:  $\psi(v) \leftarrow \sum_{u \in \mathcal{N}_v} \rho(u, v) \tau(u), \forall v \in \mathcal{V}$ 
2: Compute  $\phi_{\mathcal{F}}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$ 
3:  $r \leftarrow J(R_{\phi_{\mathcal{F}}}, k)$ 
4: while  $|\mathcal{F}| > 0$  do
5:   for  $i = 1$  to  $|\mathcal{F}|$  do
6:     Compute  $\phi_{\mathcal{F} \setminus f_i}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$ 
7:     Compute  $J_i(R'_{\phi_{\mathcal{F}}}, k)$ 
8:   end for
9:   Find feature  $f_i$  with the largest  $J_i(R'_{\phi_{\mathcal{F}}}, k)$ 
10:   $\mathcal{F} \leftarrow \mathcal{F} \setminus f_i$  and update  $\psi(v)$  for all  $v$  contained in  $\mathcal{F}$ 
11:  if  $J_i(R'_{\phi_{\mathcal{F}}}, k) \geq r$  then
12:     $R \leftarrow R', \mathcal{S} \leftarrow \mathcal{F}$  and  $r \leftarrow J_i(R'_{\phi_{\mathcal{F}}}, k)$ 
13:  end if
14: end while
15: return  $R$  and  $\mathcal{S}$ 
```

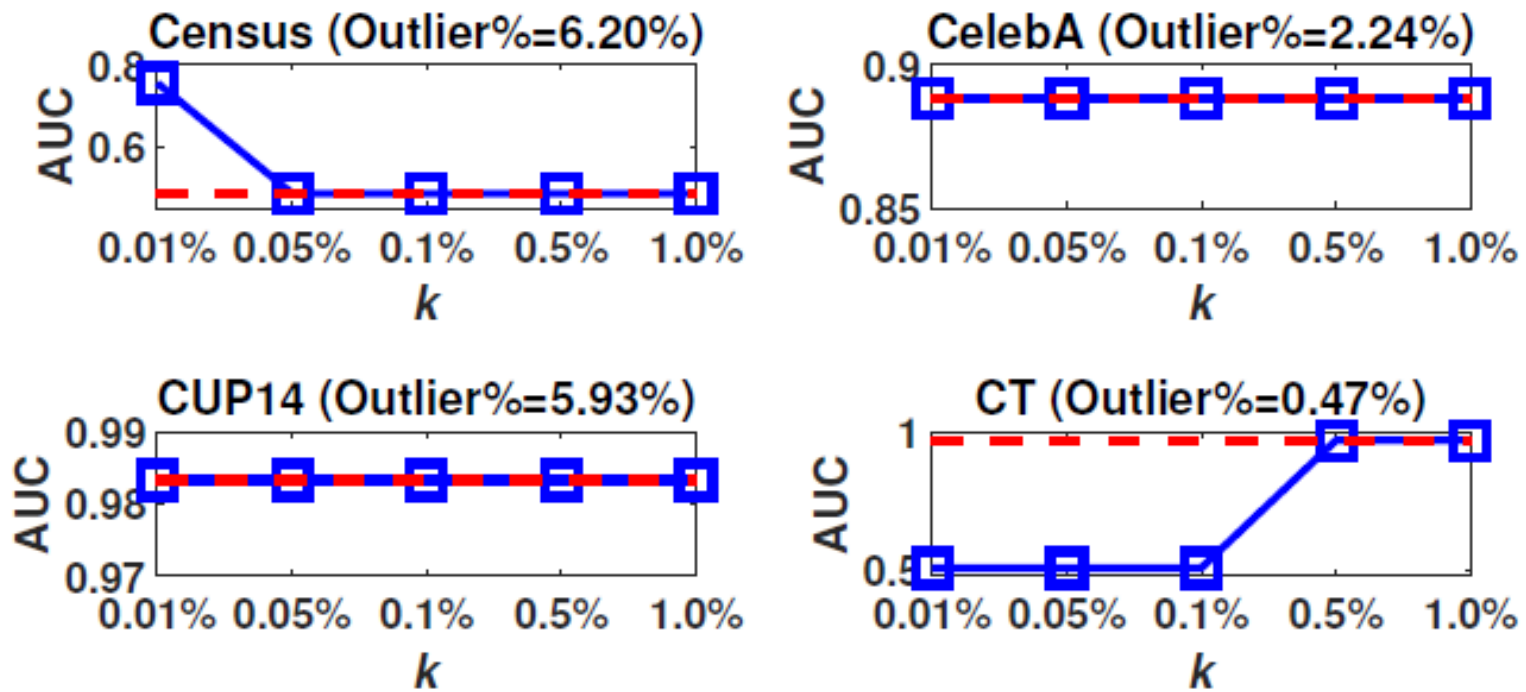
Performance Evaluation I: Comparing to State-of-the-art Detectors

					AUC				$P@n$			
Data	N	$ \mathcal{F} $	$ \mathcal{S} (\nabla)$	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

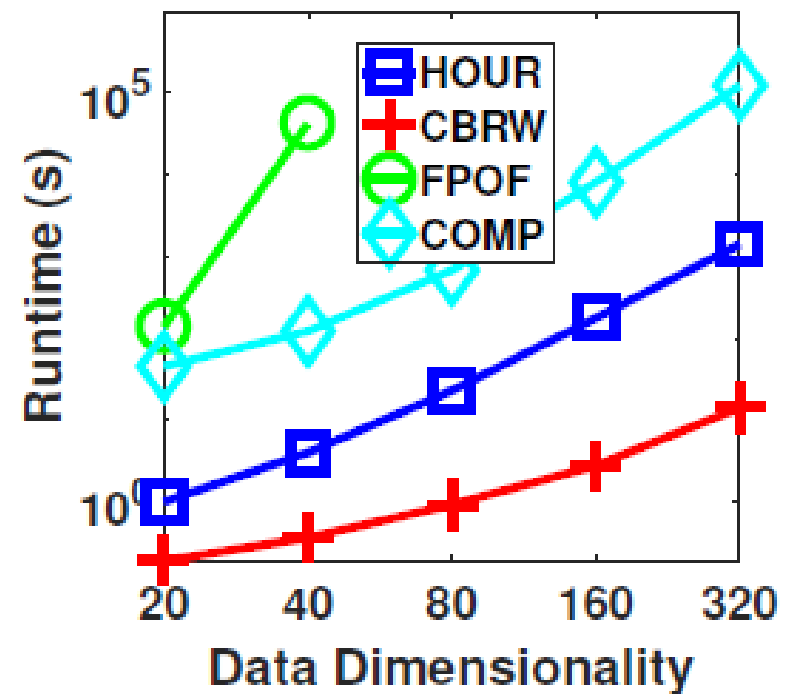
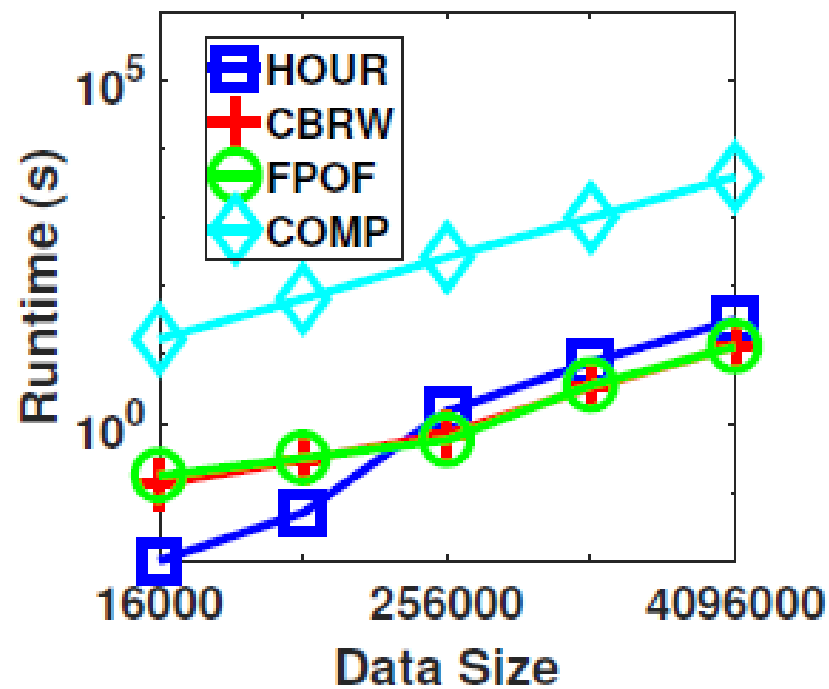
Performance Evaluation II: Comparing to State-of-the-art FS + Detectors

Data	AUC				
	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

Performance Evaluation III: Sensitivity Test



Performance Evaluation IV: 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>

Non-IID Statistical Learning

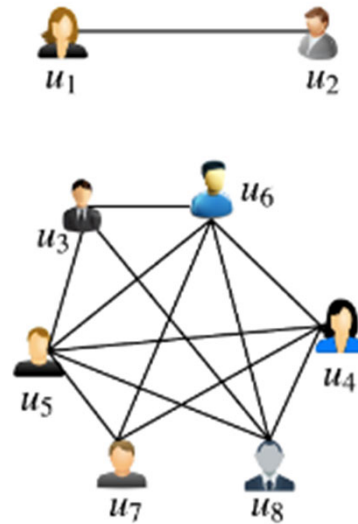
PAKDD2019 Tutorial on Large-scale statistical learning

www.datasciences.org

Large-scale, sparse, multi-source data: Non-IID

	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:

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

Bayesian probabilistic models

- $X = \{x_1, x_2, \dots, x_n\}$ 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

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

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

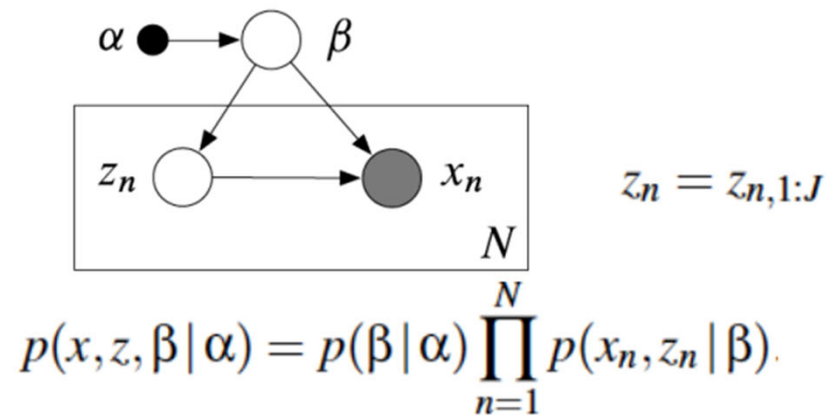
E.g., Gamma-Gamma-Gamma-Poisson distribution Compound models

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



- Our goal: approximate the posterior

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

- Locally independence

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

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

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

Stochastic variational inference (SVI)

- ELBO

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

Copula Mixed-Membership Stochastic Blockmodel

Fan, X., Da Xu, R. Y., & Cao, L. (2016). Copula Mixed-Membership Stochastic Blockmodel. In *IJCAI* (pp. 1462-1468).

Motivation

- Group members may have higher correlated interactions towards the ones within the same group.
 - For instance, in a company, IT support team members tend to co-interact with each other more than with employees of other departments.
- In reality, within a social networking context, it is important to incorporate group member interactions (here called intra-group correlations) into the modeling of membership indicators.

Our Model

- Mixed Membership Stochastic Model (MMSB) which focuses on detecting overlapping communities of the complex networks.
- Further coupling learning of members in the same group using Copula.

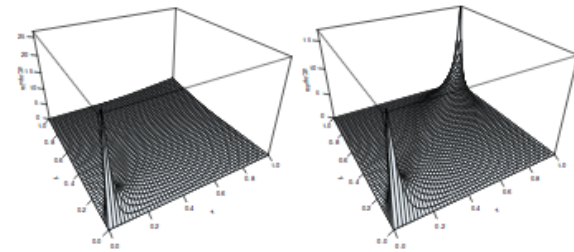


Figure 1: Clayton Copula (2) and Gaussian Copula (0.9) visualization.

$$H(x, y) = C(F(x), G(y))$$

$$h(x, y) = c(F(x), G(y)) \cdot f(x)g(y)$$

The Graphical Model

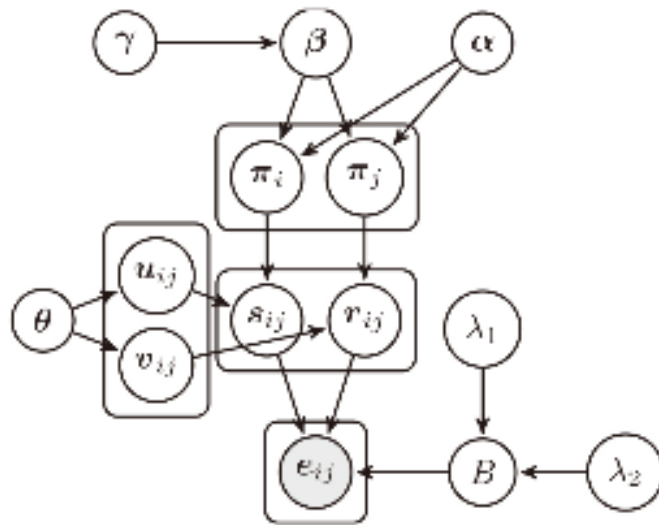


Figure 2: Graphical model of Copula MMSB

$$C1: \beta \sim GEM(\gamma)$$

$$C2: \{\pi_i\}_{i=1}^n \sim DP(\alpha \cdot \beta)$$

$$C3: \begin{cases} (u_{ij}, v_{ij}) \sim Copula(\theta), & g_{ij} = 1; \\ u_{ij}, v_{ij} \sim U(0, 1), & g_{ij} = 0. \end{cases}$$

$$C4: s_{ij} = \Pi_i^{-1}(u_{ij}), r_{ij} = \Pi_j^{-1}(v_{ij})$$

$$C5: B_{k,l} \sim Beta(\lambda_1, \lambda_2), \forall k, l;$$

$$C6: e_{ij} \sim Bernoulli(B_{s_{ij}, r_{ij}}).$$

Empirical Results

Table 3: Model Performance (Mean \pm Standard Deviation) on Real-world Datasets.

Dataset		Train error	Test error	Test log likelihood	AUC
NIPS co-author	IRM	0.0317 \pm 0.0004	0.0423 \pm 0.0014	-135.0467 \pm 7.3816	0.8901 \pm 0.0162
	LFRM	0.0473 \pm 0.0794	0.0540 \pm 0.0735	-105.2166 \pm 179.5505	0.9348 \pm 0.1667
	MMSB	0.0132 \pm 0.0042	0.0301 \pm 0.0064	-86.2134 \pm 10.1258	0.9524 \pm 0.0215
	iMMM	0.0061 \pm 0.0019	0.0253 \pm 0.0035	-83.4264 \pm 9.4293	0.9574 \pm 0.0155
	cMMSB $^{\pi}$	0.0066 \pm 0.0038	0.0231 \pm 0.0043	-83.4261 \pm 9.4280	0.9569 \pm 0.0159
	cMMSB uv	0.0097 \pm 0.0047	0.0240 \pm 0.0065	-83.4257 \pm 9.4292	0.9581 \pm 0.0153
MIT reality	IRM	0.0627 \pm 0.0002	0.0665 \pm 0.0004	-133.8037 \pm 1.1269	0.8261 \pm 0.0047
	LFRM	0.0397 \pm 0.0017	0.0629 \pm 0.0037	-143.6067 \pm 10.0592	0.8529 \pm 0.0179
	MMSB	0.0263 \pm 0.0105	0.0716 \pm 0.0043	-129.4354 \pm 7.6549	0.8561 \pm 0.0176
	iMMM	0.0297 \pm 0.0055	0.0625 \pm 0.0015	-126.7876 \pm 3.4774	0.8617 \pm 0.0124
	NMDR	0.0386 \pm 0.0040	0.0668 \pm 0.0013	-139.5227 \pm 2.9371	0.8569 \pm 0.0138
	cMMSB $^{\pi}$	0.0246 \pm 0.0016	0.0489 \pm 0.0016	-125.3876 \pm 3.2689	0.8794 \pm 0.0159
Lazega lawfirm	cMMSB uv	0.0283 \pm 0.0035	0.0438 \pm 0.0015	-123.3876 \pm 3.1254	0.8738 \pm 0.0364
	IRM	0.0987 \pm 0.0003	0.1046 \pm 0.0012	-201.7912 \pm 3.3500	0.7056 \pm 0.0167
	LFRM	0.0566 \pm 0.0024	0.1051 \pm 0.0064	-222.5924 \pm 16.1985	0.8170 \pm 0.0197
	MMSB	0.0391 \pm 0.0071	0.0913 \pm 0.0030	-212.1256 \pm 3.2145	0.7989 \pm 0.0102
	iMMM	0.0487 \pm 0.0068	0.1096 \pm 0.0026	-202.7148 \pm 5.3076	0.8074 \pm 0.0141
	NMDR	0.0640 \pm 0.0055	0.1133 \pm 0.0018	-207.7188 \pm 3.4754	0.8285 \pm 0.0114
	cMMSB $^{\pi}$	0.0246 \pm 0.0050	0.1023 \pm 0.0056	-201.0154 \pm 5.2167	0.8273 \pm 0.0148
	cMMSB uv	0.0276 \pm 0.0043	0.1143 \pm 0.0019	-204.0289 \pm 9.5460	0.8215 \pm 0.0167

Incorporating Node Information into BNP Models

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.

Motivation

- The metadata (e.g., the node information in the social network) may affect the relations between nodes (e.g., the friendship).

MMSB and LFRM Models

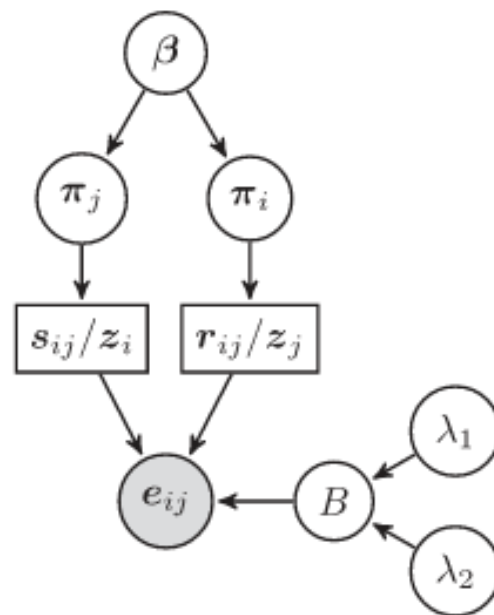
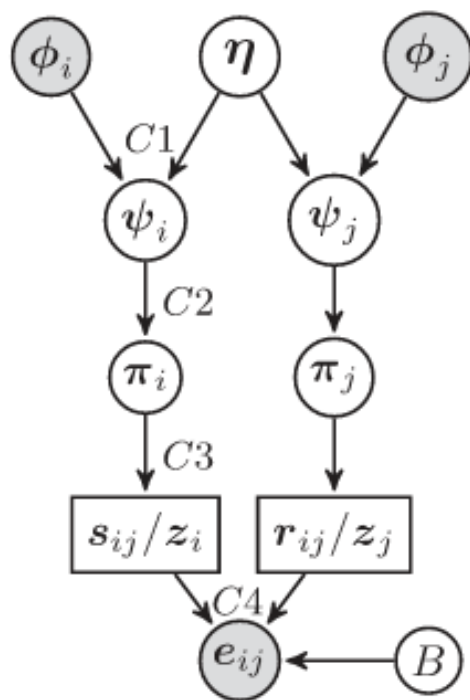


Fig. 1. Graphical model for the MMSB and the LFRM. Here, s_{ij} and r_{ij} in the rectangular nodes represent the latent variable in MMSB, and z_i and z_j are in the LFRM context.

Node-Information Involved Mixed-Membership Model: niMM, niLF



The generative process for the niMM model is defined as follows (w.l.o.g. $\forall i, j = 1, \dots, n, k \in N^+$).

$$C1: \psi_{ik} \sim \text{Beta}(1, \prod_f \eta_{fk}^{\phi_{if}}).$$

$$C2: \pi_{ik} = \psi_{ik} \prod_{l=1}^{k-1} (1 - \psi_{il}).$$

$$C3: s_{ij} \sim \text{Multi}(\pi_i), r_{ij} \sim \text{Multi}(\pi_j).$$

$$C4: e_{ij} \sim \text{Bernoulli}(B_{s_{ij}r_{ij}}).$$

Fig. 2. Generative model for the niMM and niLF models.

Empirical Results

TABLE III
PERFORMANCE ON REAL-WORLD DATA SETS (MEAN \pm STANDARD DEVIATION)

Datasets	Models	Training error	Testing error	Testing log likelihood	AUC
Lazega	IRM	0.0987 \pm 0.0003	0.1046 \pm 0.0012	-201.7912 \pm 3.3500	0.7056 \pm 0.0167
	LFRM	0.0566 \pm 0.0024	0.1051 \pm 0.0064	-222.5924 \pm 16.1985	0.8170 \pm 0.0197
	iMMM	0.0487 \pm 0.0068	0.1096 \pm 0.0026	-202.7148 \pm 5.3076	0.8074 \pm 0.0141
	NMDR	0.0640 \pm 0.0055	0.1133 \pm 0.0018	-207.7188 \pm 3.4754	0.8285 \pm 0.0114
	niMM	0.0334 \pm 0.0056	0.1067 \pm 0.0021	-196.0503 \pm 4.3962	0.8369 \pm 0.0122
	niLF	0.0389 \pm 0.0126	0.1012 \pm 0.0034	-213.5246 \pm 12.3249	0.8123 \pm 0.0135
	cniMM	0.0466 \pm 0.0092	0.1119 \pm 0.0020	-205.0673 \pm 4.5321	0.8314 \pm 0.0119
Reality	IRM	0.0627 \pm 0.0002	0.0665 \pm 0.0004	-133.8037 \pm 1.1269	0.8261 \pm 0.0047
	LFRM	0.0397 \pm 0.0017	0.0629 \pm 0.0037	-143.6067 \pm 10.0592	0.8529 \pm 0.0179
	iMMM	0.0297 \pm 0.0055	0.0625 \pm 0.0015	-126.7876 \pm 3.4774	0.8617 \pm 0.0124
	NMDR	0.0386 \pm 0.0040	0.0668 \pm 0.0013	-139.5227 \pm 2.9371	0.8569 \pm 0.0138
	niMM	0.0269 \pm 0.0047	0.0621 \pm 0.0015	-127.7377 \pm 3.1313	0.8507 \pm 0.0134
	niLF	0.0379 \pm 0.0046	0.0732 \pm 0.0049	-131.0326 \pm 9.4521	0.8645 \pm 0.0139
	cniMM	0.0553 \pm 0.0023	0.0641 \pm 0.0011	-126.9091 \pm 2.6459	0.8597 \pm 0.0099

Motivation

- We extend the existing benchmark models (i.e., MMSB and LFRM) to incorporate the node information. The experimental results seem quite promising while the node information is closely related to the link data.
- Our extension to MMSB retrieves the conjugate property during the MCMC inference, which mixes much faster in the Markov Chain than the previous approaches. Also, we find that in the experiments, our method converges much earlier than the previous one.
- Our model is under the Bayesian nonparametrics setting (achieved through the methods similar to the stick-breaking constructions), which can deal with the problem of an unknown number of communities.

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

Combination of Multiple Sources of Data

- Static

	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

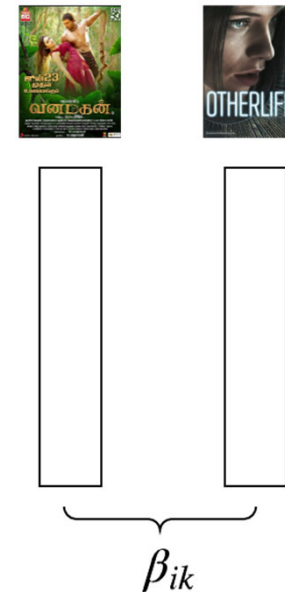
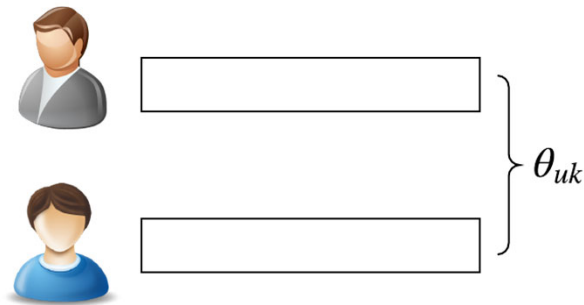
Overview of Statistical Models for Large and Sparse Data

- Poisson Factorization (PF)
 - In matrix factorization, we decompose the rating matrix Y into the vector of the user's preference and item's feature.
 - Similarly, Poisson Factorization (PF) assumes the rating matrix Y follows the Poisson distribution and can be factorized to a vector of K latent preferences for each user and a vector of K latent features for each item.

$$\begin{array}{ccccc} \text{Observations} & & \text{Users weight} & & \text{Items weight} \\ \boxed{Y} & = & \boxed{\theta} & \times & \boxed{\beta} \\ U \times I & & U \times K & & K \times I \end{array}$$

Overview of Statistical Models for Large and Sparse Data

- Matrix Factorization (MF):
 - Users are represented by vectors of latent preferences.
 - Items are represented by vectors of latent features.
 - Latent user preferences and latent item features are estimated based on their own distributions.



Poisson Factorization (PF)

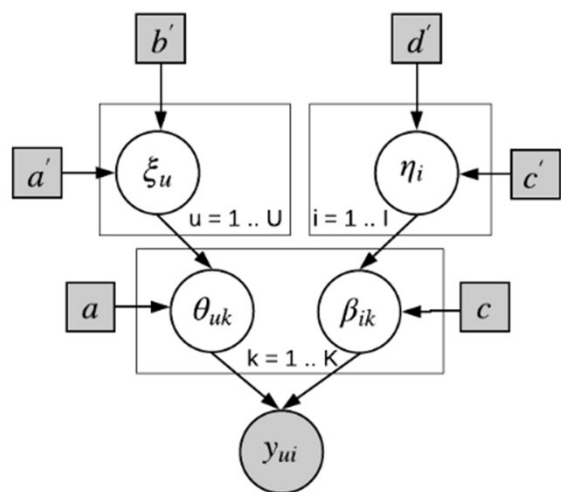


Figure 1.6: Graphical Model of Poisson Factorization (PF).

1. For each user u :
 - (a) Sample latent activity $\xi_u \sim \text{Gamma}(a', a'/b')$.
 - (b) Sample latent preference $\theta_{uk} \sim \text{Gamma}(a, \xi_u)$.
2. For each item i :
 - (a) Sample latent popularity $\eta_i \sim \text{Gamma}(c', c'/d')$.
 - (b) Sample latent attribute $\beta_{ik} \sim \text{Gamma}(c, \eta_i)$.
3. For each user u and item i , sample rating:

$$y_{ui} \sim \text{Poisson}(\sum_k \theta_{uk} \beta_{ik}).$$

Overview of Statistical Models for Large and Sparse Data

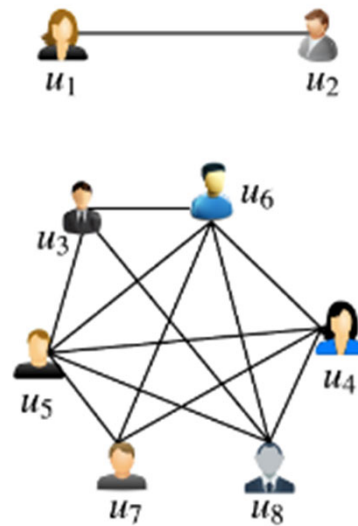
- Properties of PF:
 - PF captures sparse factors. It is based on the way of PF compute only on the **non-missing** data. Since the real-world rating data is often sparse (e.g., Netflix data has more than 98% missing data), this makes PF strong.
 - PF models the **long-tail** of users and items. It is also fitted with the real-world data in which the majority users tend to rate for the minority of items.
 - PF downweights the effect of **zeros**. As there are many missing values (i.e., zero value), this property is critical in the real-world situation.
 - Fast inference with sparse matrices. Since Bayesian models strongly depend on the inference methods, we need to have a good method to boost the computational time of PF.

Combination of Multiple Sources of Data

- Static

	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



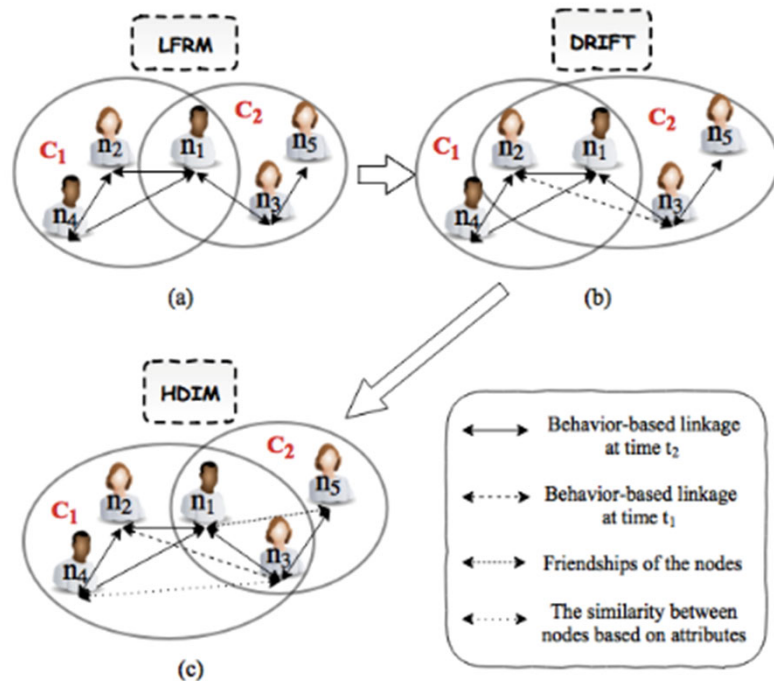
(b) User friendship

	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

Combination of Multiple Sources of Data

- Dynamic



		Attributes				Friendships	Behavior-based Linkages
		Age	Gender	Location	Education		
t_1	n_1	34	Male	SYD	Master	$\{n_2, n_4\}$	$\{n_2, n_3\}$
	n_2	35	Female	SYD	Master	$\{n_1, n_4\}$	$\{n_1, n_3\}$
	n_3	24	Female	NYC	Bachelor	$\{n_4\}$	$\{n_1, n_2\}$
	n_4	25	Male	SYD	Bachelor	$\{n_1, n_2, n_3\}$	$\{\}$
t_2	n_1	34	Male	SYD	Master	$\{n_2, n_4, n_5\}$	$\{n_2, n_3, n_4\}$
	n_2	35	Female	SYD	Master	$\{n_1, n_3, n_4\}$	$\{n_1, n_4\}$
	n_3	24	Female	NYC	Bachelor	$\{n_2, n_4\}$	$\{n_1, n_5\}$
	n_4	25	Male	NYC	Bachelor	$\{n_1, n_2, n_3\}$	$\{n_1, n_2\}$
	n_5	24	Female	NYC	Bachelor	$\{n_1\}$	$\{n_3\}$

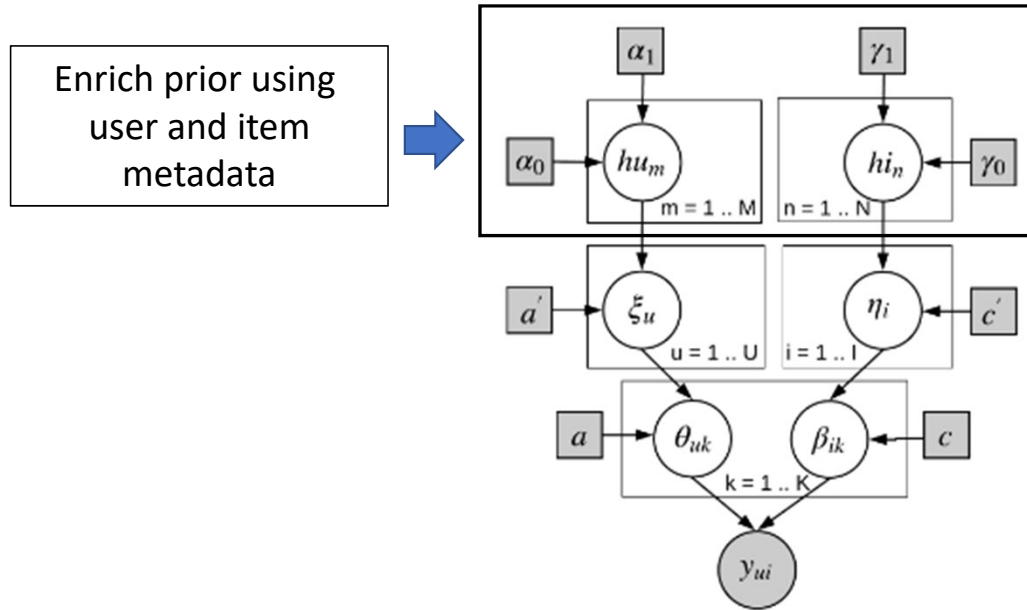
Statistical Learning of Large-scale, Sparse and Multi-source Data

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)



(a) MPF

Metadata-integrated Poisson Factorization (MPF)

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

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

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

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

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

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

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

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

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

(a) Sample user's latent preference:

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

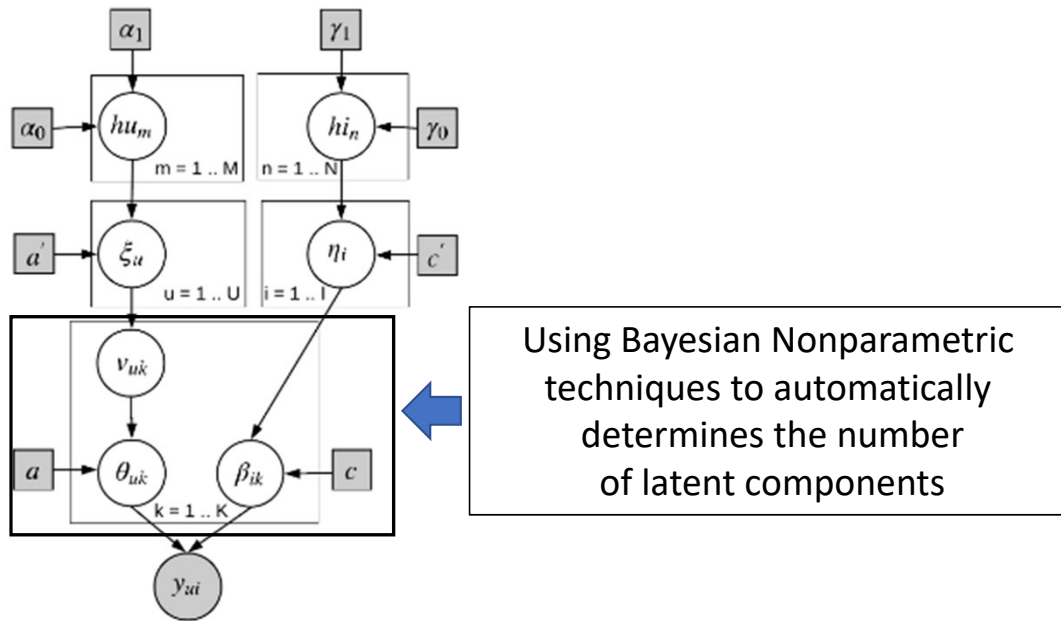
(b) Sample item's latent feature:

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

(6) Sample rating:

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

Metadata-integrated Infinite Poisson Factorization (MIPF)



(b) MIPF

Metadata-integrated Infinite Poisson Factorization (MIPF)

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

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

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

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

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

(a) Draw the user's latent behavior:

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

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

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

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

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

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

(a) Draw the item's latent attractiveness:

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

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

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

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

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

Inference

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

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

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*.

Inference

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

$$q(hu, hi, v, \beta, \xi, \eta, z) = \prod_m q(hu_m | \zeta_m) \prod_n q(hi_n | \rho_n) \\ \prod_{k=1}^{\infty} \prod_u q(v_{uk} | \sigma_{uk}) \prod_{k=1}^{\infty} \prod_i q(\beta_{ik} | \mu_{ik}) \prod_u q(\xi_u | \kappa_u) \\ \prod_i q(\eta_i | \tau_i) \prod_{k=1}^{\infty} \prod_{u,i} q(z_{ui,k} | \phi_{ui,k})$$

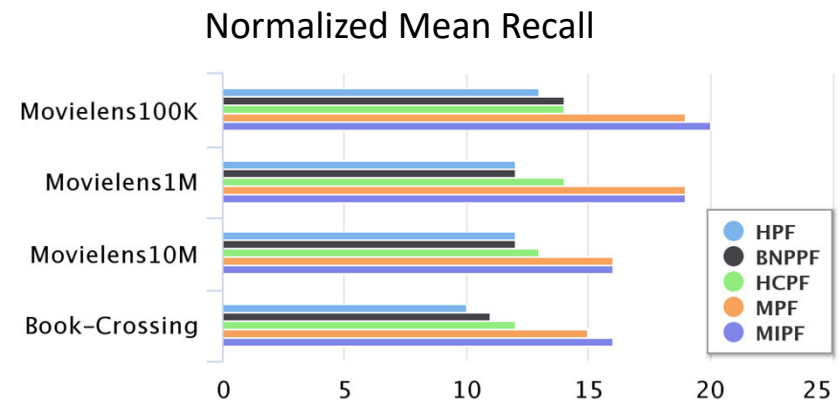
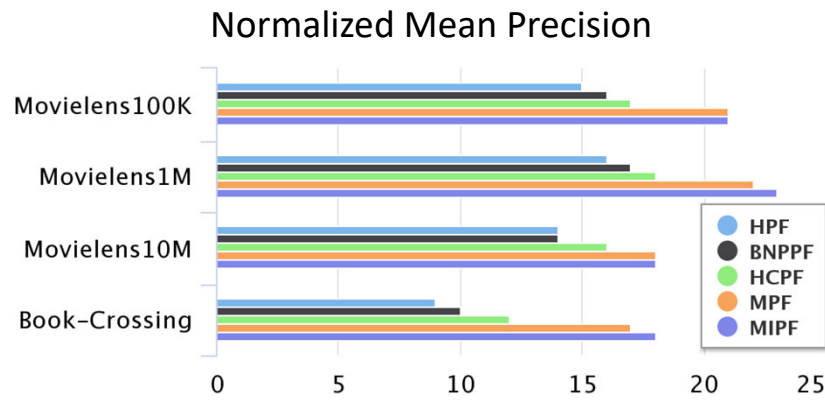
Algorithm 1 Variational Inference for MPF

- 1: Initialize the variational parameters $\{\zeta, \rho, \nu, \mu, \kappa, \tau, \phi\}$.
 - 2: Set the number of components K .
 - 3: Sample *shape* of user's latent behavior, and *shape* of item's latent attractiveness, as in Eqs. (22) and (24).
 - 4: Sample *shape* of the weight of user's attribute (in metadata), and *shape* of the weight of item's attribute (in metadata), as in Eqs. (18) and (20).
 - 5: **repeat**
 - 6: **for** each rating of user u to item i that $y_{ui} \neq 0$ **do**
 - 7: Update the multinomial as in Eq. (26).
 - 8: **end for**
 - 9: **for** each user **do**
 - 10: Update the latent preference as in Eqs. (27) and (28)
 - 11: Update *rate* of latent behavior as in Eq. (23).
 - 12: **for** each user attribute in metadata **do**
 - 13: Update *rate* of the weight as in Eq. (19)
 - 14: **end for**
 - 15: **end for**
 - 16: **for** each item **do**
 - 17: Update the latent feature as in Eqs. (29) and (30).
 - 18: Update *rate* of latent attractiveness as in Eq. (25).
 - 19: **for** each item attribute **do**
 - 20: Update *rate* of the weight as in Eq. (21).
 - 21: **end for**
 - 22: **end for**
 - 23: **until** convergence
-

Experiments

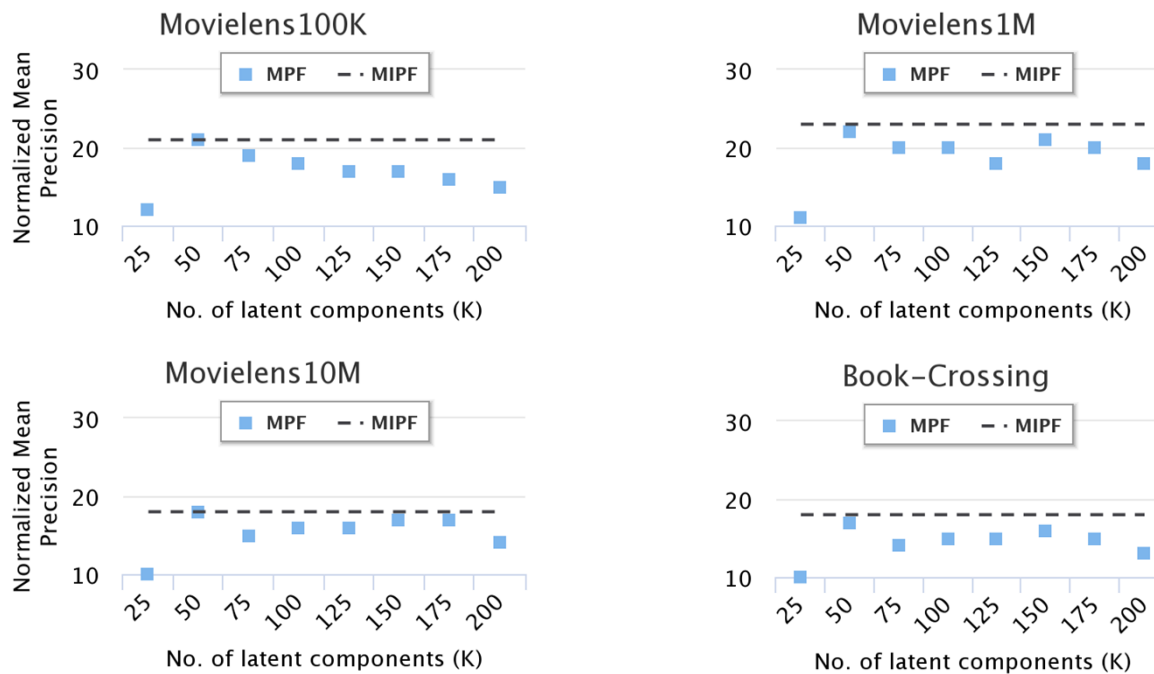
- Datasets:
 - (1) Movielens100K, Movielens1M and Movielens10M [Harper and Konstan, 2016].
 - (2) Book-Crossing [Ziegler et al., 2005].
- Baseline methods:
 - **HPF** [Gopalan et al., 2015] as it outperforms many baselines in MF including NMP, LDA and PMF.
 - **Bayesian Nonparametric PF (BNPPF)** [Gopalan et al., 2014a].
 - The latest PF: **Hierarchical Compound PF (HCPF)** [Basbug and Engelhardt, 2016].

How do MPF/MIPF significantly outperform other PF models?



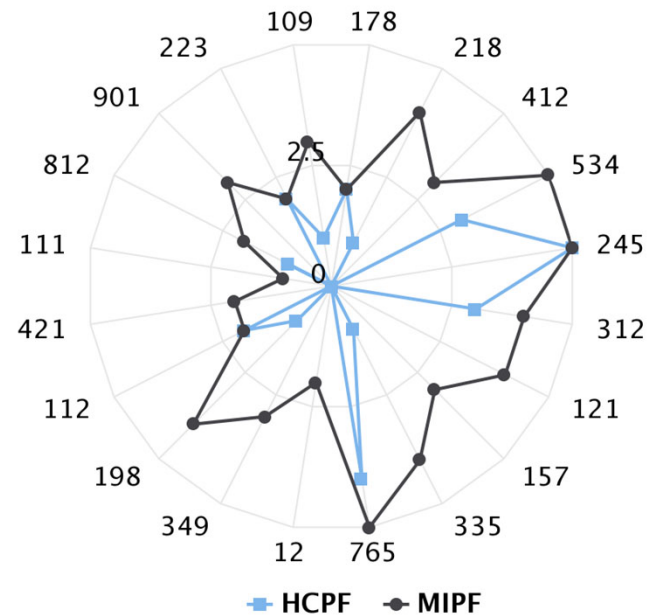
Top-20 Recommendation Compared with baselines

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.

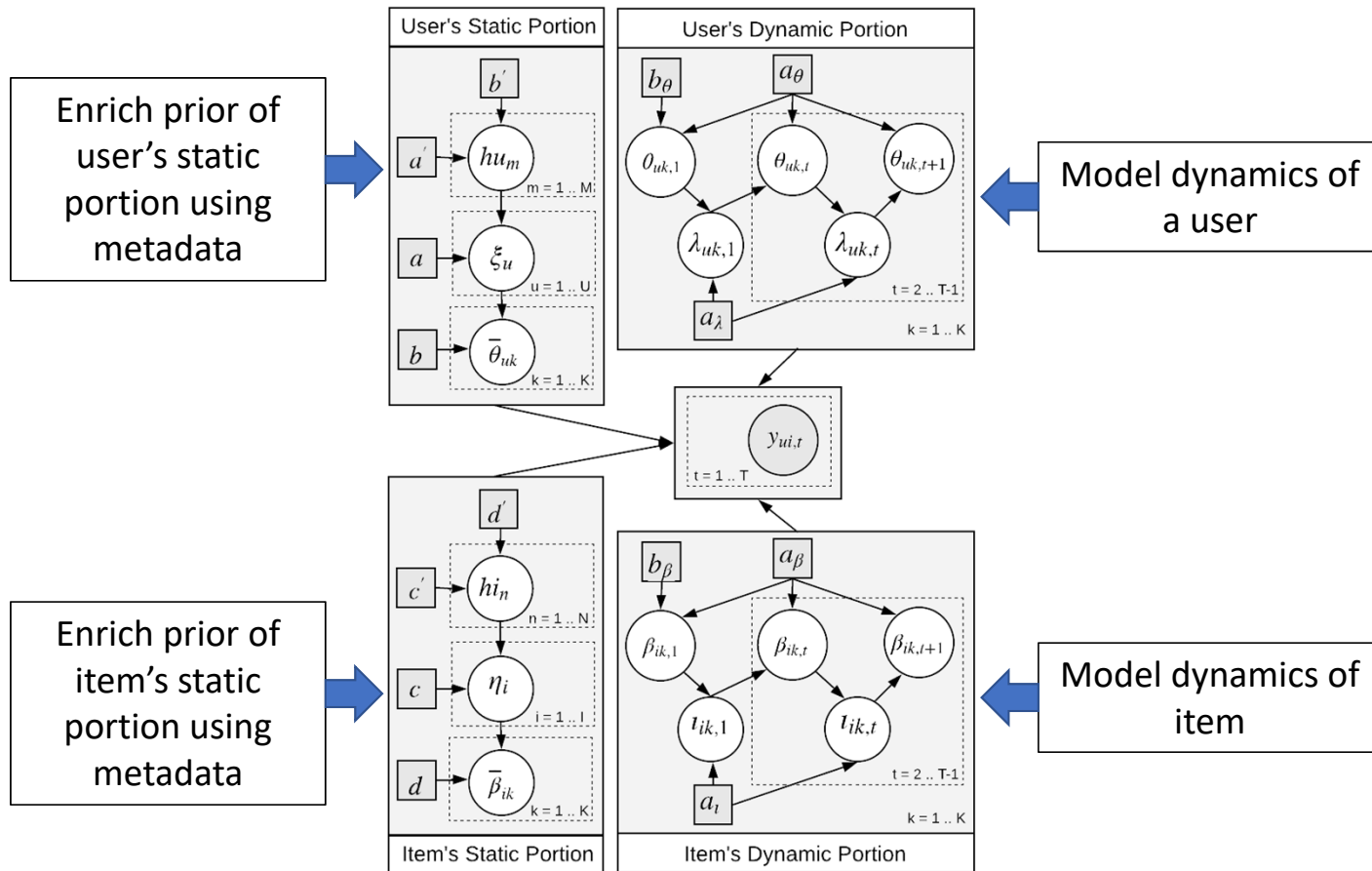
Statistical Learning of Large-scale, Sparse and Multi-source Data

Trong Dinh Thac Do and Longbing Cao. Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence, NIPS2018

Motivation

- Deal with large and sparse data.
- Solve the problem of sparse users/items and cold-start.
- Capture the dynamics of data.

Gamma-Poisson Dynamic Matrix Factorization model incorporated with metadata influence (mGDMMF)



Gamma-Poisson Dynamic Matrix Factorization model incorporated with metadata influence (mGD MF)

1. Metadata Integration:

(a) For each user:

- i. Draw the weight of m^{th} attribute in user metadata $hu_m \sim \text{Gamma}(a', b')$
- ii. Draw latent user preference $\xi_u \sim \text{Gamma}(a, \prod_{m=1}^M hu_m^{f_{u,m}})$
- iii. Draw global static factor $\bar{\theta}_{uk} \sim \text{Gamma}(b, \xi_u)$

(b) For each item:

- i. Draw the weight of n^{th} attribute in item metadata $hi_n \sim \text{Gamma}(c', d')$
- ii. Draw latent item attractiveness $\eta_i \sim \text{Gamma}(c, \prod_{n=1}^N hi_n^{f_{i,n}})$
- iii. Draw global static factor $\bar{\beta}_{ik} \sim \text{Gamma}(d, \eta_i)$

2. Dynamic Modeling:

(a) For each user:

- i. Draw initialized state of local dynamic factor $\theta_{uk,1} \sim \text{Gamma}(a_\theta, a_\theta b_\theta)$
- ii. For each time slice $t > 1$:
 - A. Draw auxiliary variable $\lambda_{uk,t-1} \sim \text{Gamma}(a_\lambda, a_\lambda \theta_{uk,t-1})$
 - B. Draw local dynamic factor $\theta_{uk,t} \sim \text{Gamma}(a_\theta, a_\theta \lambda_{uk,t-1})$

(b) For each item:

- i. Draw initialized state of local dynamic factor $\beta_{ik,1} \sim \text{Gamma}(a_\beta, a_\beta b_\beta)$
- ii. For each time slice $t > 1$:
 - A. Draw auxiliary variable $\iota_{ik,t-1} \sim \text{Gamma}(a_\iota, a_\iota \beta_{ik,t-1})$
 - B. Draw local dynamic factor $\beta_{ik,t} \sim \text{Gamma}(a_\beta, a_\beta \iota_{ik,t-1})$

3. For each rating:

- (a) Draw $y_{ui,t} \sim \text{Poisson}(\sum_k (\theta_{uk,t} + \bar{\theta}_{uk})(\beta_{ik,t} + \bar{\beta}_{ik}))$

inference

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

$$\begin{aligned}
 q(hu, hi, \xi, \eta, \bar{\theta}, \bar{\beta}, \lambda, \iota, \theta, \beta, z) = & \prod_m q(hu_m | \zeta_m) \prod_n q(hi_n | \rho_n) \prod_u q(\xi_u | \kappa_u) \prod_i q(\eta_i | \tau_i) \\
 & \prod_{u,k} q(\bar{\theta}_{uk} | \bar{\nu}_{uk}) \prod_{i,k} q(\bar{\beta}_{ik} | \bar{\mu}_{ik}) \prod_{u,k,t} q(\theta_{uk,t} | \nu_{uk,t}) \prod_{i,k,t} q(\beta_{ik,t} | \mu_{ik,t}) \\
 & \prod_{u,k,t} q(\lambda_{uk,t} | \gamma_{uk,t}) \prod_{i,k,t} q(\iota_{ik,t} | \omega_{ik,t}) \prod_{u,i,t,k} q(z_{ui,t,k} | \phi_{ui,t,k})
 \end{aligned} \tag{3}$$

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

inference

Table 1: Latent Variables, Type, Variational Variables and Variational Update for Users. Similar variables for items (i.e., $h_{i,n}$, η_i , $\bar{\beta}_{ik}$, β_{ik} , $\iota_{ik,t}$) can be found in the supplementary. \aleph_m is the number of users having the m^{th} attribute, K is the number of latent components, and $\Psi(\cdot)$ is the *digamma* function. The Gamma distribution is parameterized by *shape* (*shp*) and *rate* (*rte*).

Latent Variable	Type	Variational Variable	Variational Update
hu_m	Gamma	$\zeta_m^{shp}, \zeta_m^{rte}$	$a' + \aleph_m a, b' + \sum_u \frac{\kappa_u^{shp}}{\kappa_u^{rte}}$
ξ_u	Gamma	$\kappa_u^{shp}, \kappa_u^{rte}$	$a + Kb, \prod_{m=1}^M \left(\frac{\zeta_m^{shp}}{\zeta_m^{rte}} \right)^{fu_{u,m}} + \sum_k \frac{\bar{\nu}_{uk}^{shp}}{\bar{\nu}_{uk}^{rte}}$
$z_{ui,t,k}$	Mult	$\phi_{ui,t,k}$	$(\exp\{\Psi(\nu_{uk,t}^{shp}) - \log(\nu_{uk,t}^{rte})\} + \exp\{\Psi(\bar{\nu}_{uk}^{shp}) - \log(\bar{\nu}_{uk}^{rte})\})$ $\ast (\exp\{\Psi(\mu_{ik,t}^{shp}) - \log(\mu_{ik,t}^{rte})\} + \exp\{\Psi(\bar{\mu}_{ik}^{shp}) - \log(\bar{\mu}_{ik}^{rte})\})$
$\bar{\theta}_{uk}$	Gamma	$\bar{\nu}_{uk}^{shp}, \bar{\nu}_{uk}^{rte}$	$b + \sum_{i,t} y_{ui,t} \phi_{ui,t,k}, \frac{\kappa_u^{shp}}{\kappa_u^{rte}} + \sum_i \left(\frac{\bar{\mu}_{ik}^{shp}}{\bar{\mu}_{ik}^{rte}} + \sum_t \frac{\mu_{ik,t}^{shp}}{\mu_{ik,t}^{rte}} \right)$
$\theta_{uk,t}$	Gamma	$\nu_{uk,t}^{shp}$ $\nu_{uk,1}^{rte}$ $\nu_{uk,t,(t>1)}^{rte}$	$a_\theta + a_\lambda + \sum_i y_{ui,t} \phi_{ui,t,k}$ $a_\theta b_\theta + a_\lambda \frac{\gamma_{uk,1}^{shp}}{\gamma_{uk,1}^{rte}} + \sum_i \left(\frac{\bar{\mu}_{ik}^{shp}}{\bar{\mu}_{ik}^{rte}} + \frac{\mu_{ik,1}^{shp}}{\mu_{ik,1}^{rte}} \right)$ $a_\theta \frac{\gamma_{uk,t-1}^{shp}}{\gamma_{uk,t-1}^{rte}} + a_\lambda \frac{\gamma_{uk,t}^{shp}}{\gamma_{uk,t}^{rte}} + \sum_i \left(\frac{\bar{\mu}_{ik}^{shp}}{\bar{\mu}_{ik}^{rte}} + \frac{\mu_{ik,t}^{shp}}{\mu_{ik,t}^{rte}} \right)$
$\lambda_{uk,t}$	Gamma	$\gamma_{uk,t}^{shp}, \gamma_{uk,t}^{rte}$	$a_\lambda + a_\theta, a_\lambda \frac{\nu_{uk,t}^{shp}}{\nu_{uk,t}^{rte}} + a_\theta \frac{\nu_{uk,t+1}^{shp}}{\nu_{uk,t+1}^{rte}}$

SVI

Algorithm 1 SVI for mGDMF

Initialize $\{\zeta, \rho, \kappa, \tau, \bar{\nu}, \bar{\mu}, \nu, \mu, \gamma, \omega, \phi\}$.

Set K : # latent components, U : # users, I : # items, $iter_0$ and ϵ .

repeat

for each time slice $t = 1 \dots T$ **do**

 Sample a rating $y_{ui,t}$ uniformly from the dataset.

 Update the local variational parameter of multivariate parameter ϕ .

 Update all intermediate variational parameters similar to Eq. (4).

 Update all global variational parameters similar to Eq. (5).

 Update the learning rates $iter$.

end for

until convergence

Experiments

- Datasets:
 - (1) Netflix-Time, Netflix-Full [Li et al., 2011].
 - (2) Yelp-Active [Jerfel et al., 2017].
 - (3) LFM-Tracks, LFM-Bands [Ò. Celma Herrada, 2009].
- Baseline methods:
 - Static:
 - **HPF** [Gopalan et al., 2015], **HCPF** [Basbug and Engelhard, 2016] as it outperforms many baselines in MF including NMP, LDA and PMF.
 - PF-last and HCPF-last are trained by using the last time slice in the training set as the observations.
 - HPF-all and HCPF-all are trained on all training ratings.
 - Dynamic:
 - **dPF** [Charlin et al., 2016] and **DCPF** [Jerfel et al., 2017].
 - dPF was shown to outperform state-of-the-art dynamic collaborative filtering algorithms, specifically, BPTF and TimeSVD++.

Effect of metadata and dynamic data modeling

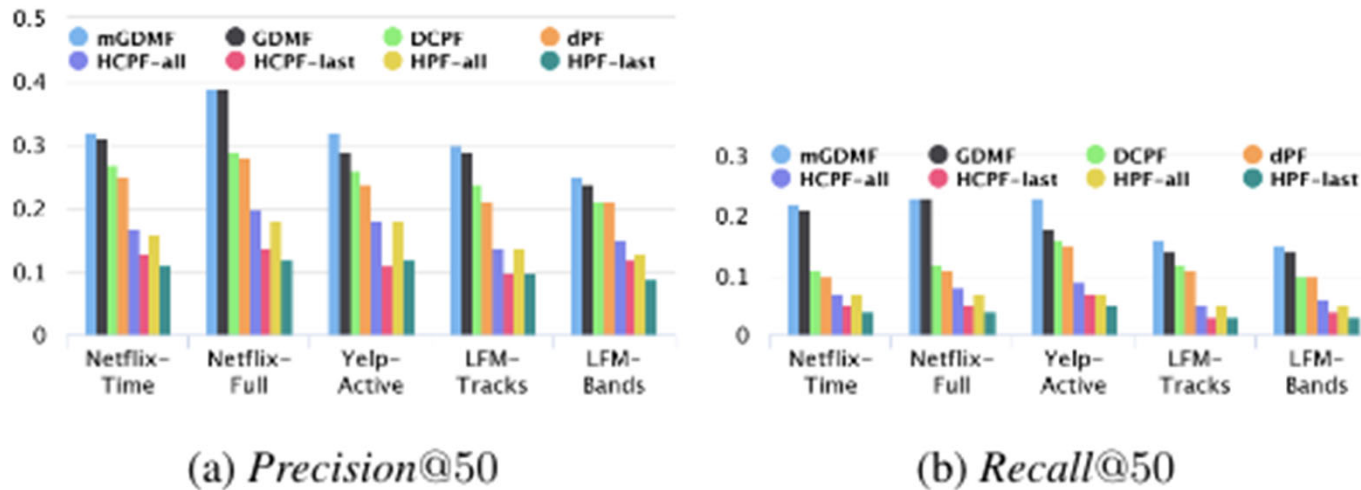


Figure 1: Top-50 Recommendations Compared with Baselines.

Effect of metadata and dynamic data modeling

Table 2: Predictive Performance on Five Datasets w.r.t. NDCG and AUC.

	Netflix-Time		Netflix-Full		Yelp-Active		LFM-Tracks		LFM-Bands	
	NDCG	AUC	NDCG	AUC	NDCG	AUC	NDCG	AUC	NDCG	AUC
mGDMF	0.389	0.9145	0.403	0.9321	0.494	0.8650	0.310	0.8245	0.367	0.8217
GDMF	0.367	0.9121	0.398	0.9320	0.416	0.8512	0.275	0.8101	0.354	0.8139
DCPF	0.293	0.9023	0.315	0.8991	0.357	0.8418	0.231	0.8098	0.275	0.8011
dPF	0.257	0.9012	0.301	0.8901	0.332	0.8321	0.210	0.8019	0.298	0.8122
HCPF-all	0.241	0.8012	0.245	0.8370	0.243	0.8032	0.209	0.7010	0.213	0.7121
HCPF-last	0.183	0.7423	0.201	0.7600	0.172	0.7312	0.132	0.5893	0.160	0.6101
HPF-all	0.231	0.8035	0.250	0.8124	0.248	0.8130	0.179	0.7084	0.184	0.7013
HPF-last	0.162	0.7213	0.198	0.7540	0.145	0.6810	0.143	0.6050	0.141	0.5982
$\delta_{min}(\%)$	32.76	1.35	27.94	3.67	38.38	2.76	34.20	1.82	23.15	1.70
$\delta_{max}(\%)$	140.12	26.78	103.54	23.62	240.69	27.12	134.85	44.83	160.28	37.36

Effect of handling sparse users/items and the 'cold-start' problem

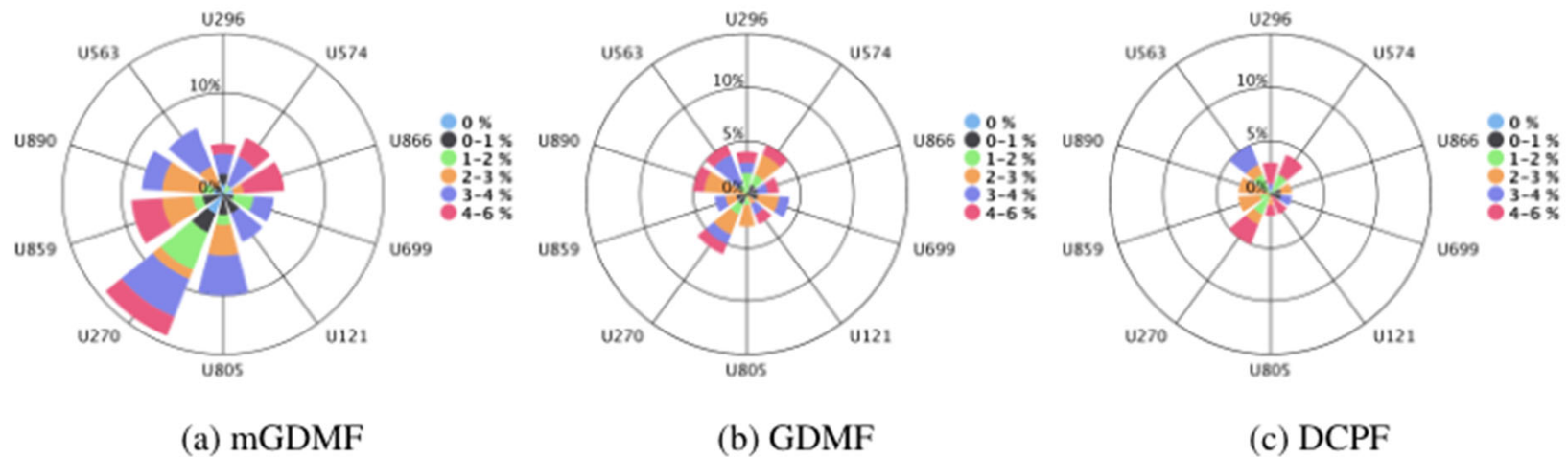


Figure 2: Percentage (%) of Sparse Items Recommended Precisely for 10 Users by mGDMF, GDMF and DCPF.

Case study of mGD MF-based recommendation

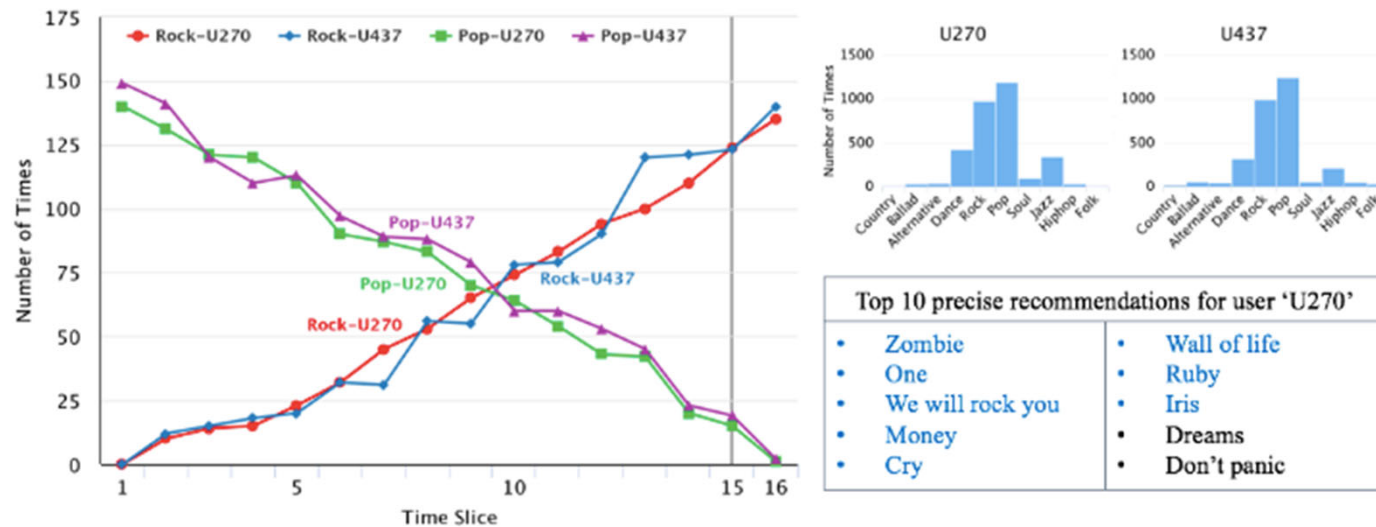


Figure 3: Analysis on two users 'U270' and 'U437' with the same metadata in Last.fm. The number of times that users listened to two 'rock' and 'pop' tracks with 16 time slices is shown on the left. The distribution of the number of times that U270 and U437 listened to top 10 'rock' and 'pop' tracks and the top10 precise recommendations by mGD MF are shown on the right.

Contributions

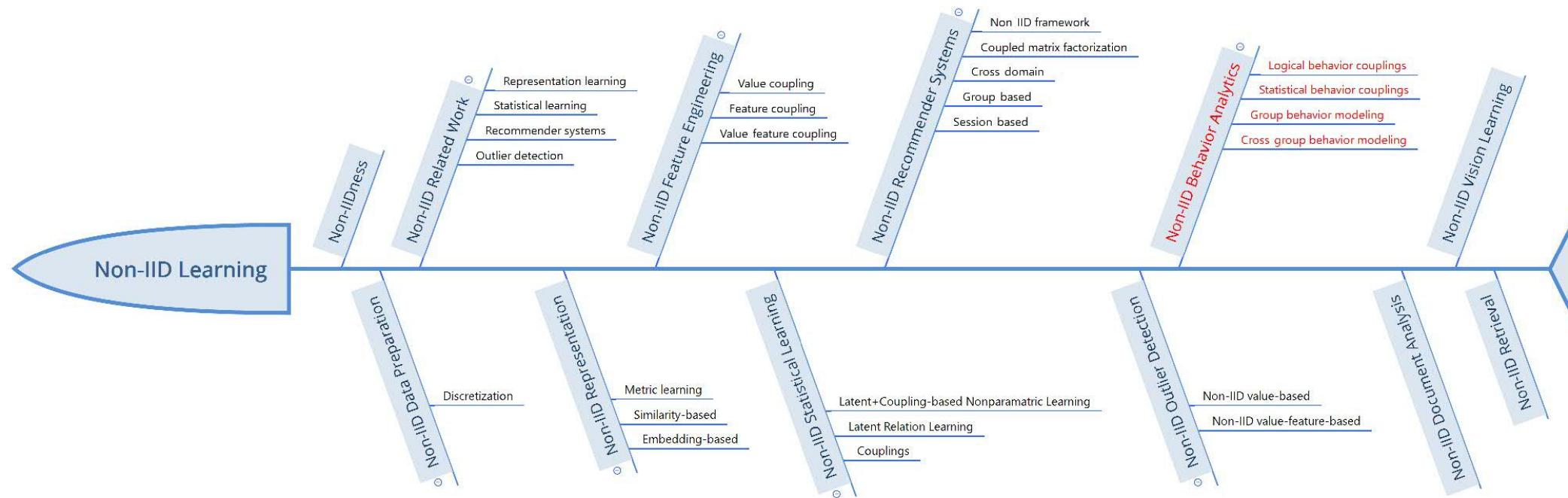
- A factorization model that uses **Gamma-Poisson** structure to model massive, sparse and dynamic data.
- A conjugate **Gamma-Gamma** of integrating the observable user/item metadata (e.g., 'age' of a user and 'genre' of a movie) with user/item latent variables to model user/item rating sparsity.
- A conjugate **Gamma-Markov chains** to model user/item latent variables that change smoothly over time.
- An efficient **stochastic variational inference** for massive, sparse and dynamic data.

Non-IID Behavior Analytics

More at KDD2018 Tutorial on Behavior Analytics

www.datasciences.org

Non-IID behavior analytics



Behavior Model

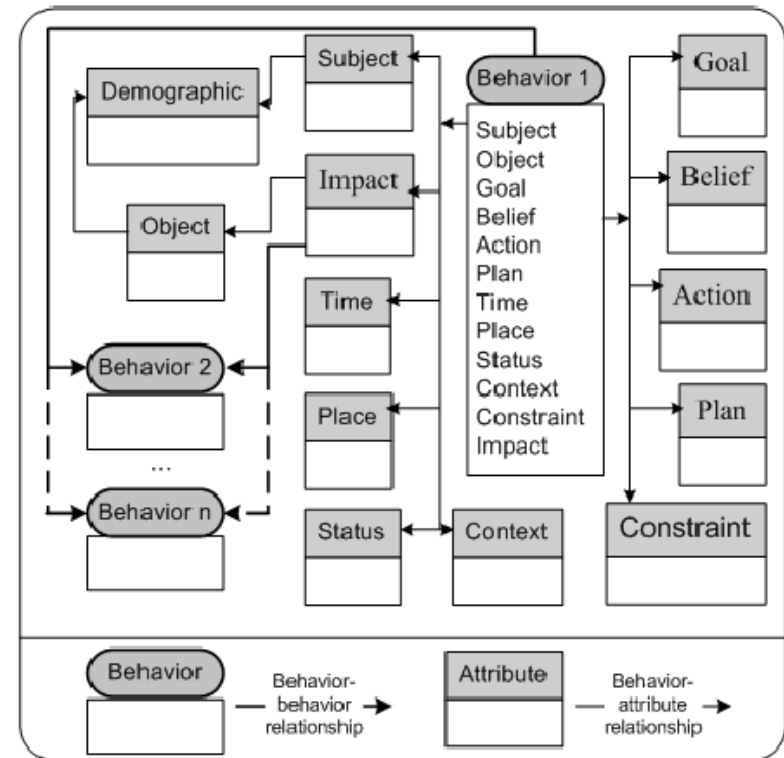
Longbing Cao, [In-depth Behavior Understanding and Use: the Behavior Informatics Approach](#), Information Science, 180(17); 3067-3085, 2010.

Examples of Coupled Objects and Behaviors



An Abstract Behavior Model

- 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



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**

$$\vec{\Gamma} = \{\vec{\gamma}_1, \vec{\gamma}_2, \dots, \vec{\gamma}_n\}$$

- **Behavior Coupling Relationships**

- ✓ Logic/semantic behavior couplings

- ✓ Statistical/Probabilistic behavior couplings

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).

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

$$FM(\mathbb{B}) = \begin{array}{c} \begin{array}{c} \text{intra-coupling} \\ \left(\begin{array}{c|cccc} \mathcal{O}_{11} & \mathcal{O}_{12} & \dots & \mathcal{O}_{1J_{max}} \\ \mathcal{O}_{21} & \mathcal{O}_{22} & \dots & \mathcal{O}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_{I1} & \mathcal{O}_{I2} & \dots & \mathcal{O}_{IJ_{max}} \end{array} \right) \end{array} \\ \text{inter-coupling} \end{array}$$

An actor \mathcal{A}_i undertakes J_i operations $\{\mathcal{O}_{i1}, \mathcal{O}_{i2}, \dots, \mathcal{O}_{iJ_i}\}$
 I actors: $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_I\}$

Intra-Coupling

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

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

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

where $\sum_{j=1}^{J_{max}} \odot$ means the subsequent behavior of \mathbb{B}_i is \mathbb{B}_{ii} 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 \mathcal{A}_i 's behaviors \mathbb{B}_{ij} ($1 \leq i \leq I$) are inter-coupled with each other in terms of coupling function $\eta_i(\mathbb{B})$,

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

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.

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

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 $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$

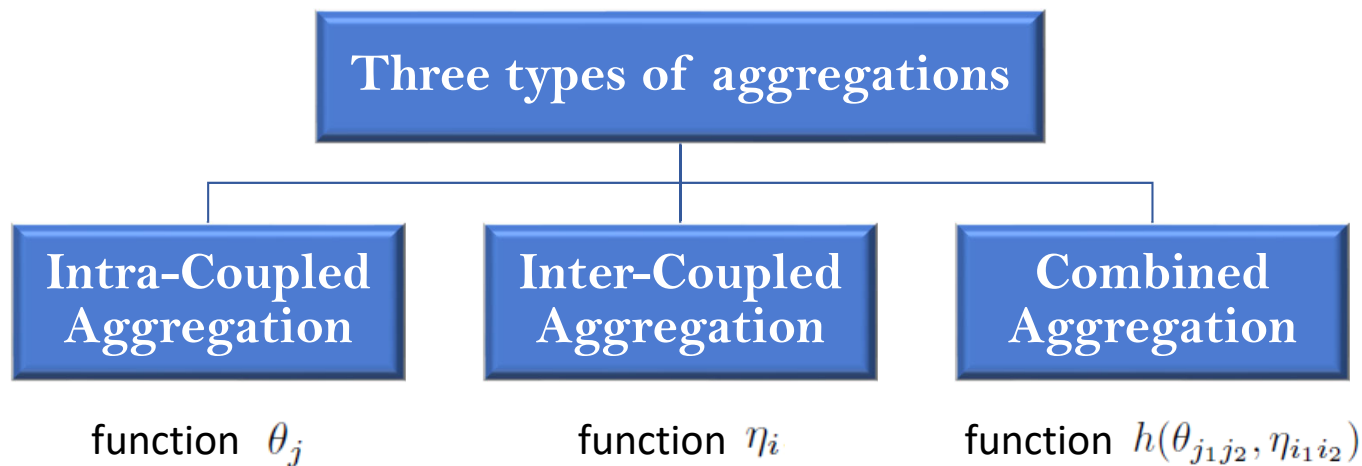
$$\mathbb{B}_c = (\mathbb{B}_{i_1 j_1}^\theta)^\eta * (\mathbb{B}_{i_2 j_2}^\theta)^\eta ::= \mathbb{B}_{ij}(\mathcal{A}, \mathcal{O}, \mathcal{C}) | \sum_{i_1, i_2=1}^I \sum_{j_1, j_2=1}^{J_{max}} h(\theta_{j_1 j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B})) \odot (\mathbb{B}_{i_1 j_1} \mathbb{B}_{i_2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

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.

Behavior Aggregator

We conduct behavior aggregations to interpret the interactions of intra-coupled and inter-coupled behaviors. The outcomes of the behavior aggregations form the basis of behavior verification.



Coupled Behavior Analysis

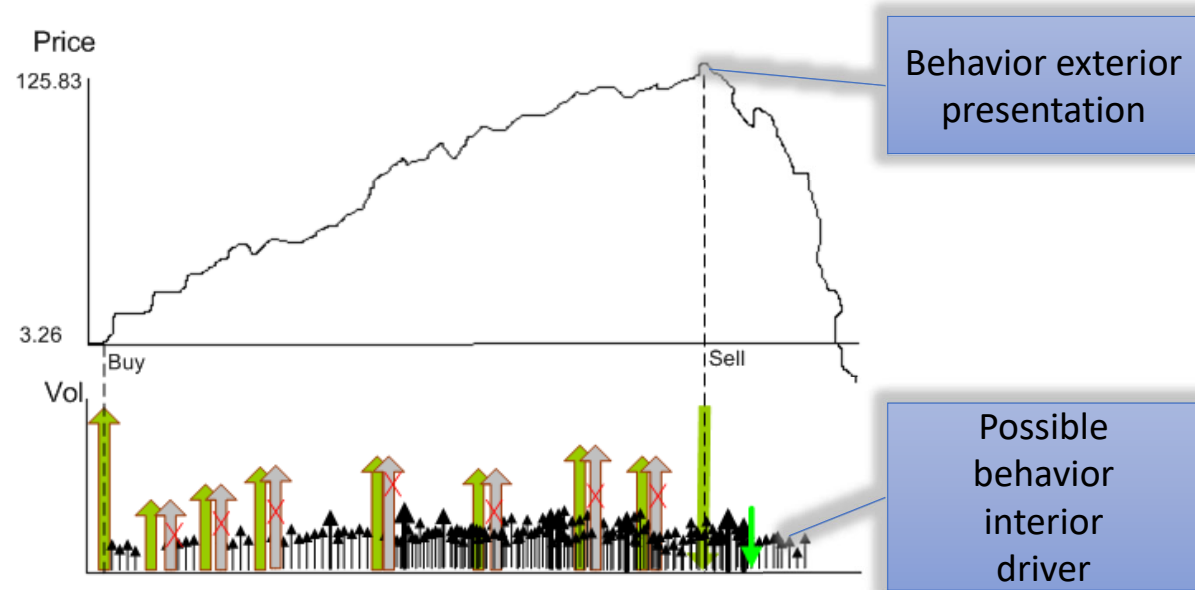
Theorem 1. *(Coupled Behavior Analysis (CBA)) The analysis of coupled behaviors (CBA Problem for short) is to build the objective function $g(\cdot)$ under the condition that behaviors are coupled with each other by coupling function $f(\cdot)$, and satisfy the following conditions.*

$$f(\cdot) ::= f(\theta(\cdot), \eta(\cdot)), \quad (9)$$

$$g(\cdot) | (f(\cdot) \geq f_0) \geq g_0 \quad (10)$$

Example of Group Behavior Analysis

- Short-term manipulation behaviors as cause



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

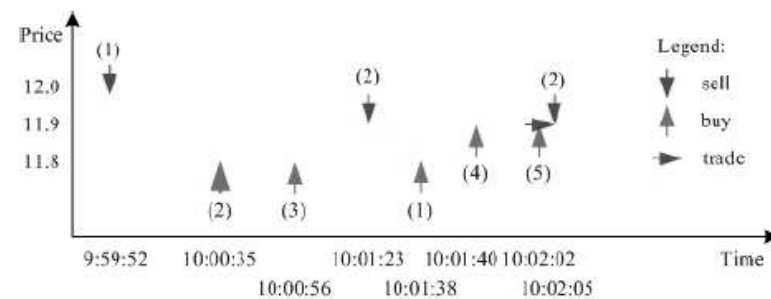


Fig. 1. Coupled Trading Behaviors

CHMM Based Coupled Sequence Modeling

- Coupled behavior sequences

- Multiple sequences

$$\Phi_1 = \{\phi_{11}, \dots, \phi_{1T}\}$$

$$\Phi_2 = \{\phi_{21}, \dots, \phi_{2F}\}$$

$$\Phi_C = \{\phi_{C1}, \dots, \phi_{CG}\}$$

- Coupling relationship

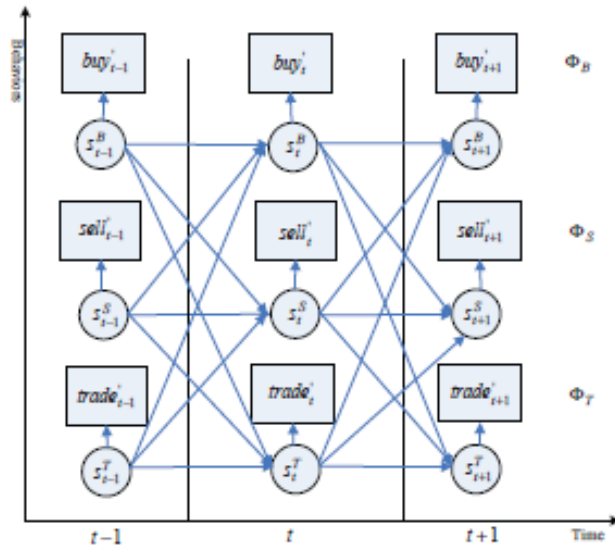
$$R_{ij}(\Phi_i, \Phi_j)$$

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

- Behavior properties

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

CBA – CHMM



(b) The Structure of the CHMM

$$CBA \text{ problem} \rightarrow CHMM \text{ model} \quad (15)$$

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

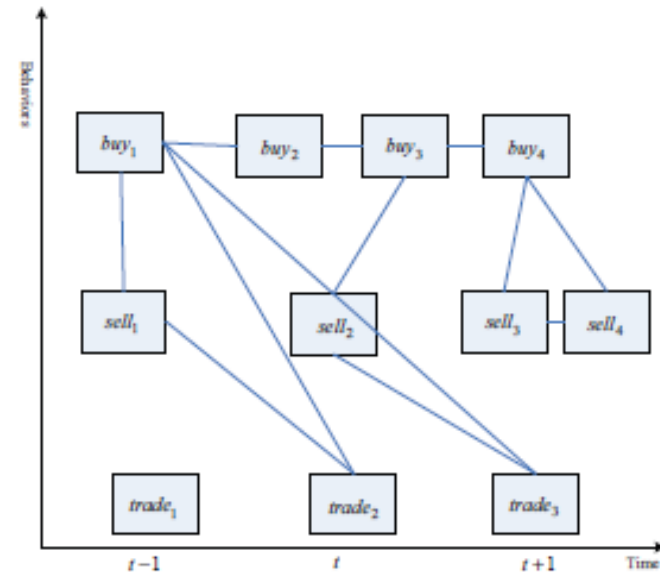
$$M(\Phi(\mathbb{B}_c)) | \phi_{ik}([p_{ij}]_1, \dots, [p_{ij}]_K) \rightarrow Y \quad (17)$$

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

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

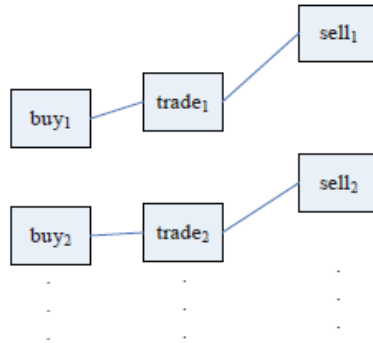
Graph-based Coupled Behavior Presentation

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



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

CBA - Conditional Probability Distribution



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

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

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

$$CBA \text{ problem} \rightarrow SRL \text{ Modeling} \quad (5)$$

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

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

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

Empirical Results

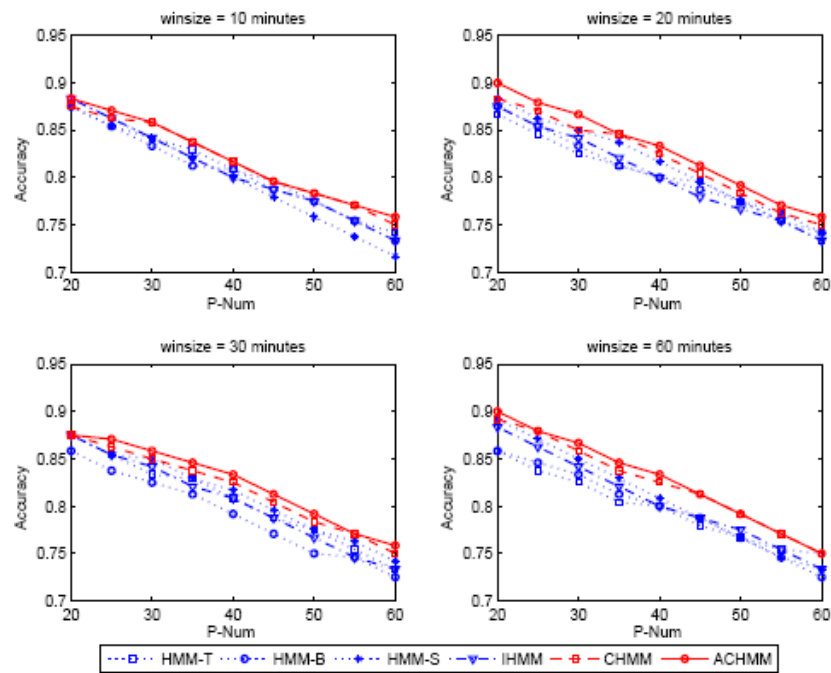


Figure 4: Accuracy of Six Models

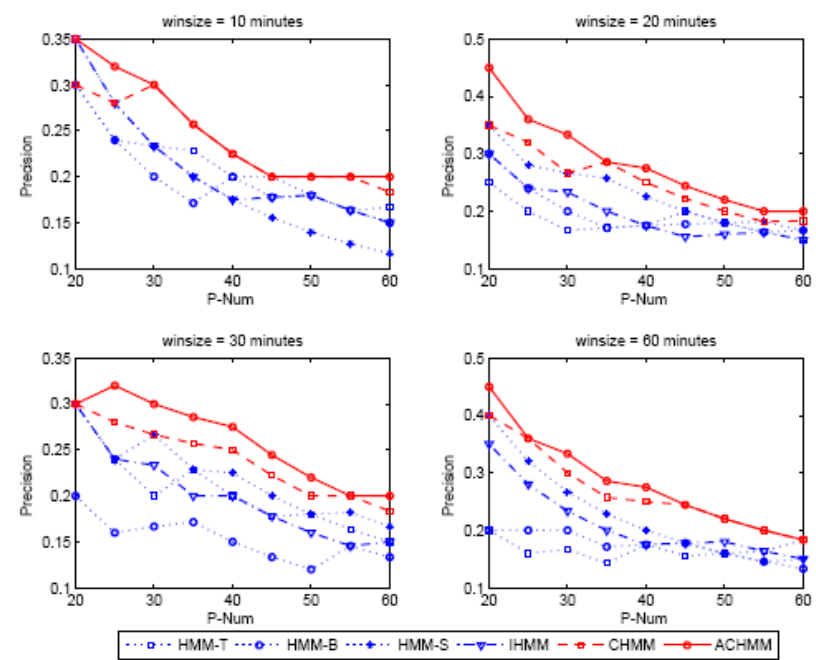


Figure 5: Precision of Six Models

Empirical Results

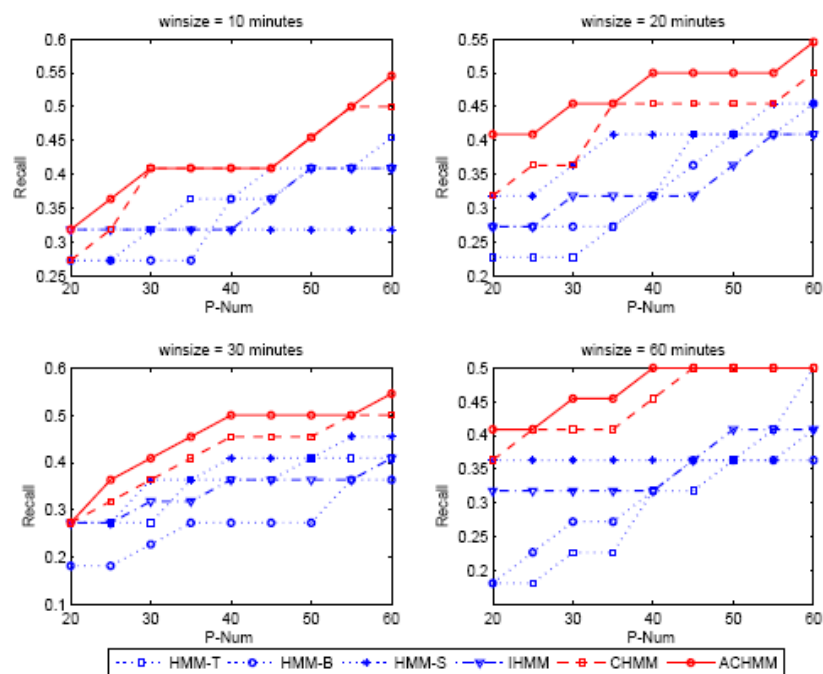


Figure 6: Recall of Six Models

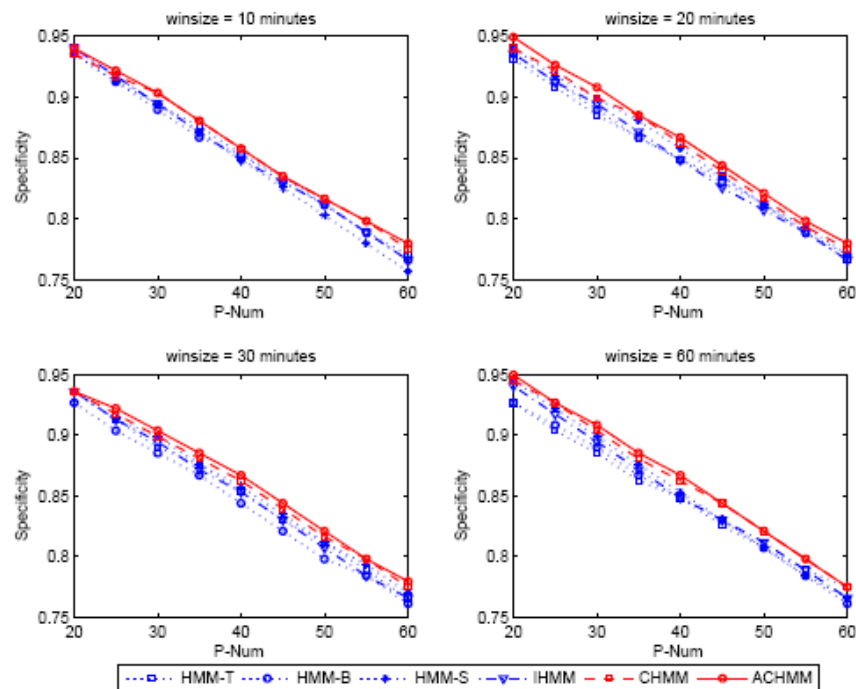


Figure 7: Specificity of Six Models

Empirical Results - Business Performance

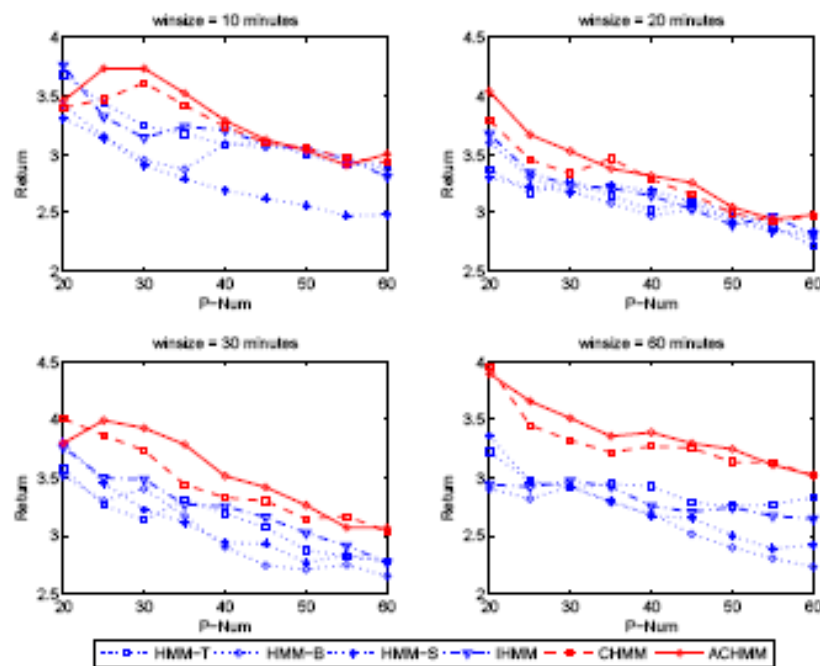


Fig. 9. Return of Six Models

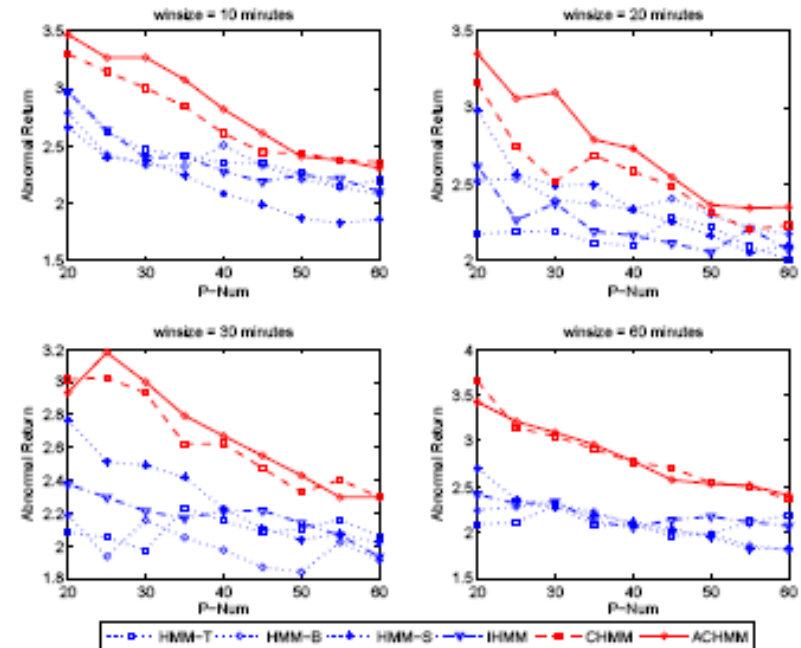
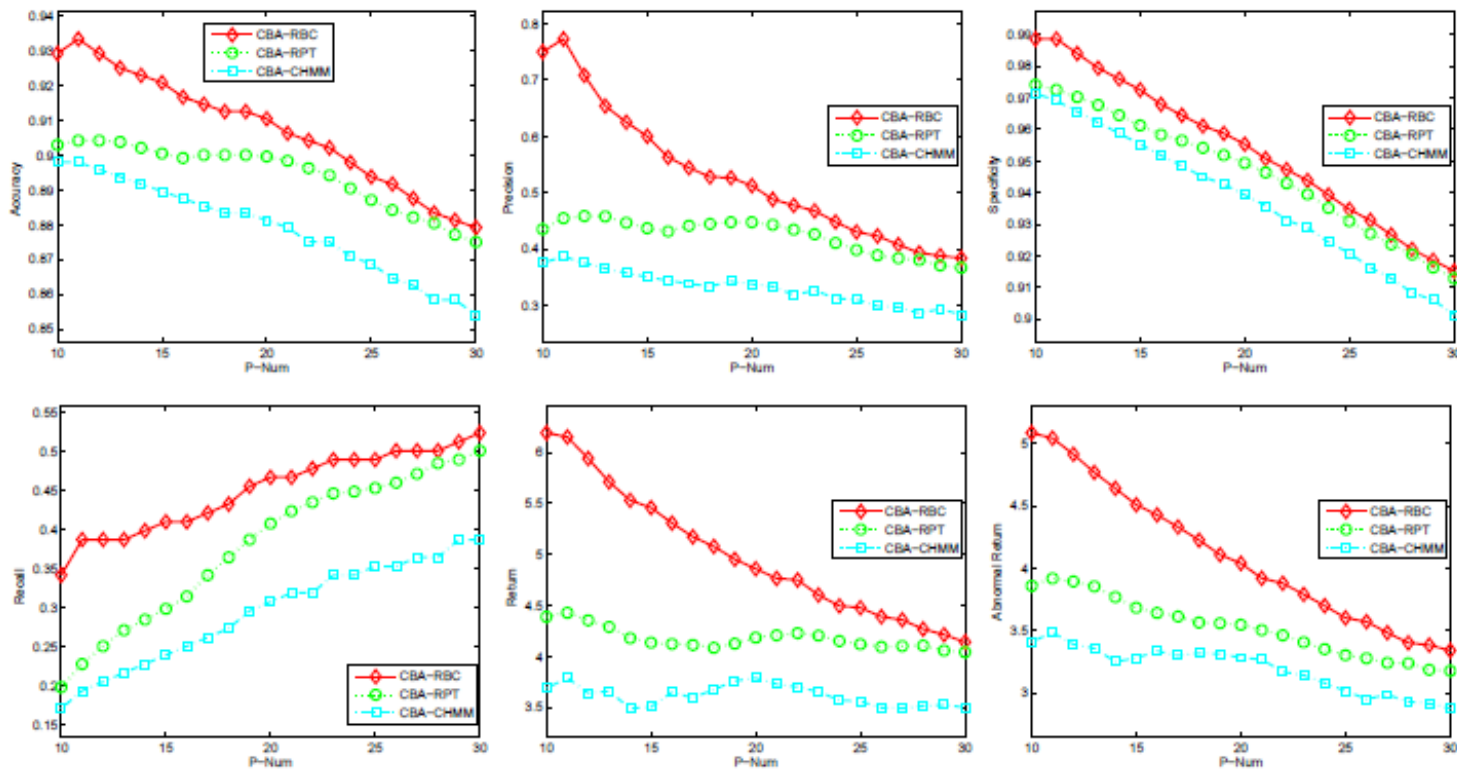


Fig. 10. Abnormal Return of Six Models

Empirical Results – Learning Group Trading Behaviors





<http://australian-animals.net/>

Non-IID Document Analysis

Xin Cheng, Duoqian Miao, Can Wang, Longbing Cao. [Coupled Term-Term Relation Analysis for Document Clustering](#), IJCNN2013.

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.

The BOW Similarity

Table 1. An Example of Document Representation: “*DM*”, “*ML*”, “*DB*” and “*CS*” denote “*Data mining*”, “*Machine learning*”, “*Database*” and “*Computer science*”, respectively.

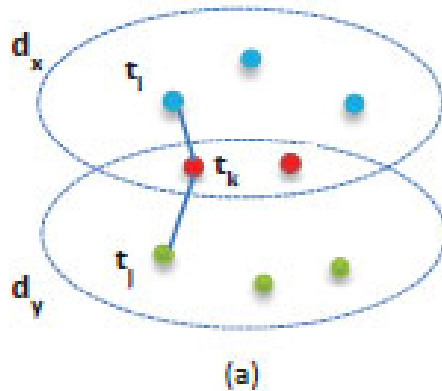
	<i>DM</i>	<i>ML</i>	<i>DB</i>	<i>CS</i>
d_1	0.5	0.0	0.1	0.3
d_2	0.0	0.5	0.1	0.25
d_3	0.0	0.0	0.8	0.1

- The cosine similarity between d_1 and d_2 is 0.253, and 0.231 for d_1 and d_3
- The similarity values are approximate, thus, it is unable to identify which two documents are more alike if the relation between terms is not captured.

Coupled Term-Term Relation Learning

Xin Cheng, Duoqian Miao, Can Wang, Longbing Cao. [Coupled Term-Term Relation Analysis for Document Clustering](#), IJCNN2013.

Intra-term Relations



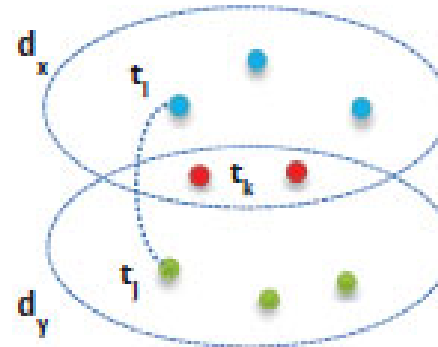
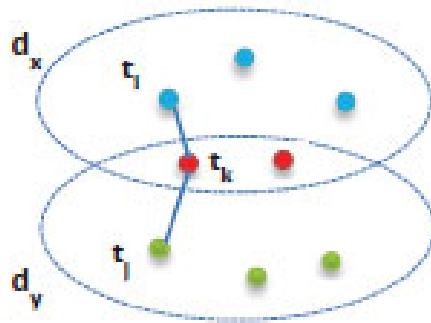
Terms are **relational** if they co-occur in the same document.

- Terms t_i and t_k co-occur in document d_x , while t_j is the co-occurrence term of t_k in document d_y .
- Then, term t_i is considered to be associated with t_k in document d_x , and term t_j is related with t_k in document d_y .

Inter-term Relations

Definition 3. Terms t_i and t_j are said to be *inter-related*, if there exists at least one term t_k such that both $IaR(t_k, t_i) > 0$ and $IaR(t_k, t_j) > 0$ hold. The term t_k is called the *link term* between them. The *relative inter-relation* between terms t_i and t_j linked by the term t_k is formalized as:

$$R_IeR(t_i, t_j | t_k) = \min(IaR(t_i, t_k), IaR(t_j, t_k)), \quad (3.3)$$





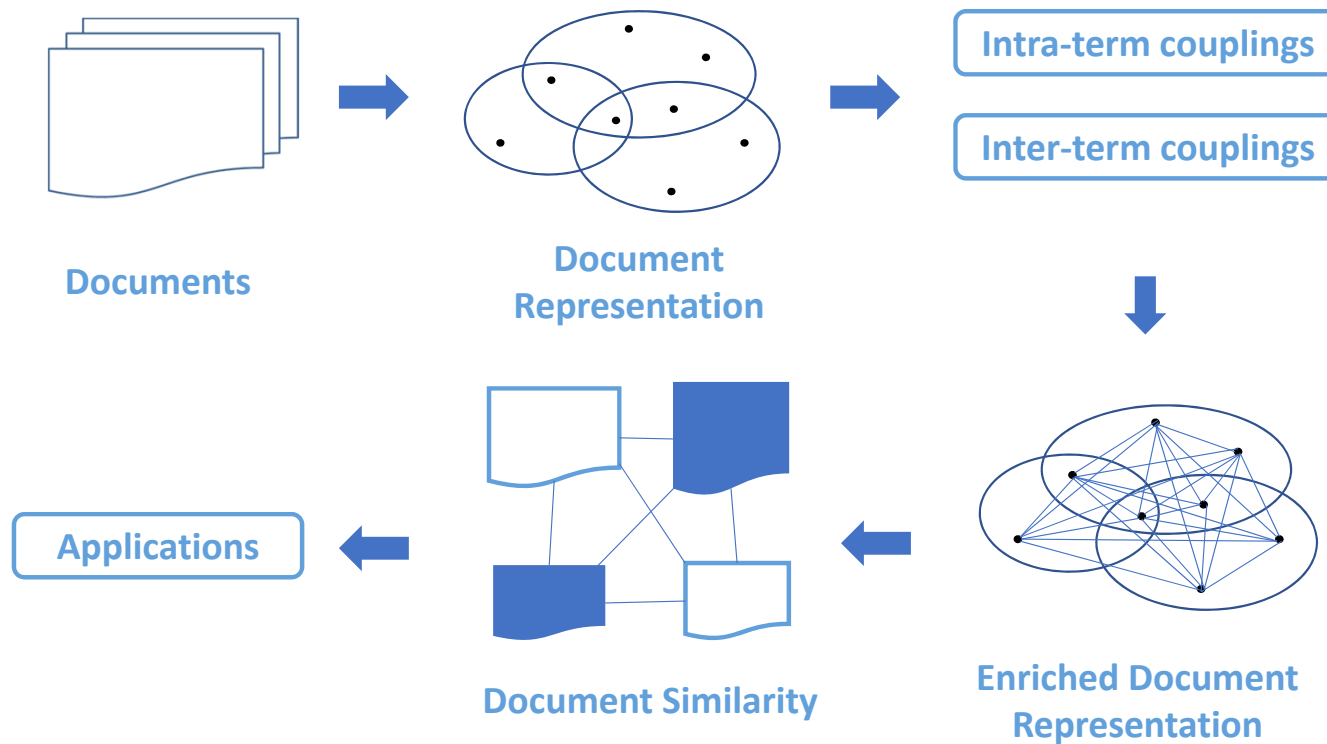
<http://australian-animals.net/>

Document Similarity by Learning Term Pair Couplings

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.

Main Ideas

Semantic Couplings of Term Pairs




Intra-Term Pair Couplings

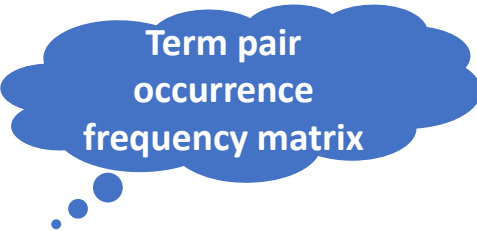
1. Semantic Intra-couplings within Term Pairs

1) **DEFINITION 1** *tpf-idf*, short for *term pair occurrence frequency - inverse document frequency*, reflects the importance of a term pair to a document in a collection or corpus. *tpf* counts the number of times a term pair occurs in a document. The *tpf-idf* scheme is formatted as:

$$tpfidf((t_i, t_j), d, D) = tpf((t_i, t_j), d) \times idf((t_i, t_j), D)$$

where (t_i, t_j) stands for a term pair, and d is a single document in a document collection D .


$$M_{tpf} = \begin{matrix} & \begin{matrix} t_1 & t_2 & \cdots & t_K \end{matrix} \\ \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_K \end{matrix} & \begin{pmatrix} 0 & tpf_{12} & \cdots & tpf_{1K} \\ tpf_{21} & 0 & \cdots & tpf_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ tpf_{K1} & tpf_{K2} & \cdots & 0 \end{pmatrix} \end{matrix}$$

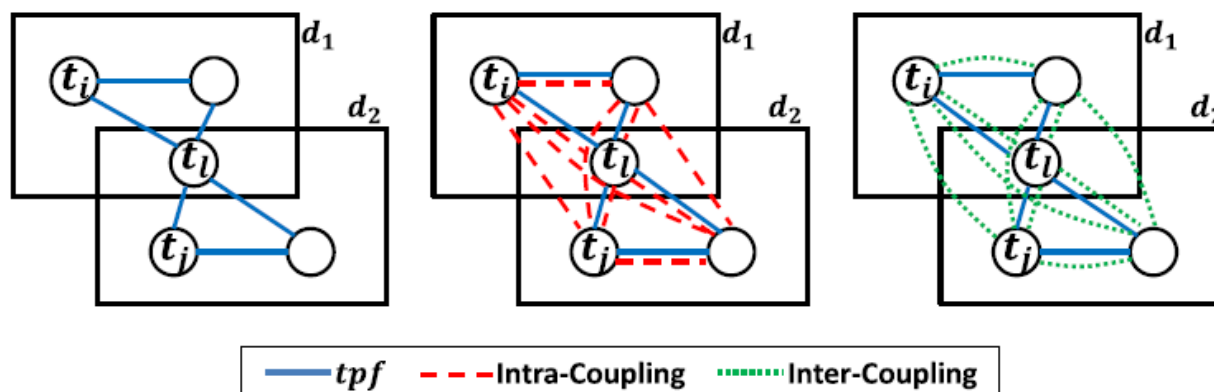


Term pair
occurrence
frequency matrix

Inter-Term Pair Couplings

2. Semantic Inter-couplings between Term Pairs

- 1) Based on M_{tpf} , the term pair frequency graph G_{tpf} is an ordered pair, $G_{tpf} = (T, E_{tpf})$, comprising a set T of terms as vertexes, $T = \{t_k | k \in [1, K]\}$, together with a set E_{tpf} as edges to reflect the tpf of every term pair.



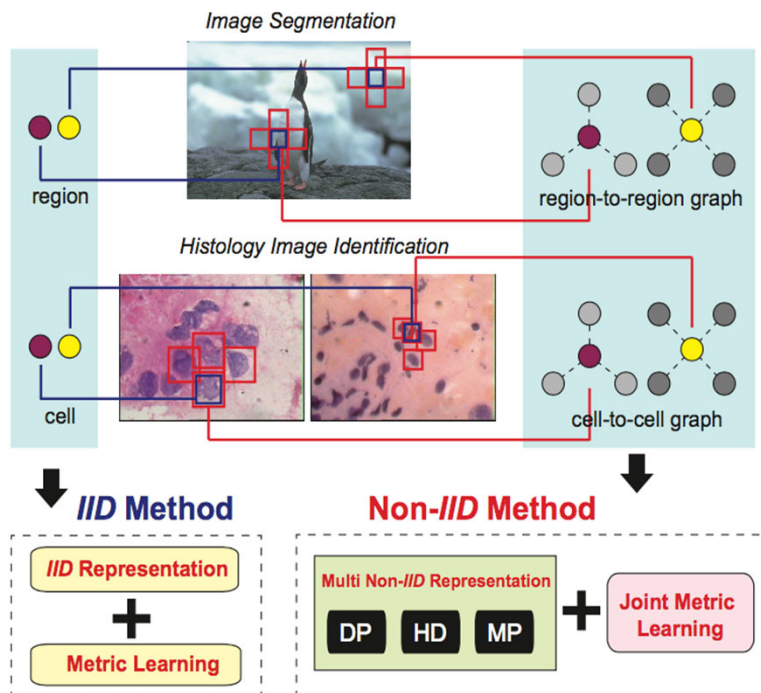
Intra-coupling: counts the explicit relation of each directly connected term pair on G_{tpf}

Inter-coupling: counts the implicit relation of each term pair on G_{tpf} through other terms

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

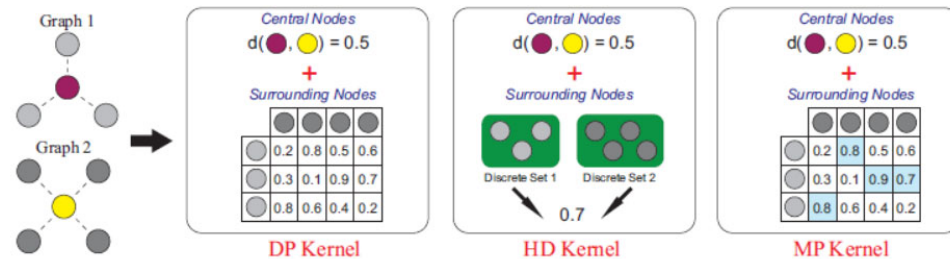
★ Many features, representations and classifiers can be integrated

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)



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

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

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

Learning/combining Multiple Non-IID Representations

Objective function for combined non-IID metrics

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

$$\begin{aligned} \arg \min_{w^p} \sum_{i,j} \psi_{ij} & \underbrace{\left\| \Omega \left(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_j^p \right) \right\|^2}_{\text{Pair-wise Constraint}} + \\ & \lambda \sum_{i,j,l} \psi_{ij} (1 - y_{il}) h \left[\underbrace{\left\| \Omega \left(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_j^p \right) \right\|^2}_{\text{Triplet Constraint}} \right. \\ & \quad \left. - \underbrace{\left\| \Omega \left(\sum_p w^p \mathbf{k}_i^p - \sum_p w^p \mathbf{k}_l^p \right) \right\|^2}_{\text{Triplet Constraint}} + 1 \right]. \\ \text{s.t.} \quad & \sum_p w^p = 1, w^p \geq 0. \end{aligned}$$

Feature Construction

Feature construction

Hand-crafted features (HC):

- Those features whose effectiveness are already validated are chosen, including height, width, RGB, HSI, area, circumference, Fourier descriptor, entropy, and central moment.
- In total, to represent a cell region, 37-dimensional features are used.

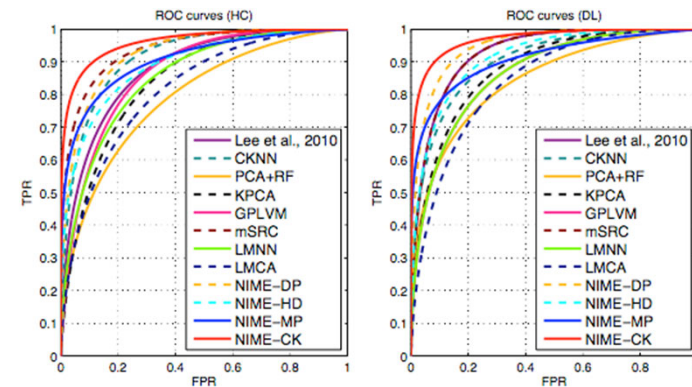
Deeply-learned features (DL):

For relative small-scale cell regions compared with natural images,

- use the bounding box to bound the irregular segmented cell regions, resize them into 32×32 patches,
- employ the *LeNet* model to automatically learn the deep features,
- form a 64-dimensional feature for each cell region.

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.*

Convergence

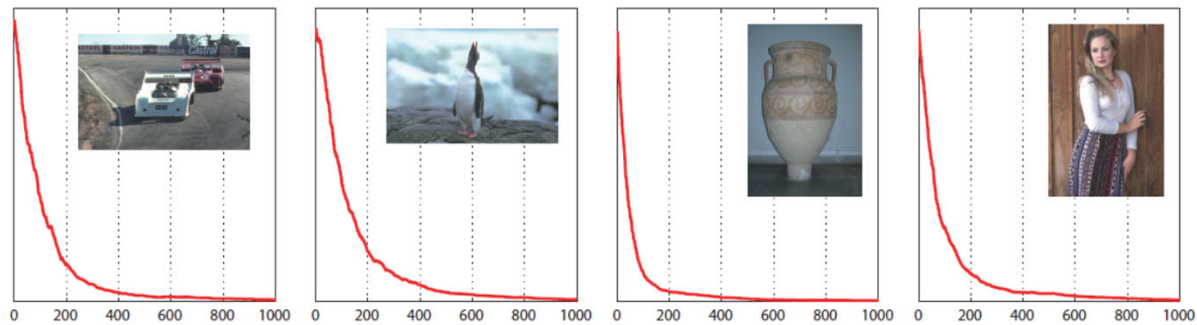


Figure 3: *Illustration of the convergence of NIME-CK.*

Pattern Relation Analysis/ Combined Pattern Mining

Combined pattern pairs

- Pair patterns

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

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

$$\mathcal{E}: \begin{cases} X_p \rightarrow T_1 \\ X_p \wedge X_e \rightarrow T_2 \end{cases}$$

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

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

$$\begin{aligned} Cont_e(P) &= \frac{Lift(X_p \wedge X_e \rightarrow T)}{Lift(X_p \rightarrow T)} \\ &= \frac{Conf(X_p \wedge X_e \rightarrow T)}{Conf(X_p \rightarrow T)} \end{aligned}$$

$$I_{\text{rule}}(X_p \wedge X_e \rightarrow T) = \frac{Cont_e(X_p \wedge X_e \rightarrow T)}{Lift(X_e \rightarrow T)}$$

$$Cps(X_e \rightarrow T|X_p) = Prob(X_e \rightarrow T|X_p) - Prob(X_e|X_p) \times Prob(T|X_p)$$

$$= \frac{Prob(X_p \wedge X_e \rightarrow T)}{Prob(X_p)} - \frac{Prob(X_p \wedge X_e)}{Prob(X_p)} \times \frac{Prob(X_p \rightarrow T)}{Prob(X_p)}$$

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 pairs

Traditional Association Rules

<i>V</i>		<i>T</i>	<i>Conf</i> (%)	<i>Count</i>	<i>Lift</i>
Arrangement	Repayment	Class			
irregular	cash or post office	A	82.4	4088	1.8
withholding	cash or post office	A	87.6	13354	1.9
withholding & irregular	cash or post office	A	72.4	894	1.6
withholding & irregular	cash or post office & withholding	B	60.4	1422	1.7

An Example of Combined Patterns

Rules	<i>X_p</i>	<i>X_e</i>			<i>Cnt</i>	<i>Conf</i> (%)	<i>I_r</i>	<i>Lift</i>	<i>Cont_p</i>	<i>Cont_e</i>	<i>Lift of</i> <i>X_p → T</i>	<i>Lift of</i> <i>X_e → T</i>
	Demographics	Arrangements	Repayments	Class								
<i>P</i> ₁	age:65+	withholding & irregular	withholding	C	50	63.3	2.91	3.40	2.47	4.01	0.85	1.38
<i>P</i> ₂	income:0 & remote:Y & marital:sep & gender:F	withholding	cash or post & withholding	B	20	69.0	1.47	1.95	1.34	2.15	0.91	1.46
<i>P</i> ₃	income:0 & age:65+	withholding	cash or post & withholding	A	1123	62.3	1.38	1.35	1.72	1.09	1.24	0.79
<i>P</i> ₄	income:0 & gender:F & benefit:P	withholding	cash or post	A	469	93.8	1.36	2.04	1.07	2.59	0.79	1.90

Combined pattern clusters

- Cluster patterns

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

$$\mathcal{C}: \begin{cases} X_1 \rightarrow T_1 \\ \dots \\ X_k \rightarrow T_k \end{cases}$$

$$\mathcal{S}: \begin{cases} X_p \rightarrow T_1 \\ X_p \wedge X_{e,1} \rightarrow T_2 \\ X_p \wedge X_{e,1} \wedge X_{e,2} \rightarrow T_3 \\ \dots \\ X_p \wedge X_{e,1} \wedge X_{e,2} \wedge \dots \wedge X_{e,k-1} \rightarrow T_k \end{cases}$$

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

Combined pattern clusters

An Example of Combined Pattern Clusters

Clusters	Rules	X_p	X_e		T	Cnt	$Conf$ (%)	I_r	I_c	$Lift$	$Cont_p$	$Cont_e$	$Lift$ of $X_p \rightarrow T$	$Lift$ of $X_e \rightarrow T$
		demographics	arrangements	repayments										
\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

- 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

Structural pattern relations

- Peer-to-peer patterns

$$\mathcal{P} ::= P_1 \cup P_2$$

- Master-slave patterns

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

- Hierarchical patterns

$$\{\mathcal{P} ::= P_i \cup P'_i \cup P_j \cup P'_j, P_j = \mathcal{G}(P_i), \dots, P'_j = \mathcal{G}'(P_i)'\}$$

Temporal pattern relations

- Independent patterns

$$\{P_1 : P_2\}$$

- Sequential patterns

$$\{P_1; P_2\}$$

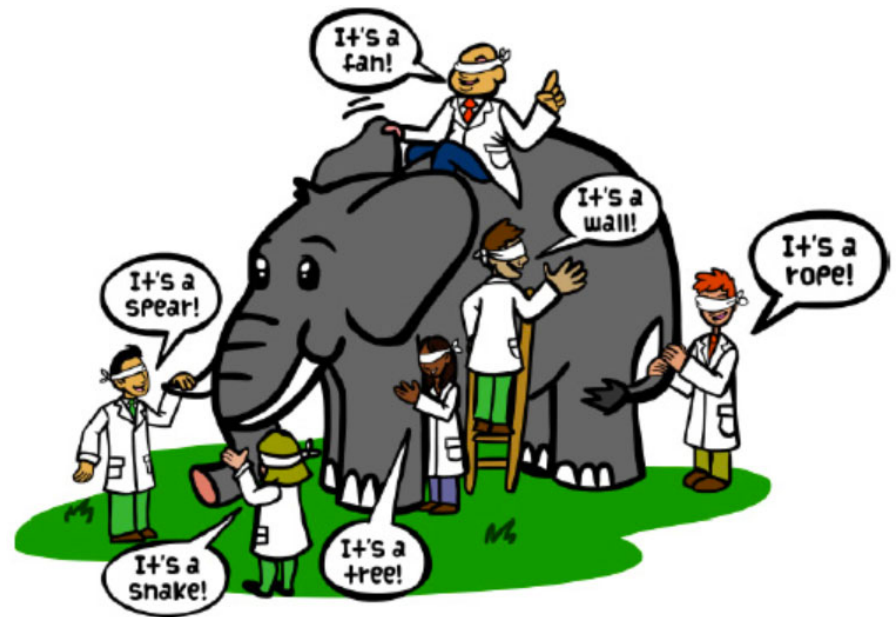
- Hybrid patterns

$$\{P_1 \otimes P_2 \cdots \otimes P_n; \otimes \in \{:, \parallel, ;\}\}$$

Conclusions & Prospects

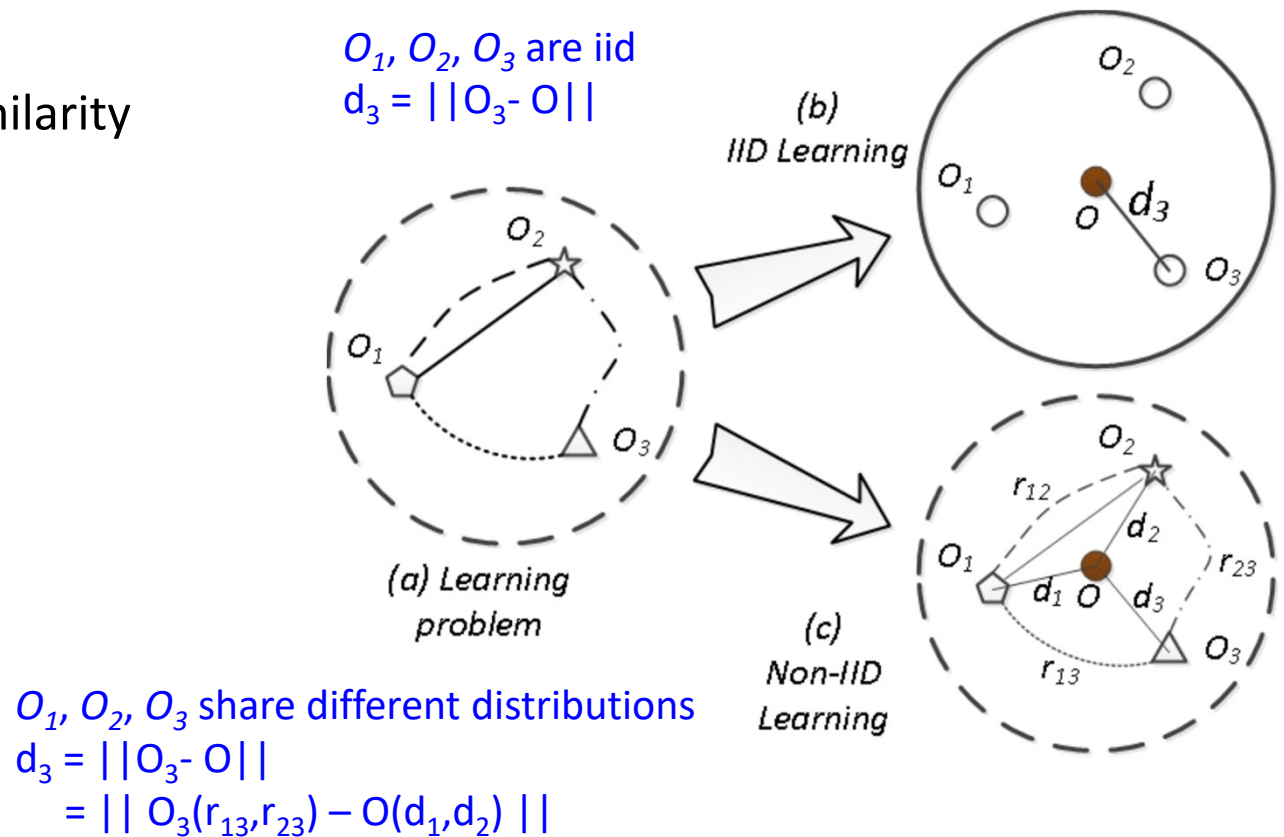
How Can Blind People Recognize An Elephant?

- How can blind people tell a genuine story about elephant?
 - Non-IID learning?
 - Couplings between parts
 - Heterogeneity between parts
 - From touching/representation → analysis → reasoning/inference → summarization
 - Local – global picture (known → unknown)/optimization

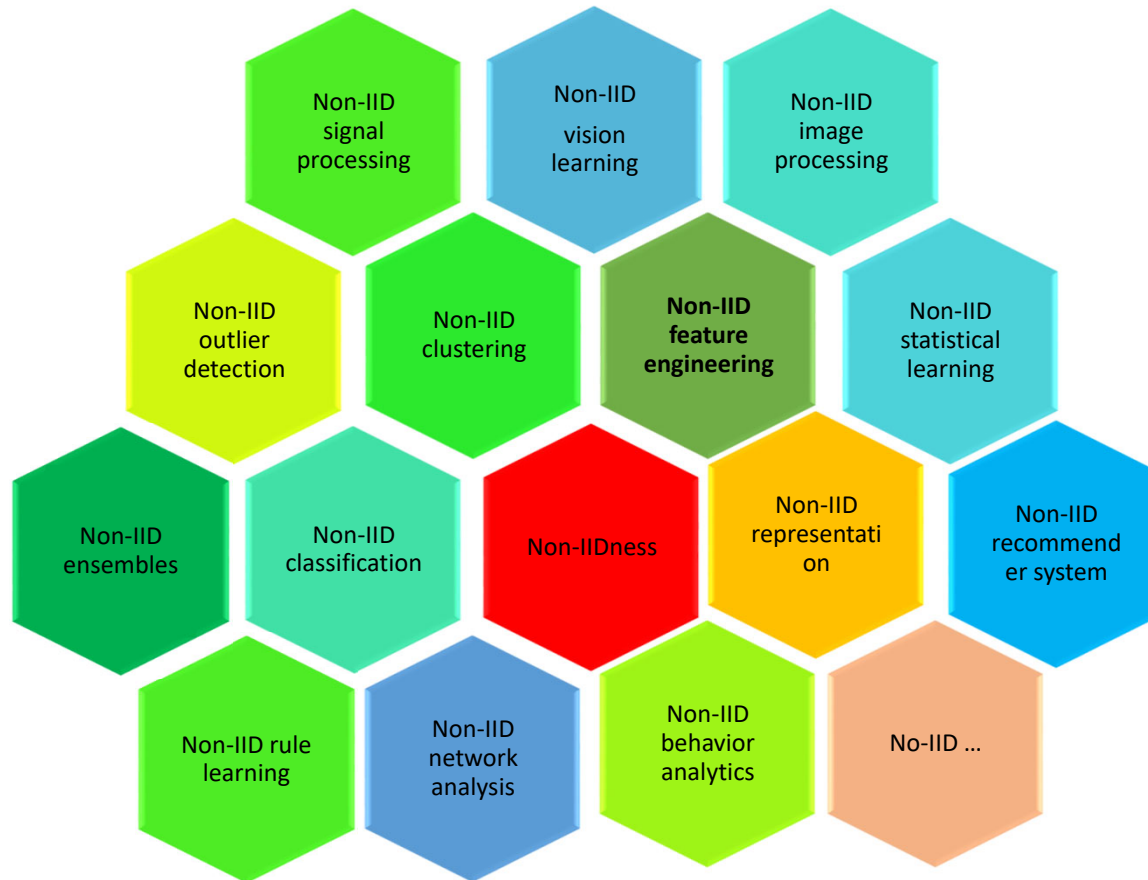


Non-IID Learning: A Challenging Problem

- Data non-IIDness
- Non-IID similarity/dissimilarity metrics/measures
- Non-IID representations
- New objective functions
- New perspectives



Non-IID Learning: A Significant Area



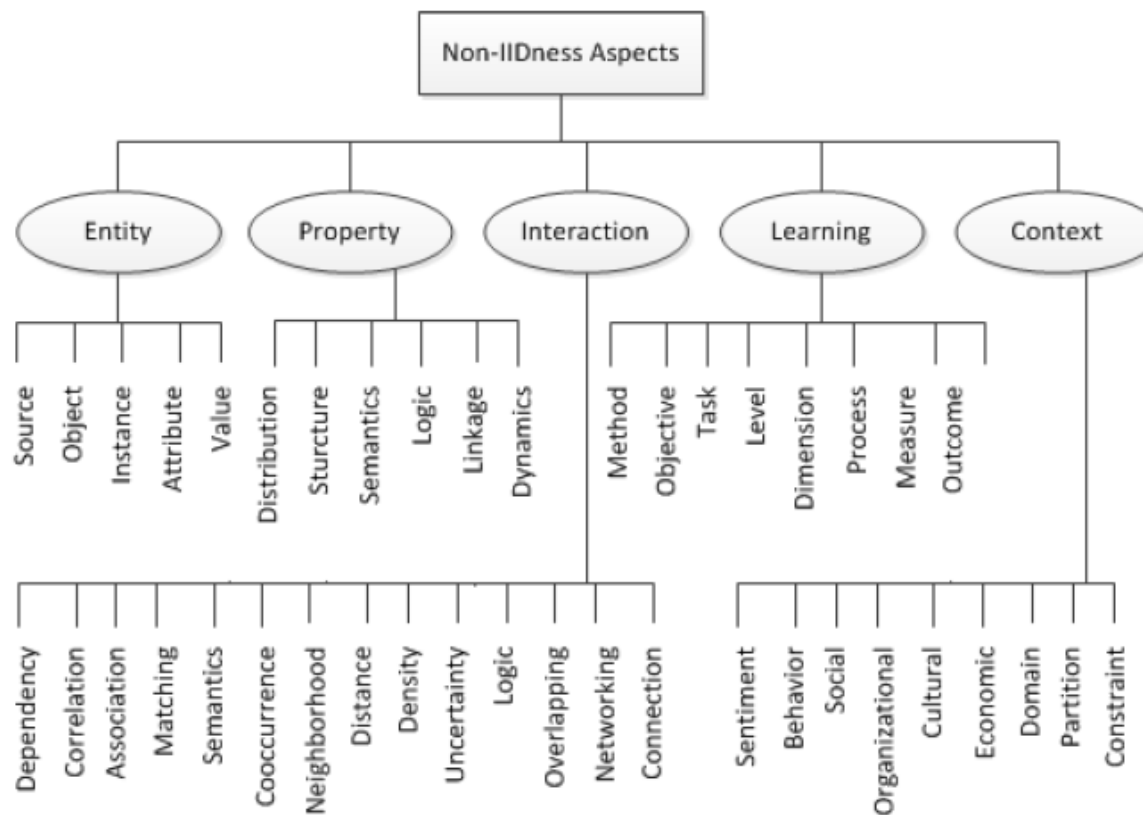
Some Fundamental Issues

- 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

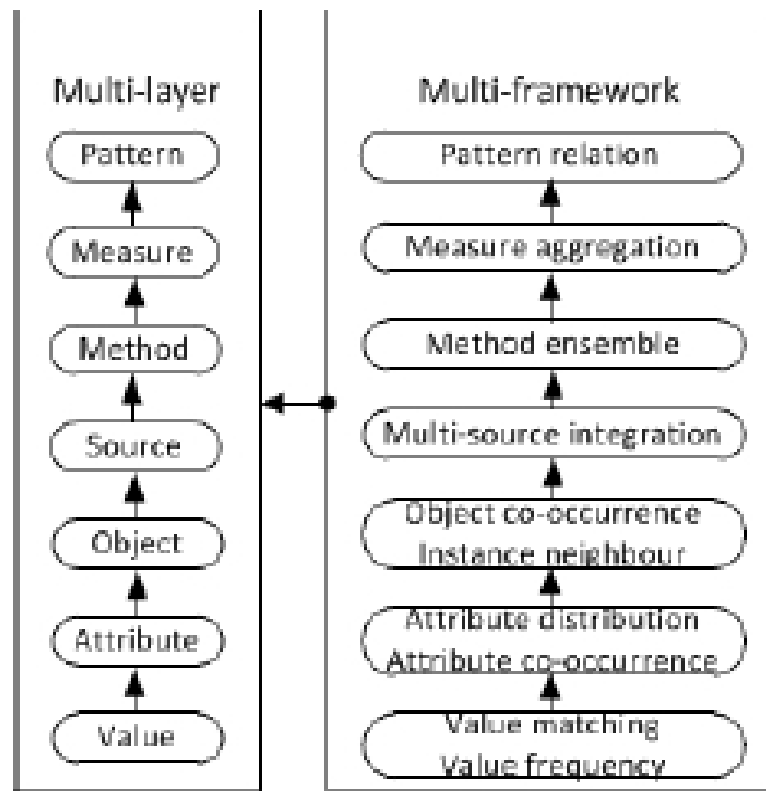
- Are KNN, SVM, decision tree, classic ensemble methods IID?
- Does classic transfer learning capture non-IIDness?
- In probabilistic graphical modeling, how non-IIDness is modelled?
- Do deep neural networks capture non-IIDness? To what extent?
- ...

Aspects of Non-IIDness



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)

References

Not all references are listed here

References

Paper download: www.datasciences.org

- **Non-IID learning concepts**

- 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**

- 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.

- **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**

- Shoujin Wang, Longbing Cao. [Inferring Implicit Rules by Learning Explicit and Hidden Item Dependency](#). IEEE Transactions on Systems, Man, and Cybernetics: Systems
- 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**

- 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)

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

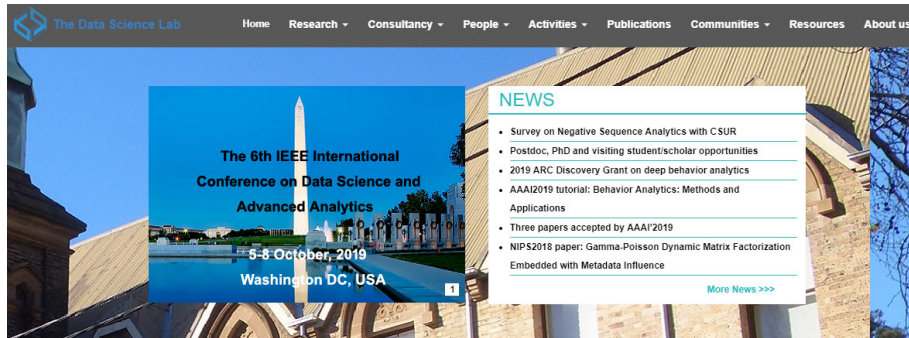
- Jian, S., Cao, L., Pang, G., & Lu, K., Gao, H. (2017 August). Embedding-based Representation of Categorical Data by Hierarchical Value Coupling Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- Kim, D. I., Hughes, M., & Sudderth, E. (2012). 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.

PhD Scholarship Opportunities

- 3 PhD scholarships are available for gifted master students to study at the Data Science Lab on data science/AI/ML frontiers
- AUD\$27k or more p.a. for 3-3.5 years, \$34k for tuition fee p.a.
- Master by research
- Major in statistics, applied mathematics, or computing science
- Published some good papers as first-author
- Outstanding performance in ungraduated and postgraduate studies
- English: IELTS Band 6.5
- For more information, Data Science Lab www.datasciences.org



Thank You Very Much

Comments & suggestions:

Longbing.Cao@uts.edu.au

Data Science Lab:

www.datasciences.org

Learn More ►

Enterprise Data Innovation

Enterprise data are growing increasingly bigger and bigger, more and more complex, and more and more valuable. Data science and intelligence science have played critical roles in discovering the intelligence, value and insight and in recommending smarter decision-making actions for enterprise innovation, productivity transformation and competitive strength upgrading. Our team has been well known for its leadership in industry and corporate engagement, high standard and demonstrated impact in assisting major industry and government organizations in building



the thinking and foundation

The thinking and foundation to design, implement, manage, review and optimize enterprise data science innovation decision-making, plans, policies, mechanisms and specifications;



the competencies and skills

The competencies and skills to create, undertake and optimize enterprise data science infrastructure, systems, models, case studies, and practice;



the qualifications

the qualifications for next-generation data science professionals through offering high quality Master's/doctoral courses and corporate workshop/training to undertake and lead actionable enterprise data science.



IEEE DSAA'2019
5-8 Oct, Washington DC

