Noniidness Learning: Coupled Object and Pattern Analysis

Professor Longbing CAO

Director,
Advanced Analytics Institute
University of Technology Sydney
References

• Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery.
• Can Wang, Zhong She, Longbing Cao. Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects, ICDE2013.
• Zhong She, Can Wang, and Longbing Cao. A Coupled Framework of Clustering Ensembles, AAAI2012 (poster)
• Can Wang, Mingchun Wang, Zhong She, Longbing Cao. CD: A Coupled Discretization Algorithm, PAKDD2012, 407-418
• Can Wang, Longbing Cao, Minchun Wang, Jinjiu Li, Wei Wei, Yuming Ou. Coupled Nominal Similarity in Unsupervised Learning, CIKM 2011, 973-978.
• 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.
• Yanchang Zhao, Huafeng Zhang, Fernando Figueiredo, Longbing Cao Chengqi Zhang, Mining for Combined Association Rules on Multiple Datasets, Proc. of 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM 07), 2007, pp. 18-23.
The Challenge of Big Data: Noniidness
**Big data—a growing torrent**

- $600 to buy a disk drive that can store all of the world’s music
- 5 billion mobile phones in use in 2010
- 30 billion pieces of content shared on Facebook every month
- 40% projected growth in global data generated per year vs. 5% growth in global IT spending
- 235 terabytes data collected by the US Library of Congress by April 2011
- 15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

**Big data—capturing its value**

- $300 billion potential annual value to US health care—more than double the total annual health care spending in Spain
- €250 billion potential annual value to Europe’s public sector administration—more than GDP of Greece
- $600 billion potential annual consumer surplus from using personal location data globally
- 60% potential increase in retailers’ operating margins possible with big data
- 140,000–190,000 more deep analytical talent positions, and
- 1.5 million more data-savvy managers needed to take full advantage of big data in the United States
Big Data Analytics

• Learning relevant complex resources for complex/actionable knowledge in full context
“What enterprise areas does your big data initiative address?”

- Marketing: 45%
- Operations: 43%
- Sales: 38%
- Risk management: 35%
- IT analytics: 33%
- Finance: 32%
- Product development: 32%
- Customer service: 30%
- Logistics: 22%
- HR: 12%
- Other: 12%
- Brand management: 8%

Base: 60 IT professionals

Source: June 2011 Global Big Data Online Survey
What make “Big” challenge?

• Size:
  – Number of transactions
  – Number of dimensions

• Heterogeneity:
  – Number of data types
  – Types of data sources

• Coupling:
  – Coupling between objects, attributes, values, relations, methods and impacts etc

Not identically distributed.
Not independent.
The Trends

Accessibility

Visibility

Operability

Data size

Analytics complexity

Business value
The Trends

What drives the gap?
- Heterogeneity
- Coupling
  ...
  ...
Noniidness learning
What is Noniidness

- Noniidness:
  - Dependency
  - Heterogeneity
Iidness learning
Dominates Current KDD/ML Research
$O_1, O_2, O_3$ are independent:
$d = ||O_3 - O||$
Iidness Learning

Traders are independent

Behaviors of a trader are treated independent or loosely dependent

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

- **K-means**
  \[
  \arg \min_{\mathbf{s}} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2
  \]

- **FCM**
  \[
  J_{FCM}(\mu, A) = \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^m \|x_j - a_i\|^2
  \]
  \[
  \sum_{i=1}^{c} \mu_{ij} = 1 \quad \text{for all } j \in J.
  \]

**Note:**
- $X_j$ Individual objects only!

**Question:**
- How about $X_{j1}$ and $X_{j2}$ dependent?
Objective functions:

- Decision tree
  \[(x, Y) = (x_1, x_2, x_3, ..., x_k, Y)\]
  \[I_G(f) = \sum_{i=1}^{m} f_i (1 - f_i) = \sum_{i=1}^{m} (f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2\]
  \[I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i\]

- KNN
  Euclidean distance: \[d(x_1, x_2)\]
  Hamming distance: \[d(s_1, s_2)\]

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?
Iidness vs Noniidness

\[ O_1, O_2, O_3 \text{ are iid} \]
\[ d_3 = ||O_3 - O|| \]

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

**FIGURE 1.** Iidness learning vs. noniidness learning.
### TABLE 1. The Information Table

<table>
<thead>
<tr>
<th>O</th>
<th>A</th>
<th>A₁</th>
<th>A₂</th>
<th>…</th>
<th>Aⱼ</th>
<th>M₁</th>
<th>…</th>
<th>Mᵠ₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>O₁</td>
<td>V₁₁</td>
<td>V₁₂</td>
<td>V₁₃</td>
<td>…</td>
<td>V₁ⱼ</td>
<td>C₁₁</td>
<td>…</td>
<td>C₁ᵠ₂</td>
</tr>
<tr>
<td>O₂</td>
<td>V₂₁</td>
<td>V₂₂</td>
<td>V₂₃</td>
<td>…</td>
<td>V₂ⱼ</td>
<td>C₂₁</td>
<td>…</td>
<td>C₂ᵠ₂</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Oᵠ</td>
<td>Vᵠ₁</td>
<td>Vᵠ₂</td>
<td>Vᵠ₃</td>
<td>…</td>
<td>Vᵠⱼ</td>
<td>Cᵠ₁</td>
<td>…</td>
<td>Cᵠᵠ₂</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Oᵣ</td>
<td>Vᵣ₁</td>
<td>Vᵣ₂</td>
<td>Vᵣ₃</td>
<td>…</td>
<td>Vᵣⱼ</td>
<td>Cᵣ₁</td>
<td>…</td>
<td>Cᵣᵠ₂</td>
</tr>
</tbody>
</table>

### FIGURE 3. Information table and couplings for noniiddness learning.
Noniidness in Complex Behaviors
Behavior Visual Descriptor

- **Instance Of** — — →
  Connecting instances (in Rectangle) to their corresponding classes

- **Subclass Of** — — →
  Linking a subclass (in Oval) to its parent class

- **Object Property** — — →
  Denoting the relationships between instances, between an object and its properties (in Rounded Rectangle), or between properties.
Coupling Relationships

Perspectives

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

Inferential
- Causal Coupling
- Conjunction Coupling
- Disjunction Coupling
- Exclusive Coupling

Party-based
- One-Party-Multiple-Operation
- Multiple-Party-One-Operation
- Multiple-Party-Multiple-Operation
Temporal Coupling

- **Serial coupling**, denoted by \{B_1;B_2\}, showing the situation in which behavior \(B_2\) follows behavior \(B_1\).

- **Parallel coupling**, by which behaviors happen in varying concurrent manners, including synchronous coupling and asynchronous coupling.
  
  - **Synchronous relationship**, denoted by \{B_1\parallel B_2\}, indicating that \(B_1\) and \(B_2\) present at the same time based on certain communication protocols.
Temporal Coupling

- **Asynchronous coupling**, showing that two behaviors $B_1$ and $B_2$ interact with each other at different time points.
  
  * **Interleaving**, denoted by $\{B_1 : B_2\}$, representing the involvement of independent complex behaviors by nondeterministic choice (independently).

  * **Shared-variable**, denoted by $\{B_1 | | | B_2\}$, signifying that the relevant behaviors have variables in common.

  * **Channel system**, denoted by $\{B_1 | B_2\}$, is a parallel system in which complex behaviors communicate via a channel, for instance, first-in and first-out buffers.
Inferential Coupling

• **Causal coupling**, represented as \( \{B_1 \rightarrow B_2\} \), meaning that behavior \( B_1 \) causes behavior \( B_2 \).

• **Conjunction coupling**, \( \{B_1 \land B_2\} \), specifying that \( B_1 \) and \( B_2 \) take place together.

• **Disjunction coupling**, \( \{B_1 \lor B_2\} \), by which at least one of the associated behaviors must happen.

• **Exclusive coupling**, \( \{B_1 \oplus B_2\} \), indicating that if \( B_1 \) happens, \( B_2 \) will not happen, and vice versa.
• **One-Party-Multiple-Operation**, represented as \(((B_1, B_2)^{[A_1]})\), depicts that distinct behaviors \(B_1\) and \(B_2\) are performed by the same actor \(A_1\).

• **Multiple-Party-One-Operation**, shown as \(((B_1)^{[A_1A_2]})\), represents that multiple actors \(A_1\) and \(A_2\) implement the same behavior \(B_1\) to achieve their own intentions.

• **Multiple-Party-Multiple-Operation**, presented as \(((B_1, B_2)^{[A_1A_2]})\), describes that different behaviors \(B_1\) and \(B_2\) are carried out by distinct actors \(A_1\) and \(A_2\).
The intra-coupling reveals the complex couplings within an actor’s distinct behaviors.

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

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

where $\sum_{j=1}^{J_{\text{max}}} \odot$ means the subsequent behavior of $B_i$ is $B_{ij}$ intra-coupled with $\theta_j(B)$, and so on.

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

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

\[
\mathcal{B}_{ij}^n := \mathcal{B}_{j}(\mathcal{A}, \Theta, \eta) \left| \sum_{i=1}^{I} \eta_i(\mathcal{B}) \odot \mathcal{B}_{ij} \right.
\]

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

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.
Definition 4 (Coupled Behaviors): Coupled behaviors $B_c$ refer to behaviors $B_{i_1,j_1}$ and $B_{i_2,j_2}$ that are coupled in terms of relationships $h(\theta(B), \eta(B))$, where $(i_1 \neq i_2) \lor (j_1 \neq j_2) \land (1 \leq i_1, i_2 \leq I) \land (1 \leq j_1, j_2 \leq J_{max})$

$$B_c = (B_{i_1,j_1}^\theta)^n \times (B_{i_2,j_2}^\theta)^n := B_{ij}(A, \Theta, \mathcal{C}) \bigg| \sum_{i_1,i_2=1}^{I} \sum_{j_1,j_2=1}^{J_{max}} h(\theta_{j_1,j_2}(B), \eta_{i_1,i_2}(B)) \odot (B_{i_1,j_1} B_{i_2,j_2}); \quad \text{(IV.4)}$$

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

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

Three types of aggregations:

- **Intra-Coupled Aggregation**
  - Function: $\theta_j$

- **Inter-Coupled Aggregation**
  - Function: $\eta_i$

- **Combined Aggregation**
  - Function: $h(\theta_{j1,j2}, \eta_{i1,i2})$
Behavior Coupling Types

- Logic/semantic relation based behavior coupling
  
  Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery.
  

- Statistical/Probabilistic relation based behavior coupling
  
  
  
Coupled Nominal Similarity in Unsupervised Learning

Can Wang, Longbing Cao, Mingchun Wang, Jinjiu Li, Wei Wei, Yuming Ou

University of Technology, Sydney, Australia

Wednesday, 26 Oct. 2011, Glasgow, UK
Coupled Object Analysis
Similarity Analysis

The more two objects resemble

The larger the similarity
Similarity Analysis

Attribute Values

Continuous

\[ \delta(x, y) = |x - y| \]
- L^p Distance
  - Manhattan
  - Euclidean
  - Minkowski
  - Chebyshev

Nominal

- Supervised
  - Value Distance Matrix
  - Modified Value Distance Matrix
  - Heterogeneous
- Unsupervised
  ...... (Our Focus)
Related Work

Simple Matching (SM)
Jaccard
Russell and Rao

Occurrence Frequency
Goodall
Anderberg

Iterated Contextual
Ahmad and Dey (AD)
Relevant algorithms

- **ROCK**: hierarchical clustering algorithm, use the link-based similarity measure to measure the similarity between two categorical data points and between two clusters
- **STIRR**: first construct a hypergraph according to the database, and then cluster the hypergraph
- **LIMBO**: hierarchical categorical clustering algorithm, build on the Information Bottleneck (IB) framework for quantifying the relevant information preserved when clustering
- **COOLCAT**: incremental algorithm to cluster categorical attributes using entropy, assumption of independence between the attributes
- **CLOPE**: Clustering with sLOPE, uses a global criterion function instead of a local one defined by pairwise similarity for clustering categorical data, especially transactional data
- **CLICKS**: a graphtheoretic approach to find $k$ disjoint sets of vertices in a graph constructed for a particular data set
Simple Matching Similarity (SMS) [5] and its diverse variants such as Jaccard coefficients [12]

- based on the principle that the similarity measure is
  - 1 with identical values, and
  - 0 otherwise
Value frequency distribution

• Neighborhood-based similarity
  – The frequency distribution of attribute values is considered for similarity measures
  – They present the similarity between a pair of objects by considering only the relationships between data objects, which are built on the similarity between attribute values
  – Quantified by SMS or its variants
Feature dependency aggregation

- The Pearson correlation coefficient [14] measures only the strength of linear dependence between two variables (e.g., nominal attributes);
- Das and Mannila put forward the Iterated Contextual Distances algorithm, believing that the attribute, object, and sub-relation similarities are inter-dependent [6].
• Andritsos et al. [16] introduced a context sensitive dissimilarity measure between attribute values based on the Jensen-Shannon divergence (for two probability distributions).

• Similarly, Ahmad and Dey [11] proposed an algorithm *ADD to compute the dissimilarity between* attribute values by considering the co-occurrence probability between each attribute value and the values of another attribute.
Limitation

• Within and between attribute couplings
• Multi-level of couplings
• Very few approaches
• No comprehensive frameworks and mechanisms
• High computational cost
• Separation between coupling and heterogeneity
• ...

4/16/2013
Motivation: Example

<table>
<thead>
<tr>
<th>Movie</th>
<th>Director</th>
<th>Actor</th>
<th>Genre</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Godfather II</td>
<td>Scorsese</td>
<td>De Niro</td>
<td>Crime</td>
<td>$G_1$</td>
</tr>
<tr>
<td>Good Fellas</td>
<td>Coppola</td>
<td>De Niro</td>
<td>Crime</td>
<td>$G_1$</td>
</tr>
<tr>
<td>Vertigo</td>
<td>Hitchcock</td>
<td>Stewart</td>
<td>Thriller</td>
<td>$G_2$</td>
</tr>
<tr>
<td>N by NW</td>
<td>Hitchcock</td>
<td>Grant</td>
<td>Thriller</td>
<td>$G_2$</td>
</tr>
<tr>
<td>Bishop’s Wife</td>
<td>Koster</td>
<td>Grant</td>
<td>Comedy</td>
<td>$G_2$</td>
</tr>
<tr>
<td>Harvey</td>
<td>Koster</td>
<td>Stewart</td>
<td>Comedy</td>
<td>$G_2$</td>
</tr>
</tbody>
</table>

**Matching Coefficient:**
- $\text{Sim}(\text{Scorsese, Coppola}) = 0$;
- $\text{Sim}(\text{Koster, Hitchcock}) = \text{Sim}(\text{Koster, Coppola})$.

**Value Frequency Distribution:**
- $\text{Sim}(\text{Scorsese, Coppola}) < \text{Sim}(\text{Koster, Hitchcock})$

**Feature Dependency Aggregation:**
- $\text{Sim}(\text{Koster, Koster}) = \text{Sim}(\text{Scorsese, Coppola})$
Challenges: Categorical clustering vs. numerical clustering

• Much less research outcomes
• More difficult
  – Local vs. global perspective
  – Multiple aspects
• Global picture through
  – similarity,
  – dependency, and
  – dissimilarity
Our solution: Framework

Coupled Nominal Similarity

Integration

Value Frequency Distribution
Intra-coupled Similarity within an Attribute

Feature Dependency Aggregation
Inter-coupled Similarity between Attributes

2013/4/16
Basic concepts

• **intra-coupled/intra-coupling**
  – indicate the involvement of attribute value occurrence frequency within one attribute
  – how often the value occurs

• **inter-coupled/inter-coupling**
  – refer to the interaction of other attributes with this attribute
  – reflect the extent of the value difference brought by other attributes
Set information function

assigns the associated value set of attribute $a_j$ to the object set

relates each value of attribute $a_j$ to the corresponding object set

assigns a particular value of attribute $a_j$ to every object.

maps the value set of attribute $a_j$ to the dependent object set

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

$$f_j^*\left(\{u_{k_1}, \ldots, u_{k_t}\}\right) = \{f_j(u_{k_1}), \ldots, f_j(u_{k_t})\},$$

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

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

where $u_i, u_{k_1}, \ldots, u_{k_t} \in U$, and $V_j' \subseteq V_j$. 
Coupled Nominal Similarity

\[ f_2^*(\{u_1, u_2, u_3\}) = \{B_1, B_2\} \]

\[ g_2(B_1) = \{u_1, u_2\} \]

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

$$\delta_j^A(x, y) = \delta_j^{Ia}(x, y) \cdot \delta_j^{Ie}(x, y)$$  \hspace{1cm} (4.1)$$

where $\delta_j^{Ia}$ and $\delta_j^{Ie}$ are IaAVS and IeAVS, respectively.
Intra-coupled Interaction

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:

$$\delta_j^{Ia}(x, y) = \frac{|g_j(x)| \cdot |g_j(y)|}{|g_j(x)| + |g_j(y)| + |g_j(x)| \cdot |g_j(y)|}.$$  \hspace{1cm} (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.
\[ \delta_2^{I_a}(B_1, B_2) = \frac{|B_1| \times |B_2|}{|B_1| + |B_2| + |B_1| \times |B_2|} = \frac{2 \times 2}{2 + 2 + 2 \times 2} = 0.5 \]
Dissimilarity between two categorical values:

Modified Value Distance Matrix

The dissimilarity $D_{j|L}$ between two attribute values $x$ and $y$ for a specific attribute $a_j$ regarding labels $L$ is:

$$D_{j|L}(v^x_j, v^y_j) = \sum_{i \in L} |P_{L|j}(\{l\}|v^x_j) - P_{L|j}(\{l\}|v^y_j)|,$$

(4.3)

- $D_{j|L}$ indicates that two values are identified as being similar if they occur with the same relative frequency for all classes.

- MVDM takes into account the overall similarity of classification of all objects on each possible value of each attribute.

**Principle:** total variation dissimilarity and half of $L1$ dissimilarity are equivalent for categorical data distribution:

$$D_{j|L}(v^x_j, v^y_j) = 2 \cdot \max_{L' \subseteq L} |P_{L|j}(L'|v^x_j) - P_{L|j}(L'|v^y_j)|.$$

(4.4)
Inter-coupled Interaction

Modified Value Distance Matrix:

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

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

Object Co-occurrence Probability

\[ \delta_{jk}^P = \min_{V_k \subseteq V} \{ 2 - P_{k|j}(V_k|v_j^x) - P_{k|j}(V_k|v_j^y) \}, \quad (4.5) \]
\[ \delta_{jk}^U = 2 - \sum_{v_k \in V_k} \max \{ P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y) \}, \quad (4.6) \]
\[ \delta_{jk}^J = 2 - \sum_{v_k \in U} \max \{ P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y) \}, \quad (4.7) \]
\[ \delta_{jk}^I = \sum_{v_k \in \cap} \min \{ P_{k|j}(\{v_k\}|v_j^x), P_{k|j}(\{v_k\}|v_j^y) \}, \quad (4.8) \]
**Inter-coupled Interaction**

**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:

$$
\delta_{ij}^I(x, y) = \sum_{k=1, k\neq j}^{n} \alpha_k \delta_{j|k}(x, y),
$$

where $\alpha_k$ 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 $v^x_j$ and $v^y_j$ of attribute $a_j$ is:

$$
\delta^A_j(v^x_j, v^y_j, \{V_k\}_{k=1}^n) = \delta^{Ia}_j(v^x_j, v^y_j) \cdot \delta^{Ie}_j(v^x_j, v^y_j, \{V_k\}_{k \neq j}),
$$

(5.10)
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 $\text{CASO}(u_x, u_y)$:

$$\text{CASO}(u_x, u_y) = \sum_{j=1}^{n} \delta^A_j(v^x_j, v^y_j, \{V_k\}_{k=1}^n), \quad (7.1)$$
### TABLE 4
**Example of Computing Similarity Using IRSP**

| $V'_i$ | $V'_i$ | $P_{1/2}(V'_i|B_1)$ | $P_{1/2}(V'_i|B_2)$ | $2 - P_{1/2}(V'_i|B_1) - P_{1/2}(V'_i|B_2)$ |
|--------|--------|----------------------|----------------------|------------------------------------------|
| $\emptyset$ | $\{A_1, A_2, A_3, A_4\}$ | 0                    | 1                    | 1                                        |
| $\{A_1\}$ | $\{A_2, A_3, A_4\}$ | 0.5                  | 1                    | 0.5                                      |
| $\ldots$ | $\ldots$ | $\ldots$             | $\ldots$             | $\ldots$                                |
| $\{A_1, A_2, A_3, A_4\}$ | $\emptyset$ | 1                    | 0                    | 1                                        |

### TABLE 5
**Computing Similarity Using IRSU**

| $v_k$ | $P_{1/2}(\{v_k\}|B_1)$ | $P_{1/2}(\{v_k\}|B_2)$ | max |
|-------|-------------------------|-------------------------|-----|
| $A_1$ | 0.5                     | 0                       | 0.5 |
| $A_2$ | 0.5                     | 0.5                     | 0.5 |
| $A_3$ | 0                       | 0                       | 0   |
| $A_4$ | 0                       | 0.5                     | 0.5 |

### TABLE 6
**Computing Similarity Using IRSJ**

| $v_k$ | $P_{1/2}(\{v_k\}|B_1)$ | $P_{1/2}(\{v_k\}|B_2)$ | max |
|-------|-------------------------|-------------------------|-----|
| $A_1$ | 0.5                     | 0                       | 0.5 |
| $A_2$ | 0.5                     | 0.5                     | 0.5 |
| $A_3$ | 0                       | 0.5                     | 0.5 |

### TABLE 7
**Computing Similarity Using IRSI**

| $v_k$ | $P_{1/2}(\{v_k\}|B_1)$ | $P_{1/2}(\{v_k\}|B_2)$ | min |
|-------|-------------------------|-------------------------|-----|
| $A_2$ | 0.5                     | 0.5                     | 0.5 |

\[
\text{CASO}(u_2, u_3) = \sum_{j=1}^{3} \delta_{j}^3(v_j^2, v_j^3, \{v_k\}_{k=1}^3) = 0.5 + 0.125 + 0.125 = 0.75.
\]
Theoretical Analysis

- Computational Accuracy Equivalence:

Theorem 5.1. IRSP, IRSU, IRSJ and IRSI are all equivalent to one another.\(^2\)

IRSP $\iff$ IRSU $\iff$ IRSJ $\iff$ IRSI
### Computational Complexity Comparison:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Calculation Steps</th>
<th>Flops per Step</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRSP</td>
<td>$nR(R-1)/2$</td>
<td>$2(n-1)2^R$</td>
<td>$O(n^2R^22^R)$</td>
</tr>
<tr>
<td>IRSU</td>
<td>$nR(R-1)/2$</td>
<td>$2(n-1)R$</td>
<td>$O(n^2R^2R)$</td>
</tr>
<tr>
<td>IRSJ</td>
<td>$nR(R-1)/2$</td>
<td>$2(n-1)P$</td>
<td>$O(n^2R^2R)$</td>
</tr>
<tr>
<td>IRSI</td>
<td>$nR(R-1)/2$</td>
<td>$2(n-1)Q$</td>
<td>$O(n^2R^2R)$</td>
</tr>
</tbody>
</table>

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

**R**: The maximal number of attribute values.

IRSP $\geq$ IRSU $\geq$ IRSJ $\geq$ IRSI
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 $\text{CASO}(u_x, u_y)$.

begin

// Compute pairwise similarity for any two values of the same attribute.
for attribute $a_j$, $j = 1 : n$ do
  for every value pair $(v^x_j, v^y_j) \in [1, |V_j|])$ do
    $U_1 \leftarrow \{i | v^x_i = v^x_j\}$, $U_2 \leftarrow \{i | v^y_i = v^y_j\}$;
    // Compute intra-coupled similarity for two values $v^x_j$ and $v^y_j$.
    $\delta^I_{a_j}(v^x_j, v^y_j) = (|U_1| + |U_2|)/(|U_1||U_2|)$;
    // Compute coupled similarity for two attribute values $v^x_j$ and $v^y_j$.
    $\delta^A_j(v^x_j, v^y_j, \{V_k\}_{k=1}^n) \leftarrow$$\delta^I_{a_j}(v^x_j, v^y_j) \cdot \text{IeASV}(v^x_j, v^y_j, \{V_k\}_{k \neq j})$;
  end

// Compute coupled similarity between two objects $u_x$ and $u_y$.
$\text{CASO}(u_x, u_y) \leftarrow \text{sum}(\delta^A_j(v^x_j, v^y_j, \{V_k\}_{k=1}^n))$;

end

Function $\text{IeASV}(v^x_j, v^y_j, \{V_k\}_{k \neq j})$

begin

// Compute inter-coupled similarity for two attribute values $v^x_j$ and $v^y_j$.
for attribute $(k = 1 : n) \land (k \neq j)$ do
  $\{v^x_k\}_{k \neq j} \leftarrow \{v^x_k\}_{x \in U_1} \cap \{v^y_k\}_{y \in U_2}$;
  for intersection $z = U_3(1) : U_3(2)$ do
    $U_0 \leftarrow \{i | v^x_i = v^x_k\}$;
    $\text{ICP}_x \leftarrow |U_0 \cap U_1|/|U_1|$;
    $\text{ICP}_y \leftarrow |U_0 \cap U_2|/|U_2|$;
    $\text{Min}(x, y) \leftarrow \min(\text{ICP}_x, \text{ICP}_y)$;
    // Compute IRSI for $v^x_j$ and $v^y_j$.
    $\delta^I_{jk}(v^x_j, v^y_j, V_k) = \text{sum}(\text{Min}(x, y))$;
    $\delta^I_{e}(x, y) = \text{sum}[\alpha(k) \times \delta^I_{jk}(v^x_j, v^y_j, V_k)]$;
  end
return $\delta^I_{e}(v^x_j, v^y_j, \{V_k\}_{k \neq j})$;
end
Coupled Nominal Similarity:

$\text{Sim(Scorsese, Coppola)} = \text{Sim(Coppola, Coppola)} = 0.33$

$\text{Sim(Koster, Hitchcock)} = 0.25 \quad \text{Sim(Koster, Coppola)} = 0$

$\text{Sim(Koster, Koster)} = \text{Sim(Hitchcock, Hitchcock)} = 0.5$

Scorsese and Coppola are very similar directors

$\text{Sim(Koster, Hitchcock)} > \text{Sim(Koster, Coppola)}$

$\text{Sim(Scorsese, Coppola)} > \text{Sim(Koster, Hitchcock)}$

$\text{Sim(Koster, Koster)} > \text{Sim(Scorsese, Coppola)}$
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.
Fig. 1. Complexity on individual attributes.
Coupled Similarity Comparison

TABLE 10
Complexity Comparison on All Attributes

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Corral</th>
<th>Voting</th>
<th>LED</th>
<th>Lenses</th>
<th>Tic</th>
<th>Chess</th>
<th>Movie</th>
<th>Hayesroth</th>
<th>Molecular</th>
<th>Solar</th>
<th>Mushroom</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>$T$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>$n$</td>
<td>6</td>
<td>16</td>
<td>24</td>
<td>4</td>
<td>9</td>
<td>36</td>
<td>3</td>
<td>4</td>
<td>57</td>
<td>10</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>$\xi(U/P)$</td>
<td>50.0%</td>
<td>50.0%</td>
<td>50.0%</td>
<td>46.4%</td>
<td>37.5%</td>
<td>49.4%</td>
<td>27.8%</td>
<td>27.1%</td>
<td>25.0%</td>
<td>20.0%</td>
<td>1.7%</td>
<td>0.1%</td>
</tr>
<tr>
<td>$\xi(I/J)$</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>99.2%</td>
<td>82.3%</td>
<td>42.5%</td>
<td>48.4%</td>
</tr>
<tr>
<td>$</td>
<td>ICP(U)</td>
<td>$</td>
<td>120</td>
<td>960</td>
<td>2208</td>
<td>78</td>
<td>1296</td>
<td>5390</td>
<td>212</td>
<td>468</td>
<td>153216</td>
<td>2544</td>
</tr>
<tr>
<td>$</td>
<td>ICP(I)</td>
<td>$</td>
<td>120</td>
<td>960</td>
<td>2208</td>
<td>78</td>
<td>1296</td>
<td>4774</td>
<td>16</td>
<td>468</td>
<td>152022</td>
<td>1998</td>
</tr>
</tbody>
</table>

$m$ objects

$n$ attributes

T their minimal

R maximal number of values
the relative complexity ratios in terms of $\xi(I/U)$, $\xi(I/J)$, and $\xi(I/U)$
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.
Fig. 4. Clustering evaluation on six data sets.
Experimental findings

In summary, all of the above experiment results clearly show that IRSI outperforms IRSU, IRSJ, and IRSP in terms of the computational complexity, no matter how small or large, simple or complicated a data set is.

In particular, with the increase of the number of either features or attribute values, IRSI demonstrates superior efficiency compared to the others. IRSJ and IRSU follow, with IRSP being the most time-consuming, especially for the large-scale data set.
We draw the following two conclusions:

- Intra-coupled relative similarity IRSI is the most efficient one when compared with IRSP, IRSU and IRSJ, especially for large-scale data.

- Our proposed object dissimilarity metric COD is better than others, such as dependency aggregation only ADD, for categorical data in terms of clustering quality.
Conclusion

Coupled Similarity

Extension

Discretization
Clustering Ensemble
Numerical Coupling
Flexible Measures
Relation Analysis for Document Clustering
Coupled Term-Term Relation Analysis for Document Clustering

Xin Cheng, Can Wang, Duoqian Miao, Longbing Cao
7. Coupled Interaction Analysis

Document Representation

**THE SILVER CHAIR**

by C.S. Lewis

CHAPTER ONE
BEHIND THE GYM

IT was a dull autumn day and Jill Pole was crying behind the gym.

She was crying because they had been bullying her. This is not going to be a school story, so I shall say as little as possible about Jill’s school, which is not a pleasant subject. It was “Co-educational,” a school for both boys and girls, what used to be called a “mixed” school; some said it was not nearly so mixed as the minds of the people who ran it. These people had the idea that boys and girls should be allowed to do what they liked. And unfortunately what ten or fifteen of the biggest boys and girls liked best was bullying the others. All sorts of things, horrid things, went on which at an ordinary school would have been found out and stopped in half a term, but at this school they weren’t. Or even if they were, the people who did them were not expelled or punished. The Head said they were interesting psychological cases and sent for them and talked to them for hours. And if you know the right sort of things to say to the Head, the main result was that you became rather a favourite than otherwise.

---

**The BOW Model**

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>...</th>
<th>Tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.21</td>
<td>0.43</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.32</td>
<td>0</td>
<td>0</td>
<td>0.38</td>
</tr>
</tbody>
</table>
7. Coupled Interaction Analysis

Document Representation

• **TF-IDF** stands for *term frequency-inverse document frequency*
  
  M is the number of times that term t occurs in document d

• Term frequency (TF): \( tf(t, d) = \frac{M}{N} \)
  
  N is the number of words in this document

• Inverse document frequency (IDF): \( idf = \frac{|D|}{|D_i|} \)
  
  \( |D| \) the total number of documents in the corpus

• \( \text{Tfidf} = tf(t, d) \times idf(t, D) \)
  
  \( |D_i| \) the number of documents where the term appears
### 7. Coupled Interaction Analysis

**Similarity Measure**

<table>
<thead>
<tr>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>0</td>
<td>0.43</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.16</td>
<td>0</td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>0</td>
<td>0.26</td>
<td>0.17</td>
<td>0</td>
<td>0.19</td>
<td>0</td>
<td>0.21</td>
</tr>
</tbody>
</table>

**Similarity Measure**

\[
similarity(d_1, d_2) = \cos(\theta) = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1||\vec{d}_2|}
\]
7. Coupled Interaction Analysis

Document Similarity

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>ML</th>
<th>DB</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.25</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- 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.
7. Coupled Interaction Analysis

Term-Term Relation

The relation between terms

- To keep the original relation between terms, various document representation models have been proposed to capture the relation between terms based on statistical analysis.

\[ \mathbf{S}: \text{the relation matrix} \]

\[ d' = dS \]

- With \( \mathbf{S} \), the document representation is refined with the term-term relation information.

- Various kinds of \( \mathbf{S} \) lead to different extensions to the BOW model.
7. Coupled Interaction Analysis

**GVSM**

\[ d' = dS \]

The entry in matrix \( W^TW \) reflects the similarity between terms which is measured by their frequency of co-occurrence across the document set.

It means two terms are similar if they frequently co-occur in the same document.

**CVM-VSM, GTCV-VSM (similar to GVSM)**

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>0</td>
<td>0.43</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>d2</td>
<td>0</td>
<td>0</td>
<td>0.26</td>
<td>0.17</td>
<td>0</td>
<td>0.19</td>
<td>0</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>
Document Similarity based on GVSM

• These approaches have consistently performed better than the BOW model for document clustering.

• The cosine similarity between $d_1$ and $d_2$ based on the GVSM model is 0.522, $d_1$ and $d_3$ is 0.473.

• It is now easier to distinguish the similarity between $d_1$ and $d_2$ from that between $d_1$ and $d_3$, compared to the BOW model.
Our Approach

7. Coupled Interaction Analysis

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>ML</th>
<th>DB</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.25</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

• Those approaches estimate the similarity between terms (e.g. “Data mining” and “Computer science”) by their co-occurrence in a simple way, while the implicit relation between terms (e.g. “Data mining” and “Machine learning”) is overlooked.

• They fail to capture the complete semantic relation between terms.
Our Approach

• We propose a novel approach to measure the relationship between terms by capturing both the \textit{intra-relation (explicit)} and \textit{inter-relation (implicit)}.

• The \textit{intra-relation} adopts the original co-occurrence based approaches,

• The \textit{inter-relation}, which has been overlooked, is also teased out to derive the complete description of the semantic relation between terms.
Terms are relational if they co-occur in the same document.

- Terms $ti$ and $tk$ co-occur in document $dx$, while $tj$ is the co-occurrence term of $tk$ in document $dy$.

- Then, term $ti$ is considered to be associated with $tk$ in document $dx$, and term $tj$ is related with $tk$ in document $dy$. 
• We adapt the popular co-occurrence measure Jaccard to evaluate the relation rather than simply considering the inner product of them

\[
CoR(t_i, t_j) = \frac{1}{|H|} \cdot \sum_{x \in H} \frac{w_{xi}w_{xj}}{w_{xi} + w_{xj} - w_{xi}w_{xj}},
\]

• We define the intra-relation between terms in a conditional probability manner, which captures the probability of a term that co-occurs with another observed term

\[
IaR(t_i, t_j) = \begin{cases} 
1 & i = j, \\
\frac{CoR(t_i, t_j)}{\sum_{t=1, t \neq i}^{n} CoR(t_i, t_j)} & i \neq j,
\end{cases}
\]
Exploiting the co-occurrence of terms helps to discover the explicit relation between terms.

The intra-relation matrix is shown in the Table.

However, it lacks the ability to reveal the underlying relation other than co-occurrence frequency.
In our paper, we introduce a novel approach to capture this kind of underlying relation (e.g. between “Machine learning" and “Data mining").

The inter-relation analysis is inspired by the fact that terms with the similar meaning likely appear in a similar context.

Therefore, we explore the relation between a pair of terms by their context, which is captured by their interaction with other terms across the entire document set.
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 $I_aR(t_k, t_i) > 0$ and $I_aR(t_k, t_j) > 0$ 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:

$$R_{IeR}(t_i, t_j | t_k) = \min(I_aR(t_i, t_k), I_aR(t_j, t_k)),$$

(3.3)
7. Coupled Interaction Analysis

Inter-relation

Definition 4. The inter-relation between two terms $t_i$ and $t_j$ is defined by their interaction with all the link terms, formalized as:

$$I_{rR}(t_i, t_j) = \begin{cases} 
0 & i = j \\
\frac{1}{|L|} \sum_{\forall t_k \in L} R_{I_{rR}}(t_i, t_j | t_k) & i \neq j
\end{cases}$$

(3.4)

where $|L|$ denotes the number of link terms in $L = \{t_k | (I_{aR}(t_k, t_i) > 0) \land (I_{aR}(t_k, t_j) > 0)\}$, and $R_{I_{rR}}(t_i, t_j | t_k)$ is the relative inter-relation between $t_i$ and $t_j$ linked by $t_k$. If $L = \emptyset$, we define $I_{rR}(t_i, t_j) = 0$. 
For instance, the relation between "Machine Learning" and "Data mining" is captured by using Equation (3.3).

The definition indicates that the larger the average relative inter-relations with link terms, the closer a pair of terms are inter-related.

In this way, the relation between terms that have link terms is enhanced.

Accordingly, the implicit relation revealed by the link terms makes the related documents more alike, which facilitates the clustering of documents.
7. Coupled Interaction Analysis

Coupled Relation

Coupled Relation between terms

**Definition 5.** Given a pair of terms \( t_i \) and \( t_j \) in \( D \), the coupled relation (CR) between \( t_i \) and \( t_j \) is defined as

\[
CR(t_i, t_j) = \begin{cases} 
1 & \text{if } i = j \\
\alpha \cdot IaR(t_i, t_j) + (1 - \alpha) \cdot IeR(t_i, t_j) & \text{if } i \neq j
\end{cases}
\]  

(3.5)

where \( \alpha \in [0, 1] \) is the parameter that decides the weight of intra-relation, \( IaR(t_i, t_j) \) and \( IeR(t_i, t_j) \) are the respective intra-relation and inter-relation between terms \( t_i \) and \( t_j \).

\[
\begin{array}{c|cccc}
 & DM & ML & DB & CS \\
\hline
DM & 1.00 & 0.24 & 0.25 & 0.26 \\
ML & 0.24 & 1.00 & 0.25 & 0.24 \\
DB & 0.22 & 0.24 & 1.00 & 0.25 \\
CS & 0.40 & 0.39 & 0.33 & 1.00 \\
\end{array}
\]
### Table 1. An Example of Document Representation: “DM”, “ML”, “DB” and “CS” denote “Data mining”, “Machine learning”, “Database” and “Computer science”, respectively.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>ML</th>
<th>DB</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.25</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

### Table 2. An Example of the Enriched Document Representation with CR ($\alpha = 0.5$)

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>ML</th>
<th>DB</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0.64</td>
<td>0.26</td>
<td>0.32</td>
<td>0.46</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0.24</td>
<td>0.62</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0.22</td>
<td>0.23</td>
<td>0.83</td>
<td>0.30</td>
</tr>
</tbody>
</table>
D1, D2, and D3 are the respective subsets of 20 Newsgroups, R8 and R52, and D4 is the WebKB benchmark document collection.

Before conducting the document representation, all data sets are pre-processed to apply word stemming. We also discard the documents that are less than 10 words which means that there is less information for document clustering.

### Table 3. Characteristics of Data Sets

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Classes</th>
<th>( m )</th>
<th>( n )</th>
<th>( n_{avg} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>5</td>
<td>1864</td>
<td>16516</td>
<td>76</td>
</tr>
<tr>
<td>D2</td>
<td>8</td>
<td>2091</td>
<td>8674</td>
<td>33</td>
</tr>
<tr>
<td>D3</td>
<td>52</td>
<td>2448</td>
<td>9728</td>
<td>36</td>
</tr>
<tr>
<td>D4</td>
<td>4</td>
<td>4087</td>
<td>7769</td>
<td>32</td>
</tr>
</tbody>
</table>
Performance Evaluation

The quality of document clustering is evaluated by four criteria: Purity, Rand Index (RI), \( F_1 \)-measure and Normalized mutual information (NMI).

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>Purity BOW</th>
<th>Purity GVSM</th>
<th>Purity CRM</th>
<th>RI BOW</th>
<th>RI GVSM</th>
<th>RI CRM</th>
<th>( F_1 )-measure BOW</th>
<th>( F_1 )-measure GVSM</th>
<th>( F_1 )-measure CRM</th>
<th>NMI BOW</th>
<th>NMI GVSM</th>
<th>NMI CRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.79</td>
<td>0.82</td>
<td>0.84</td>
<td>0.49</td>
<td>0.57</td>
<td>0.62</td>
<td>0.48</td>
<td>0.58</td>
<td>0.61</td>
<td>0.32</td>
<td>0.41</td>
<td>0.44</td>
</tr>
<tr>
<td>D2</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>0.49</td>
<td>0.58</td>
<td>0.60</td>
<td>0.62</td>
<td>0.66</td>
<td>0.67</td>
<td>0.44</td>
<td>0.48</td>
<td>0.44</td>
</tr>
<tr>
<td>D3</td>
<td>0.78</td>
<td>0.79</td>
<td>0.82</td>
<td>0.27</td>
<td>0.29</td>
<td>0.36</td>
<td>0.41</td>
<td>0.43</td>
<td>0.50</td>
<td>0.37</td>
<td>0.42</td>
<td>0.39</td>
</tr>
<tr>
<td>D4</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.63</td>
<td>0.65</td>
<td>0.63</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
<td>0.32</td>
<td>0.35</td>
<td>0.37</td>
</tr>
</tbody>
</table>

- In comparison with the BOW model, our approach achieve 3\% improvement on the average Purity score of four data sets, and the improvement in RI is around 22\%, 12\% for the \( F_1 \)-measure and 14\% for the NMI.

- Compared with the GVSM model, our approach also achieves 1.5\%, 8\%, 5.5\% and on the average Purity, RI and \( F_1 \)-measure scores, respectively.
Performance Evaluation

Comparisons of the number of discovered relations on four data sets:

- GVSM brings a lot of relations into consideration for clustering, and it achieves the performance improvement comparing with BOW.

- It does not consider the underlying relation which have proven useful in measuring document similarity.
In our approach, we consider both the intra-relation and inter-relation between terms, which discovers more relations between terms.

In other words, it integrates more semantic information into document representation. That is why it achieves the best performance in comparison to the other two models.
Conclusion

• In this paper, we present a novel approach to capture the coupled relation between terms to improve the performance of document clustering.

• Based on the combination of intra-relation and inter-relation, our approach provides more semantic information in each document for clustering.

• The experiment result shows that it can significantly improve the performance of document clustering.
Coupled Clustering Ensemble
Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects

Can Wang, Zhong She, Longbing Cao
Advanced Analytics Institute, University of Technology, Sydney, Australia
{canwang613,zhong2024,longbing.cao}@gmail.com
Outline

• **Introduction**
• Related Work
• Coupled Framework of Clustering Ensembles (CCE)
• Coupled Relationship in CCE
• Coupled Consensus Function in CCE
• Algorithm and Analysis
• Empirical Study
• Conclusion
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.
**Introduction**

**Clustering ensemble**
- largely captures the common structure of the base clusterings
- achieves a combined clustering with better quality than that of individual clusterings

It also faces several issues that have not been explored well in the consensus design.

*An example*

Fig. 1. Four possible base clusterings of 12 data objects into two clusters, different partitions use different sets of labels.
Possible cluster labels based on four base clusterings

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

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.
Introduction

The reason is that the similarity defined here is too limited to reveal the complete hidden relationship 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.
Introduction

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.
The number of common neighbors of objects $u_2$ and $u_3$ is much larger than that of $u_2$ and $u_{10}$. How to distinguish them and assign the correct label to each object?

Just Coupled Relationship between Base Clusterings

Not Enough

Discover the Discrepancy on the Common Neighborhood Domains

The number of common neighbors of objects $u_2$ and $u_3$ is much larger than that of $u_2$ and $u_{10}$. $u_2$ and $u_3$ are more probable in the same cluster and $u_{10}$ is in another one, which is the genuine partition.
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 relation with other data objects?

- **Integrated Coupling**: If there are interactions between both clusterings and objects, then how to integrate such couplings in clustering ensemble?
Outline

• Introduction
• **Related Work**
• Coupled Framework of Clustering Ensembles (CCE)
• Coupled Relationship in CCE
• Coupled Consensus Function in CCE
• Algorithm and Analysis
• Empirical Study
• Conclusion
Related Work

Process of Clustering Ensemble

1. Building Base Clusterings
2. Aggregating Base Clusterings
3. Post-processing Clustering
Related Work

Building Base Clusterings

- Random initializations
- Data sampling or re-sampling
- Random projections and random hyperplane splits
- Diverse clustering algorithms
Post-processing Clustering

- partition-based: k-means, PAM.
- hierarchy-based: single linkage
- spectrum-based: SPEC
- graph-based: METIS
Related Work

Aggregating Base Clusterings

• **Consensus function**
  It focuses on the total agreements of all the base clusterings from different perspectives.

• **Categorical clustering**
  Clustering ensemble can also be converted to the problem of clustering categorical data by viewing each attribute as a way of producing a base clustering of the data.

• **Direct optimization**
  It performs on the original objective function of clustering rather than exploring the consent among partial solutions.
Related Work

Consensus Functions

Here, we are interested in the branch of consensus functions when aggregating base clusterings.

The consensus functions seek a combination of multiple base clusterings to provide a prior superior input for the post-processing clustering.
Related Work

Consensus Functions

We can construct the consensus functions by following several approaches: the direct best matching, graph-based mappings, statistical mixture models, and pairwise comparisons, etc.

Based on

Co-associations or pairwise agreements:

- between clusterings (e.g., partition difference, EM, and QMI)
- between data objects (e.g., CSPA and ROCK),
- between clusters (e.g., MCLA, HBGF)
Outline

• Introduction
• Related Work
• **Coupled Framework of Clustering Ensembles (CCE)**
  • Coupled Relationship in CCE
  • Coupled Consensus Function in CCE
• Algorithm and Analysis
• Empirical Study
• Conclusion
Fig. 3. A coupled framework of clustering ensembles (CCE), where ←→ indicates the intra-coupling and ↔ refers to the inter-coupling.
**Coupled Framework of Clustering Ensembles**

**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
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
Coupled Framework of Clustering Ensembles

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.
• Introduction
• Related Work
• Coupled Framework of Clustering Ensembles (CCE)
• **Coupled Relationship in CCE**
• Coupled Consensus Function in CCE
• Algorithm and Analysis
• Empirical Study
• Conclusion
Coupled Relationship 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 $v^x_j$ and $v^y_j$ of base clustering $bc_j$ is:

$$
\delta^{IaC}(v^x_j, v^y_j) = \frac{|g_j(v^x_j)| \cdot |g_j(v^y_j)|}{|g_j(v^x_j)| + |g_j(v^y_j)| + |g_j(v^x_j)| \cdot |g_j(v^y_j)|}
$$

(V.1)

where $g_j(v^x_j)$ and $g_j(v^y_j)$ are the set information functions.

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

The set of objects whose cluster labels is $v^y_j$ in base clustering $bc_j$.
Coupled Relationship in CCE

Coupling of Clusterings

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

\[ \delta_{j|k}(v^x_j, v^y_j|V_k) = \sum_{v_k \in \cap} \min\{P_{k|j}(v_k|v^x_j), P_{k|j}(v_k|v^y_j)\} \]

where \( v_k \in \cap \) denotes \( v_k \in \varphi_{j \rightarrow k}(v^x_j) \cap \varphi_{j \rightarrow k}(v^y_j) \), \( \varphi_{j \rightarrow k} \) is the inter-information function, and \( P_{k|j} \) is the information conditional probability formalized in Equation (III.1).
Coupled Relationship in CCE

Coupling of Clusterings

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

\[ \delta^I_{j} (v^x_j, v^y_j | \{V_k\}_{k \neq j}) = \sum_{k=1, k \neq j}^{L} \lambda_k \delta^I_j (v^x_j, v^y_j | V_k), \quad (V.3) \]

where \( \lambda_k \) 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 \( \delta^I_j (v^x_j, v^y_j | V_k) \) is IeRSC.
Coupled Relationship 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:

\[
\delta_j^C(v_x^j, v_y^j | \{V_k\}_{k=1}^L) = \delta_j^{IaC}(v_x^j, v_y^j) \cdot \delta_j^{IeC}(v_x^j, v_y^j | \{V_k\}_{k \neq j}),
\]

where $\delta_j^{IaC}$ and $\delta_j^{IeC}$ are IaCSC and IeCSC, respectively.

*Definition 5.4: (CCSC) The Coupled Clustering Similarity for Clusters between cluster labels $v_x^j$ and $v_y^j$ of clustering $bc_j$ is:*
**Coupled Relationship in CCE**

**Coupling of Clusterings**

**TABLE I**

**AN EXAMPLE OF BASE CLUSTERINGS**

<table>
<thead>
<tr>
<th>U</th>
<th>bc₁</th>
<th>bc₂</th>
<th>bc₃</th>
<th>bc₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>u₁</td>
<td>2</td>
<td>B</td>
<td>X</td>
<td>β</td>
</tr>
<tr>
<td>u₂</td>
<td>2</td>
<td>A</td>
<td>X</td>
<td>α</td>
</tr>
<tr>
<td>u₃</td>
<td>2</td>
<td>A</td>
<td>Y</td>
<td>β</td>
</tr>
<tr>
<td>u₄</td>
<td>2</td>
<td>B</td>
<td>X</td>
<td>β</td>
</tr>
<tr>
<td>u₅</td>
<td>1</td>
<td>A</td>
<td>X</td>
<td>β</td>
</tr>
<tr>
<td>u₆</td>
<td>2</td>
<td>A</td>
<td>Y</td>
<td>β</td>
</tr>
<tr>
<td>u₇</td>
<td>2</td>
<td>B</td>
<td>Y</td>
<td>α</td>
</tr>
<tr>
<td>u₈</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>α</td>
</tr>
<tr>
<td>u₉</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>β</td>
</tr>
<tr>
<td>u₁₀</td>
<td>1</td>
<td>A</td>
<td>Y</td>
<td>α</td>
</tr>
<tr>
<td>u₁₁</td>
<td>2</td>
<td>B</td>
<td>Y</td>
<td>α</td>
</tr>
<tr>
<td>u₁₂</td>
<td>1</td>
<td>B</td>
<td>Y</td>
<td>α</td>
</tr>
</tbody>
</table>

Fig. 4. An example of the coupled similarity for cluster labels α and β, where ←→ indicates the intra-coupling and ↔ refers to the inter-coupling, the value along each line is the corresponding similarity.
Coupled Relationship 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:

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

where $\delta^C(v^x_j, v^y_j, \{V_k\}_{k=1}^L)$ refers to CCSC between cluster labels $v^x_j$ and $v^y_j$ of base clustering $bc_j$. 
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:

$$
\delta^{Sim}(u_x, u_y) \geq \theta,
$$

(V.6)

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

The neighbor set of object $u_x$: 

$$
N_{u_x} = \{u_z | \delta^{Sim}(u_x, u_z) \geq \theta\}.
$$
Intuitively, objects $u_x$ and $u_y$ more likely belong to the same cluster if they have a larger overlapping in their neighbor sets $N_{ux}$ and $N_{uy}$. 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 $U$.

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

where $N_{ux}^{Sim}$ and $N_{uy}^{Sim}$ are the neighbor sets of objects $u_x$ and $u_y$ based on $\delta^{Sim}$, respectively.
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 $\delta_{\text{Sim}}$ in (V.7) to be $\delta_{\text{IaOSO}}$ $\delta_{\text{IaO}}$ 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 $\delta_{\text{Sim}}$ is in particular regarded as $\delta_{\text{IaO}}$. Specifically:

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

where sets of objects $N_{u_x}^{IaO} = \{u_z | \delta_{\text{IaO}}(u_x, u_z) \geq \theta\}$ and $N_{u_y}^{IaO} = \{u_z | \delta_{\text{IaO}}(u_y, u_z) \geq \theta\}$. 
Coupled Relationship in CCE

Coupling of Objects

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

- Introduction
- Related Work
- Coupled Framework of Clustering Ensembles (CCE)
- Coupled Relationship in CCE
- Coupled Consensus Function in CCE
- Algorithm and Analysis
- Empirical Study
- Conclusion
Coupled Consensus Function in CCE
The **clustering-based consensus function** captures the pairwise agreement between base clusterings.

A common measure of discrepancy is partition difference:

\[
S_{Cg}(bc_{j_1}, bc_{j_2}) = \sum_{1 \leq x, y \leq m} [BC_{j_1}(x, y) - BC_{j_2}(x, y)]^2.
\]

Intuitively, \( S_{Cg} \) computes the sum of similarity between objects that belong to different base clusterings \( bc_{j_1} \) and \( bc_{j_2} \). A target clustering \( fc^* \) thus should be:

\[
fc^* = \arg \min_{c^1, \ldots, c^k} \sum_{j=1}^{L} S_{Cg}(fc, bc_j)
\]
The usual way:

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

Our proposed way CgC:

\[ BC_j^C (x, y) = \delta^C_j (v_j^x, v_j^y | \{V_k\}_{k=1}^N) \]

\[ S_{C_g}(bc_{j_1}, bc_{j_2}) = \sum_{1 \leq x, y \leq m} [BC_j^1 (x, y) - BC_j^2 (x, y)]^2 \]

Coupled Clustering Similarity for Clusters:

\[ \delta^C_j (v_j^x, v_j^y | \{V_k\}_{k=1}^N) = \delta^{IaC}_j (v_j^x, v_j^y) \cdot \delta^{IeC}_j (v_j^x, v_j^y | \{V_k\}_{k \neq j}) \]
The object-based consensus function squeezes out the co-associations between objects

A simple and obvious heuristic to describe the similarity between $u_x$ and $u_y$ is the entry-wise average of the L associated similarity matrices induced by the L base clusterings.

$$BC^*(x, y) = \frac{1}{L} \cdot \sum_{j=1}^{L} BC_j(x, y)$$

- The pairwise similarity measure $BC_j^N (x, y)$ is rather rough since only 1 and otherwise 0 are considered.

- Neither relationship within nor between base clusterings (i.e., attributes $bc_{j1}$ and $bc_{j2}$) is explicated.

- Most existing work only uses the similarity measure between objects when clustering them, it thus, does not involve the context (i.e., neighborhood) of the objects.
To solve the first two issues, we could regard the entry $BC^*(x, y)$ of the overall similarity matrix to be “The Intra-coupled Object Similarity for Objects (IaOSO)” to capture the **intra-coupled interactions within two objects** as well as both the **intra-coupled and inter-coupled Interactions among base clusterings**.

Our proposed way 1 OC-Ia:

$$S_{IaO}^{IaC}(u_x, u_y) = BC^*(x, y) = \delta^{IaO}$$

The Intra-coupled Object Similarity for Objects:

$$\delta^{IaO}(u_x, u_y) = \frac{1}{L} \cdot \sum_{j=1}^{L} \delta^C_j(v^x_j, v^y_j | \{V_k\}_{k=1}^n)$$
Object-based Coupling

Further taking into account the third issue above, we incorporate both the intra- and inter- coupling of clusterings and objects.

Our proposed way 2 OC-H:

\[ S_O^C(u_x, u_y) = \delta^{CO}(u_x, u_y | U) \]

Accordingly, the desired final clustering \( f_c^* \) with \( k^* \) clusters can be obtained by maximizing the following criterion function:

\[
f_{c^*} = \arg \max_{c^1, \ldots, c^{t^*}} \sum_{t=1}^{t^*} m_t \cdot \sum_{u_x, u_y \in c^t} \frac{S_O^C(u_x, u_y) \cdot m}{m_{t}^{1+2f(\theta)}}
\]
The cluster-based consensus function characterizes the pairwise interactions between clusters.

One of the basic approaches based on the relationship between clusters is MCLA proposed by Strehl and Ghosh (Jaccard matching coefficient):

\[
S^N_Cr(c_{j1}^{i1}, c_{j2}^{i2}) = \frac{|c_{j1}^{i1} \cap c_{j2}^{i2}|}{|c_{j1}^{i1} \cup c_{j2}^{i2}|}
\]

The above similarity measure \(S^N_Cr\) considers neither coupling between base clusterings nor interaction between objects. Therefore, it is in lack of the capability to reflect the essential link and relationship among data.

Our proposed way CrC (CrC-Ia, CrC-C):

\[
S^{Cave}_C(c_{j1}^{i1}, c_{j2}^{i2}) = \frac{1}{k_{j1}k_{j2}} \sum_{u_x \in c_{j1}^{i1}, u_y \in c_{j2}^{i2}} S_O(u_x, u_y)
\]
How to Generate Base Clusterings: There are several existing methods to provide diverse base clusterings: using different clustering algorithms, employing random or different parameters of some algorithms, and adopting random subsampling or random projection of the data. Since our focus is mainly on the consensus function, we use $k$-means on random subsampling [8] of the data as the base clustering algorithm in our experiments. The number $t^j$ of base clustering $bc_j$ is pre-defined for each data set and remains the same for all clustering runs.
How to Post-process Clustering: In the proposed CCE framework, we mainly focus on the consensus function based on pairwise interactions between base clusterings, between objects and between clusters. Those interactions are described by the corresponding similarity matrices. Thus, a common and recommended way to combine the base clusterings is to recluster the objects using any reasonable similarity-based clustering algorithm. In our experiments, we choose $k$-means, agglomerative algorithm [6] and METIS [2] due to their popularity in clustering ensemble.
Outline

• Introduction
• Related Work
• Coupled Framework of Clustering Ensembles (CCE)
• Coupled Relationship in CCE
• Coupled Consensus Function in CCE
• Algorithm and Analysis
• Empirical Study
• Conclusion
Algorithm and Analysis

**Algorithm 1: Coupled Similarity for Clusters CCSC()**

**Data:** Object set $U = \{u_1, \ldots, u_m\}$ and $u_x, u_y \in U$, base clustering set $C = \{bc_1, \ldots, bc_L\}$, and weight $\lambda = (\lambda_k)_{k \in n}$.

**Result:** Similarity $CCSC(v^x_j, v^y_j)$ between clusters $v^x_j, v^y_j$.

1. begin
   2. maximal cluster label $r(j) \leftarrow \max(\{c^x_j, \ldots, c^y_j\})$
   3. for every cluster label pair $(v^x_j, v^y_j) \in [1, r(j)]$ do
      4. $U_1 \leftarrow \{i|v^x_i = v^x_j\}$, $U_2 \leftarrow \{i|v^y_i = v^y_j\}$
         // Compute intra-coupled similarity for two cluster labels $v^x_j$ and $v^y_j$.
      5. $\delta^{i \rightarrow C}_{j}(v^x_j, v^y_j) = (|U_1||U_2|)/(|U_1| + |U_2| + |U_1||U_2|)$
      6. $\delta^{\rightarrow j}_{j}(v^x_j, v^y_j|\{V_k\}_{k=1}^L) \leftarrow \delta^{i \rightarrow j}_{j}(v^x_j, v^y_j) \cdot IeCSC(v^x_j, v^y_j)$
      7. $CCSC(v^x_j, v^y_j) \leftarrow \delta^{\rightarrow j}_{j}(v^x_j, v^y_j|\{V_k\}_{k=1}^L)$
   end

9. Function $IeCSC(v^x_j, v^y_j)$

10. begin
    11. for each base clustering $(bc_k \in C) \wedge (bc_k \neq bc_j)$ do
        12. $\varphi \leftarrow \{v^x_k|x \in U_1\} \cap \{v^y_k|y \in U_2\}$
        13. for every intersection $v^x_k \in \varphi$ do
            14. $U_0 \leftarrow \{i|v^x_k = v^x_i\}$
            15. $ICP_x \leftarrow |U_0 \cap U_1|/|U_1|$
            16. $ICP_y \leftarrow |U_0 \cap U_2|/|U_2|$
            17. $\text{Min}_{(x,y)} \leftarrow \min(I_{CP_x}, I_{CP_y})$
            18. $\delta_{L;k}(v^x_j, v^y_j|V_k) = \text{sum}(\text{Min}_{(x,y)})$
        end
        // Compute inter-coupled similarity for two cluster labels $v^x_j$ and $v^x_k$.
    19. $\delta^{\rightarrow jL,k}_{j}(v^x_j, v^y_j|V_k) \leftarrow \text{sum}(\lambda_k \cdot \delta_{L;k}(v^x_j, v^y_j|V_k))$
    20. return $IeCSC(v^x_j, v^y_j) = \delta^{\rightarrow jL,k}_{j}(v^x_j, v^y_j|\{V_k\}_{k \neq j})$

**Algorithm 2: Coupled Similarity for Objects CCOSO()**

**Data:** Object set $U = \{u_1, \ldots, u_m\}$ and $u_x, u_y \in U$, base clustering set $C = \{bc_1, \ldots, bc_L\}$, and threshold $\theta \in [0,1]$.

**Result:** Similarity $CCOSO(u_x, u_y)$ between objects $u_x, u_y$.

1. begin
   2. for each base clustering $bc_j \in C$ do
      3. $\delta^{C}_{j}(v^x_j, v^y_j|\{V_k\}_{k=1}^L) \leftarrow CCSC(v^x_j, v^y_j)$
         // Compute intra-coupled similarity for two objects $u_x$ and $u_y$.
      4. $\delta^{IAO}(u_x, u_y) = 1/L \cdot \text{sum}[\delta^{C}_{j}(v^x_j, v^y_j|\{V_k\}_{k=1}^L)]$
      5. neighbor sets $N_{ux} = N_{uy} = \emptyset$
   6. for objects $(u_{x1}, u_{z2} \in U) \wedge (u_{z1} \neq u_x) \wedge (u_{z2} \neq u_y)$ do
      7. if $\delta^{IAO}(u_x, u_{z1}) \geq \theta$ then
         8. $N_{ux} = \{u_{z1}\} \cup N_{ux}$
      9. if $\delta^{IAO}(u_y, u_{z2}) \geq \theta$ then
         10. $N_{uy} = \{u_{z2}\} \cup N_{uy}$
         // Compute inter-coupled similarity for two objects $u_x$ and $u_y$.
      11. $\delta^{CO}(u_x, u_y|u_{z2}) = 1/m \cdot |N_{ux} \cap N_{uy}|$
      12. $CCOSO(u_x, u_y) \leftarrow \delta^{CO}(u_x, u_y|u_{z2})$
   end

---

THE ADVANCED ANALYTICS INSTITUTE
Outline

- Introduction
- Related Work
- Coupled Framework of Clustering Ensembles (CCE)
- Coupled Relationship in CCE
- Coupled Consensus Function in CCE
- Algorithm and Analysis
- Empirical Study
- Conclusion
The experimental evaluation is conducted on eight data sets, including two synthetic data sets (i.e., Sy1 and Sy2, which are 2-Gaussian and 4-Gaussian with Noise, respectively) and nine real-life data sets from UCI.
Selection of Parameters

\[ \delta^{Sim}(u_x, u_y) \geq \theta \]

- \( \theta \): The neighbor threshold in (4.6) is defined to be the average \( IaOCO \) value (i.e., \( \delta^{IaO} \)) of pairwise objects \( u_x \) and \( u_y \).

- \( L \): The ensemble size (i.e., the number of base clusterings) is taken to be \( L = 10 \).

- \( k_j, k^* \): The number of clusters in the base clustering \( bc_j \) and final clustering \( fc^* \) are both regarded as the number of pre-known classes \( k_j = k^* = k \).

- \( \lambda_k \): The weight \( \lambda_k \) for base clustering \( bc_k \) in Definition 4.3 on \( IeCSC \) is simplified as \( \lambda_k = 1/L = 1/10 \).

- \( R \): The number of runs for each clustering ensemble is fixed to be \( R = 50 \) to obtain an average result.
Selection of Algorithms

Clustering-based:
  CgC+k-means VS PD(2009), EM(2005), QMI(2005), base clustering

Object-based:
  OC-H+AA

Cluster-based:
  CrC-Ia+Metis VS MCLA(02), HBGF(04), LB-Pam(11), LB-Spec(11), base clustering
  CrC-C+Metis
Experiment Results

### Table IV

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Accuracy (AC)</th>
<th>Normalized Mutual Information (NMI)</th>
<th>Combined Stability Index (CSI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Avg</td>
<td>Min</td>
</tr>
<tr>
<td>Sy1</td>
<td>0.955</td>
<td>0.950</td>
<td>0.945</td>
</tr>
<tr>
<td>Sy2</td>
<td>0.503</td>
<td>0.460</td>
<td>0.385</td>
</tr>
<tr>
<td>Iris</td>
<td>0.927</td>
<td>0.827</td>
<td>0.513</td>
</tr>
<tr>
<td>Wine</td>
<td>0.708</td>
<td>0.689</td>
<td>0.556</td>
</tr>
<tr>
<td>Seg</td>
<td>0.586</td>
<td>0.529</td>
<td>0.433</td>
</tr>
<tr>
<td>Glass</td>
<td>0.517</td>
<td>0.479</td>
<td>0.449</td>
</tr>
<tr>
<td>Ecoli</td>
<td>0.687</td>
<td>0.512</td>
<td>0.470</td>
</tr>
<tr>
<td>Ionos</td>
<td>0.712</td>
<td>0.704</td>
<td>0.650</td>
</tr>
<tr>
<td>Blood</td>
<td>0.739</td>
<td>0.709</td>
<td>0.707</td>
</tr>
<tr>
<td>Vowel</td>
<td>0.373</td>
<td>0.354</td>
<td>0.339</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.384</td>
<td>0.332</td>
<td>0.319</td>
</tr>
</tbody>
</table>

Accuracy (AC), normalized mutual information (NMI), and combined stability index (CSI). **The larger, the better.**
Fig. 5. Clustering-based comparisons.
Fig. 6. Object-based comparisons.
### TABLE V
Cluster-based Comparisons on AC, NMI and CSI

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

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**.
Group Behaviour Analysis
• Short-term manipulation behaviors as cause
CHMM Based Coupled Sequence Modeling

- Coupled behavior sequences
  - Multiple sequences
    \[ \Phi_1 = \{\phi_{11}, \ldots, \phi_{1T}\} \]
    \[ \Phi_2 = \{\phi_{21}, \ldots, \phi_{2F}\} \]
    \[ \Phi_C = \{\phi_{C1}, \ldots, \phi_{CG}\} \]
  - Coupling relationship
    \[ R_{ij}(\Phi_i, \Phi_j) \]
    \[ R_{ij} \subset R, R_{ij}(\Phi_i, \Phi_j) = \emptyset \]
  - Behavior properties
    \[ \phi_{ik}(p_{ik,1}, \ldots, p_{ik,L}) \]
(b) The Structure of the CHMM

CBA problem $\rightarrow$ CHMM model

\[
\Phi(B_c) | \text{category} \rightarrow X \tag{15}
\]

\[
M(\Phi(B_c)) | \phi_{ik}([p_{ij}]_1, \ldots, [p_{ij}]_K) \rightarrow Y \tag{16}
\]

\[
f(\theta(\cdot), \eta(\cdot)) \rightarrow Z \tag{17}
\]

Initial distribution of $\Phi(B_c) | \text{category} \rightarrow \pi \tag{19}$
CBA - Conditional Probability Distribution

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

<table>
<thead>
<tr>
<th>$X^{(t)}$</th>
<th>$RF_1$</th>
<th>$RF_2$</th>
<th>$RF_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$trade_1$</td>
<td>$x_1$</td>
<td>$r_{f11}$</td>
<td>$r_{f21}$</td>
</tr>
<tr>
<td>$trade_2$</td>
<td>$x_2$</td>
<td>$r_{f12}$</td>
<td>$r_{f22}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>

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

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

\[
CL(b^k) = \prod_{b_i^{(s)} \in b^k} p(X^{(s)}) = x_{b_i^{(s)}} | r_{f1s}, r_{f2s}, \cdots, r_{fn; s}; M)
\]

CBA problem $\rightarrow$ SRL Modeling

\[
f(\theta(\cdot), \eta(\cdot)) \rightarrow \text{the CPD } p(X^{(t)}|RF_1, \cdots, RF_n)
\]
Rule Interaction Analysis
Best rule unions analysis
Single rule evaluation

**Single Rule Performance (Profit vs FPR, unique FPR) - [Mar 2010]**

- **FPR**: Lower, Better
- **DR**: Higher, Better
## Cross fired rule group evaluation

<table>
<thead>
<tr>
<th>Crossfired group</th>
<th>Sub-ruleSet</th>
<th>fpr</th>
<th>dr</th>
<th>ddr</th>
<th>benefit</th>
<th>profit</th>
<th>benefitCostRatio</th>
<th>BenefitCostRatio Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,5,6,9,10,19,20,25]</td>
<td>[6]</td>
<td>0.9790</td>
<td>0.109</td>
<td>0.0583255</td>
<td>256789</td>
<td>233206.55</td>
<td>10.916951</td>
<td>0.01266951</td>
</tr>
<tr>
<td>[6,8,10,24,25]</td>
<td>[6]</td>
<td>0.9822</td>
<td>0.131</td>
<td>0.05955326</td>
<td>37050</td>
<td>338602.12</td>
<td>11.72634</td>
<td>0.0121634</td>
</tr>
<tr>
<td>[6,8,10,24,25]</td>
<td>[6]</td>
<td>0.9842</td>
<td>0.156</td>
<td>0.04978389</td>
<td>44455</td>
<td>425869.66</td>
<td>10.892601</td>
<td>0.012692601</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
<tr>
<td>[6,8,10,15,32]</td>
<td>[6]</td>
<td>0.9854</td>
<td>0.164</td>
<td>0.08718451</td>
<td>46152</td>
<td>44146.22</td>
<td>9.831935</td>
<td>0.012931935</td>
</tr>
</tbody>
</table>

### Notes:
- The table above presents the evaluation of cross-fired rule groups, focusing on metrics such as fpr, dr, ddr, benefit, profit, and benefit cost ratio.
- Each row represents a different sub-rule set, with associated values indicating performance metrics.
- The goal is to optimize the balance between benefit and cost, ensuring effective rule implementation.
Best rule unions analysis

Union Rule By Profit Ranking [Jun 2009]

<table>
<thead>
<tr>
<th>Best Rule Unions</th>
<th>Daily Alert</th>
<th>Profit</th>
<th>FPR</th>
<th>DR</th>
<th>DDR</th>
<th>Cost/Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>[57:10;51:39;6:25; 64:16;29;60;14:66;30]</td>
<td>127</td>
<td>10061.44</td>
<td>0.98</td>
<td>0.69</td>
<td>0.605</td>
<td>0.05</td>
</tr>
<tr>
<td>[57:10;51:39;6:25; 54:16;29;60;14:66;30;32]</td>
<td>127</td>
<td>10061.44</td>
<td>0.98</td>
<td>0.69</td>
<td>0.605</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Best rule unions analysis

• **Best rule unions with the most stable profit rank**

  - The table right lists top 20 most stable best rule unions out of these top 10 best rule unions in each month, according to CV of their monthly profit rank with corresponding CV value and average profit rank.
  - We use the relative profit rank of a rule union compared to other best rule unions.

<table>
<thead>
<tr>
<th>Optimal Rule Unions</th>
<th>Coefficient of Variation of Profit Rank</th>
<th>Average Profit Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>[81; 8, 61, 51; 24, 12, 40; 2, 5; 30; 47; 83; 41; 71]</td>
<td>0.381458584</td>
<td>60</td>
</tr>
<tr>
<td>[6; 13; 39; 61; 57; 17; 2; 60; 51; 32; 23; 44; 40; 47; 5; 18; 56]</td>
<td>0.389838479</td>
<td>9</td>
</tr>
<tr>
<td>[81; 8, 61, 51, 24, 12, 40; 2, 5; 30, 47; 83, 41]</td>
<td>0.391815917</td>
<td>57</td>
</tr>
<tr>
<td>[81; 8, 61, 51, 24, 12, 40; 2, 5; 30, 47; 83, 41]</td>
<td>0.399909985</td>
<td>55</td>
</tr>
<tr>
<td>[26; 40, 20; 8, 51, 62, 2; 1, 2; 56, 81, 83, 32, 47, 5]</td>
<td>0.406098929</td>
<td>44</td>
</tr>
<tr>
<td>[8; 13; 39; 61; 57; 17; 2; 60; 51; 32; 23; 44; 40; 47; 5; 18; 56, 62]</td>
<td>0.406736162</td>
<td>7</td>
</tr>
<tr>
<td>[8; 13; 39; 61; 57; 17; 2; 60; 51; 32; 23; 44; 40; 47; 5; 18; 56, 62]</td>
<td>0.407831498</td>
<td>14</td>
</tr>
<tr>
<td>[26; 40, 20; 8, 51, 62, 2; 1, 2; 56, 81, 83]</td>
<td>0.408656335</td>
<td>41</td>
</tr>
<tr>
<td>[69; 42; 64; 25; 2; 39; 40; 44; 30; 56; 57; 5; 47; 51; 62]</td>
<td>0.412107557</td>
<td>5</td>
</tr>
<tr>
<td>[8; 13; 39; 61; 57; 17; 2; 60; 51; 32; 23; 44; 40; 47; 5; 18; 56, 62, 41]</td>
<td>0.41707034</td>
<td>37</td>
</tr>
<tr>
<td>[8; 13; 39; 61; 57; 17; 2; 60; 51; 32; 23; 44; 40; 47; 5; 18; 56, 62, 41]</td>
<td>0.41707034</td>
<td>37</td>
</tr>
<tr>
<td>[26; 40, 20; 8, 51, 62, 2; 1, 2; 56, 81]</td>
<td>0.422434822</td>
<td>39</td>
</tr>
<tr>
<td>[8; 1; 8, 61, 51; 24, 12, 40; 2, 5; 30]</td>
<td>0.425809421</td>
<td>51</td>
</tr>
<tr>
<td>[57; 39, 29; 40; 44; 5; 56; 47; 15; 41; 38, 59]</td>
<td>0.431397101</td>
<td>39</td>
</tr>
<tr>
<td>[8; 1; 8, 61, 51; 24, 12, 40; 2, 5; 30]</td>
<td>0.431734145</td>
<td>53</td>
</tr>
<tr>
<td>[69; 42; 64; 25; 2; 39; 40; 44; 30; 56; 57; 5; 47; 51]</td>
<td>0.439297655</td>
<td>3</td>
</tr>
<tr>
<td>[57; 39, 29; 40; 44; 5; 56; 47; 15; 41; 38]</td>
<td>0.444878261</td>
<td>37</td>
</tr>
<tr>
<td>[8; 1; 8, 61, 51; 24, 12, 40; 2, 5]</td>
<td>0.445424954</td>
<td>52</td>
</tr>
</tbody>
</table>
CONCLUSIONS

Noniidenss is important but challenging
  • A new area to be explored

  • Coupling (dependency)
  • Heterogeneity
Opportunities

- Coefficient
- Frequency
- Co-occurrence
- CHMM
- Complex Bayesian network
- Copula
- Tensor
- Wishart process
- ...

UTS:AAI
THE ADVANCED ANALYTICS INSTITUTE
Your Feedback Is Appreciated!

- Longbing Cao
  longbing.cao@uts.edu.au
  www-staff.it.uts.edu.au/~lbcao

- UTS Advanced Analytics Institute
  www.analytics.uts.edu.au