NEC Laboratories America
Relentless passion for innovation

From Categorical to Numerical:
Multiple Transitive Distance Learning and Embedding

Kai Zhang, Qiaojun Wang, Zhengzhang Chen, Ivan Marsic,
Guofei Jiang, Vipin Kumar

NEC Laboratories America
Relentless passion for innovation
Rutgers, The State University of New Jersey
University of Minnesota
Northwestern University

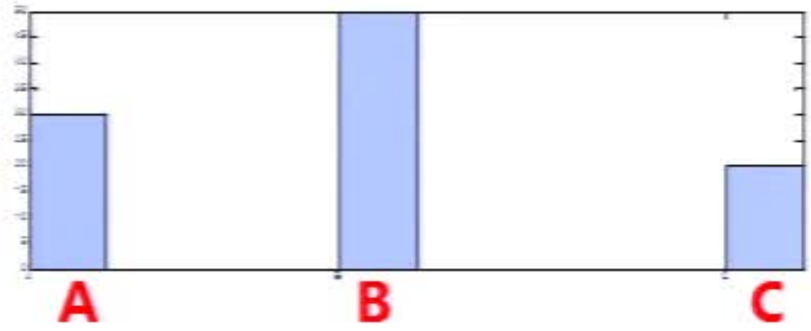
Outline

- Categorical Data and Challenges
- Why from Categorical to Numerical?
- Multiple Transitive Distance Learning
- Experimental Results

Categorical Data

Definition

- only takes limited number of values
  - levels, categories, groups, etc.
- nominal (no order) or ordinal (order)

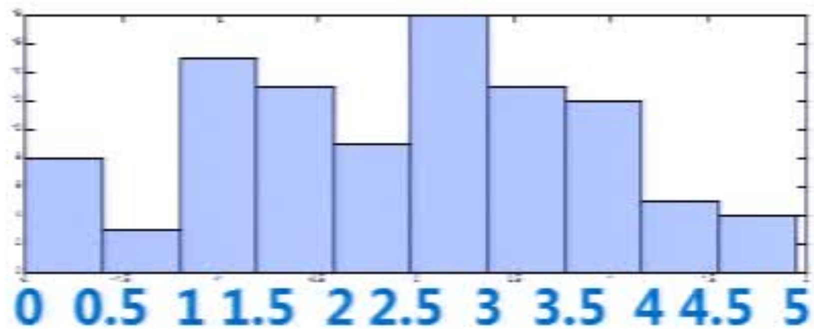


distribution of a categorical variable

Examples

- social, biology, psychology, …

attributes	symbols
Severity (symptom)	mild, moderate, severe
Weather	windy, cloudy, sunny
Blood type	A, B, AB, O



distribution of a numerical variable

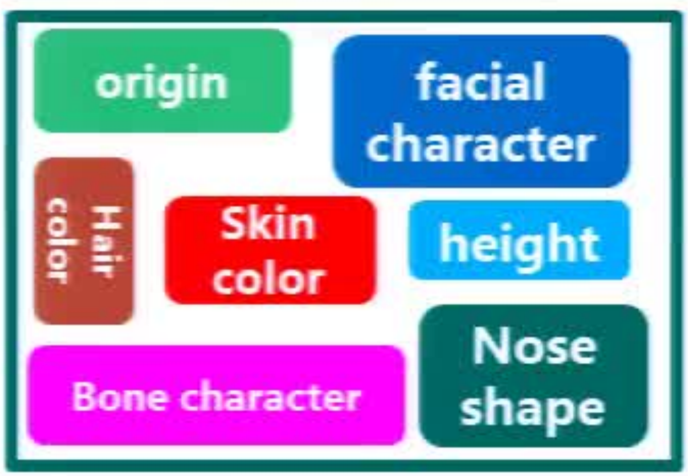
Categorical Data are Difficult

Curse of cardinality
- joint relation among symbols (and response)

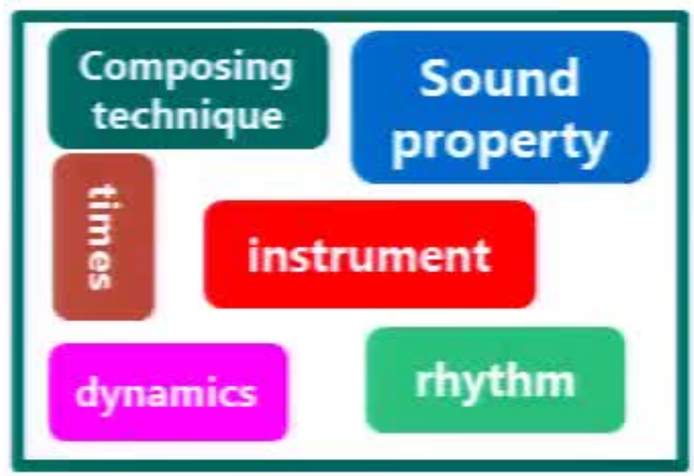
Lack of distance measures
- distance between symbols undefined

Richness of information
- one symbol can be a composite status

RACE



MUSIC



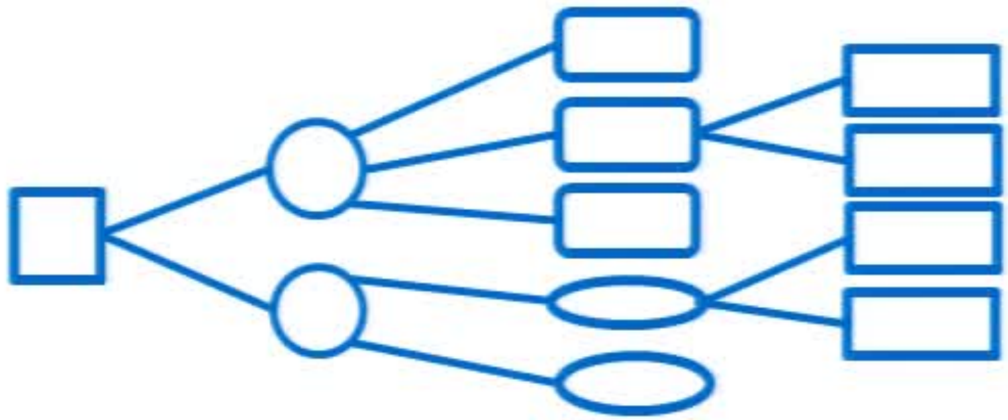
Existing Methods

Statistical modelling

- Binomial, multinomial, poisson, etc.
- Generalized Linear Models
- Logistic regression

Rule Analysis

- Decision Tree



Variable Coding

- Contrast coding
- Regression coding

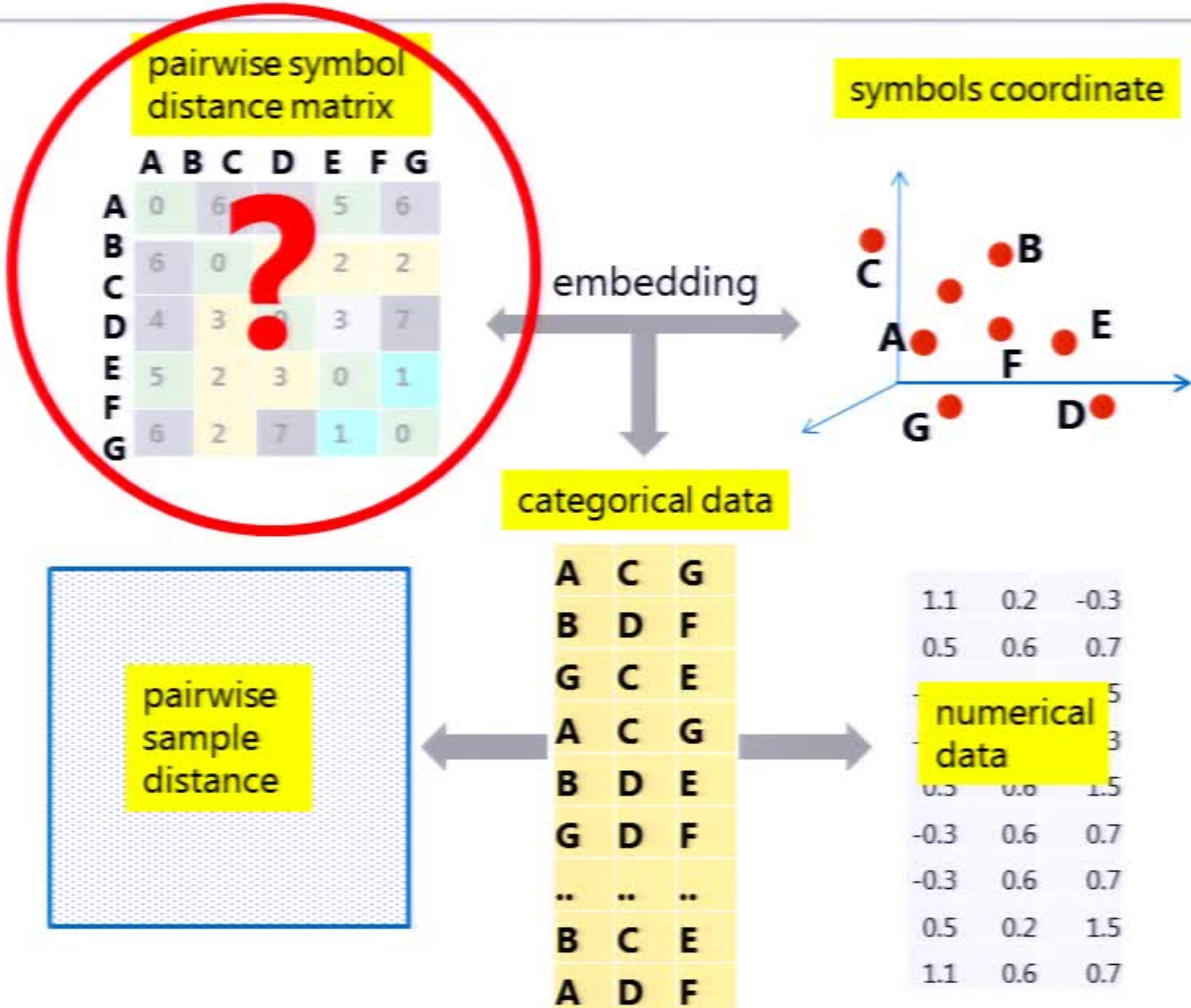
Level of race	New variable 1 (c1)	New variable 2 (c2)	New variable 3 (c3)
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (white)	-1	-1	-1

From categorical to Numerical

Why?

- continuous data sometimes easier to handle
  - distribution estimation
  - visualization
  - correlation analysis
  - ....
- abundant algorithms designed for numerical data

Key Idea



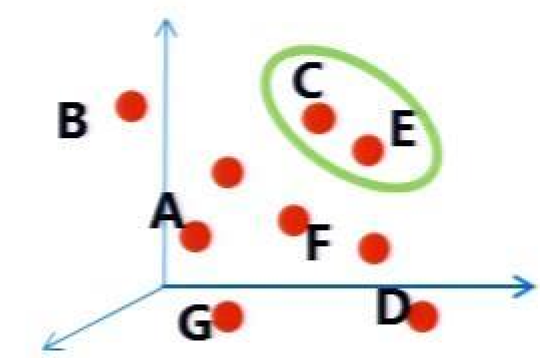
pairwise symbol distance matrix
A B C D E F G
?
embedding
symbols coordinate
categorical data
A C G
B D F
G C E
A C G
B D E
G D F
.. .. ..
B C E
A D F
pairwise sample distance
numerical data
1.1 0.2 -0.3
0.5 0.6 0.7
-0.3 0.6 0.7
-0.3 0.6 0.7
0.5 0.2 1.5
1.1 0.6 0.7

Distance and Embedding

Multidimensional Scaling [Cox and Cox 2001]

- pairwise distances of n objects $S \in R^{n \times n}$
- eigen-decomposition $-\frac{1}{2}HSH = U\Delta U$
- embedding $U\Delta^{-1/2}$ recovers distances in S

pairwise distance



low-dimensional embedding

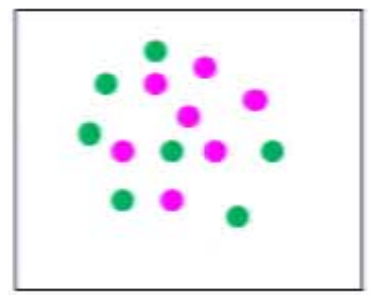
Basic Criterion

What are good symbol distances?

	A	B	C
A	0	2	3
B	2	0	4
C	3	4	0

	A	B	C
A	0	2	3
B	2	0	10
C	3	10	0

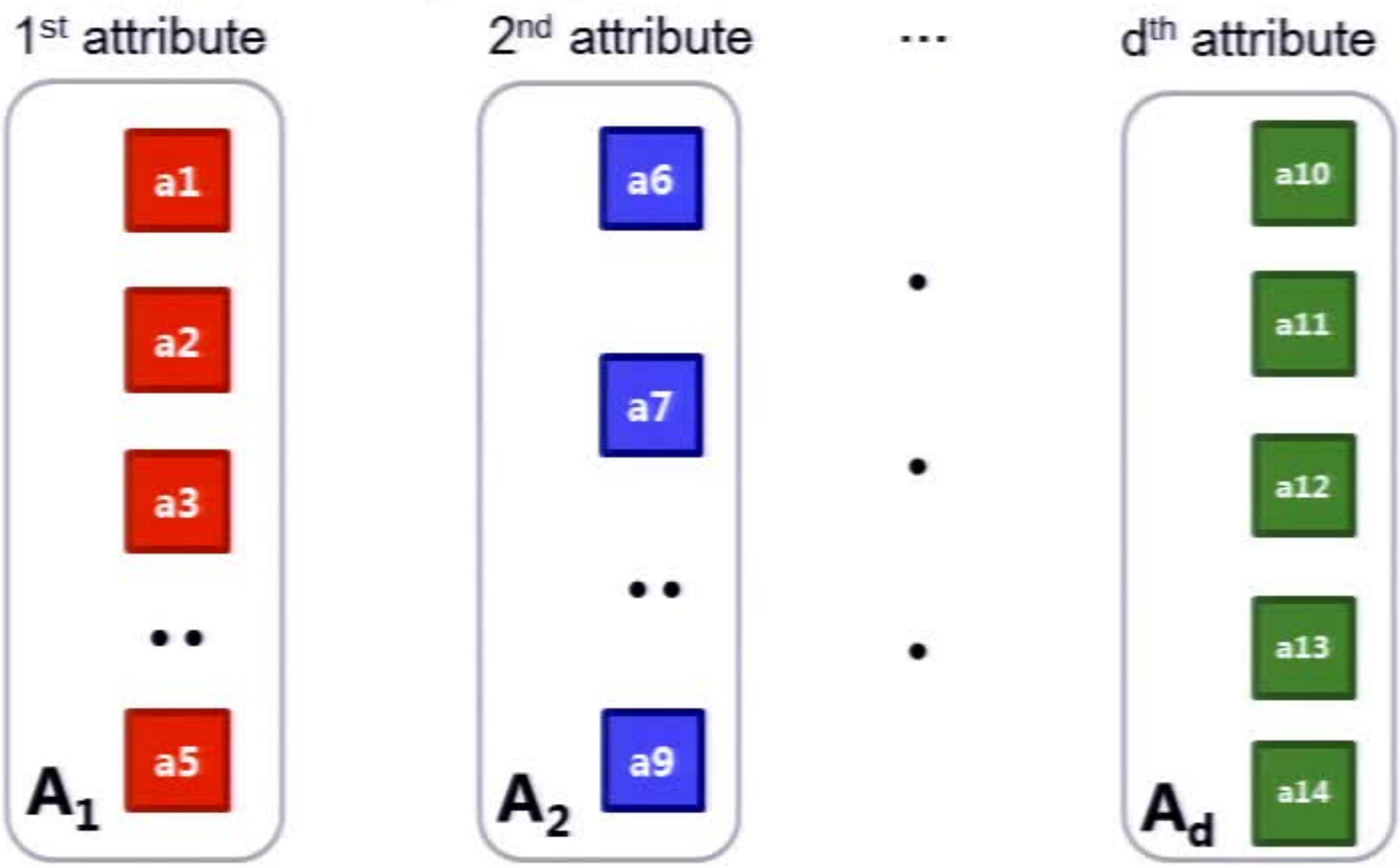
- **Transitivity**
  - if A ↔ B, and A ↔ C, then B ↔ C
  - co-occurrence statistics may fail
  - measuring proximity in categorical data [Gibson et al. 1998]



- **Consistency** (with target)
  - symbol distances determine *sample geometry*
  - induced geometry should be predictive of learning target

Multiple Transitive Distance Learning

- Notations: given n-by-d symbol matrix X



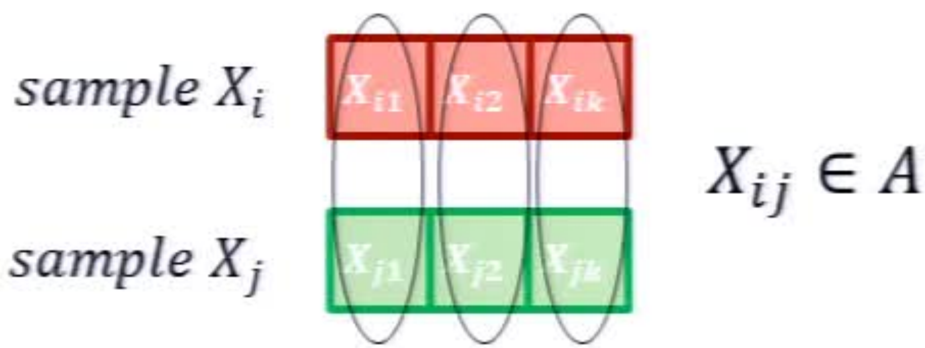
$$A = \{a_1, a_2, \dots, \quad a_6, a_7, \dots, \quad \dots, \quad a_{10}, a_{11}, \dots\}$$

Basic Assumptions

L-θ distance computation

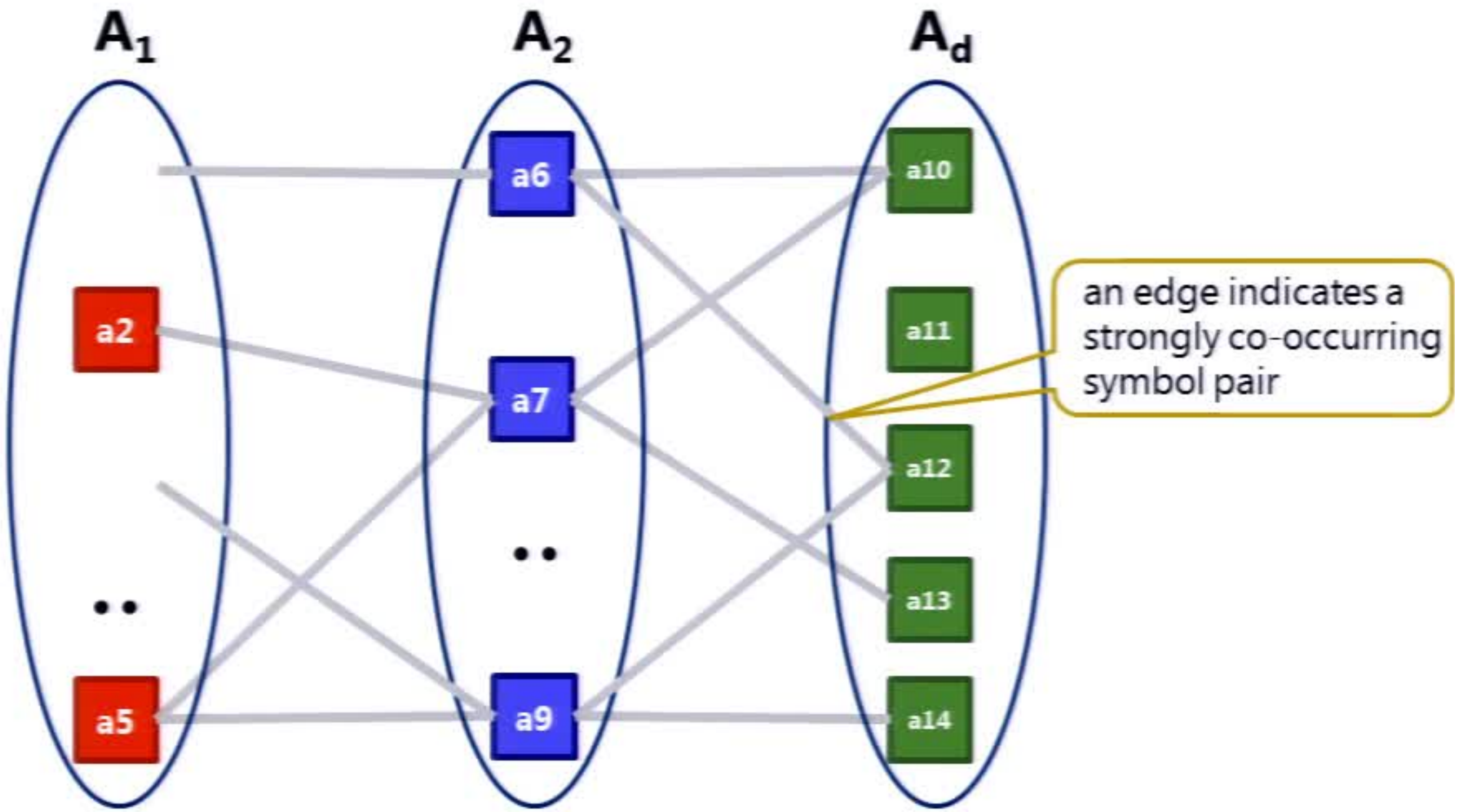
- distance between two samples is the sum of distances over each attribute

$$\mathrm{dis}(X_i, X_j)^\theta = \sum_{k=1}^{d} \mathrm{dis}(X_{ik}, X_{jk})^\theta$$



Multiple Transitive Distance Learning

- Transitive distances: shortest path



Initial d-partite graph of symbols

Basic Assumptions

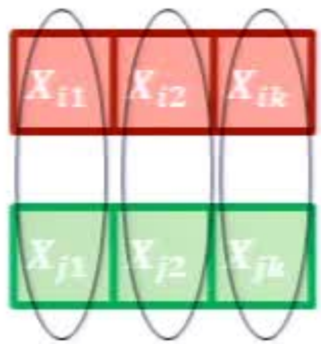
L-θ distance computation

- distance between two samples is the sum of distances over each attribute

$$\mathrm{dis}(X_i, X_j)^\theta = \sum_{k=1}^{d} \mathrm{dis}(X_{ik}, X_{jk})^\theta$$

sample X_i

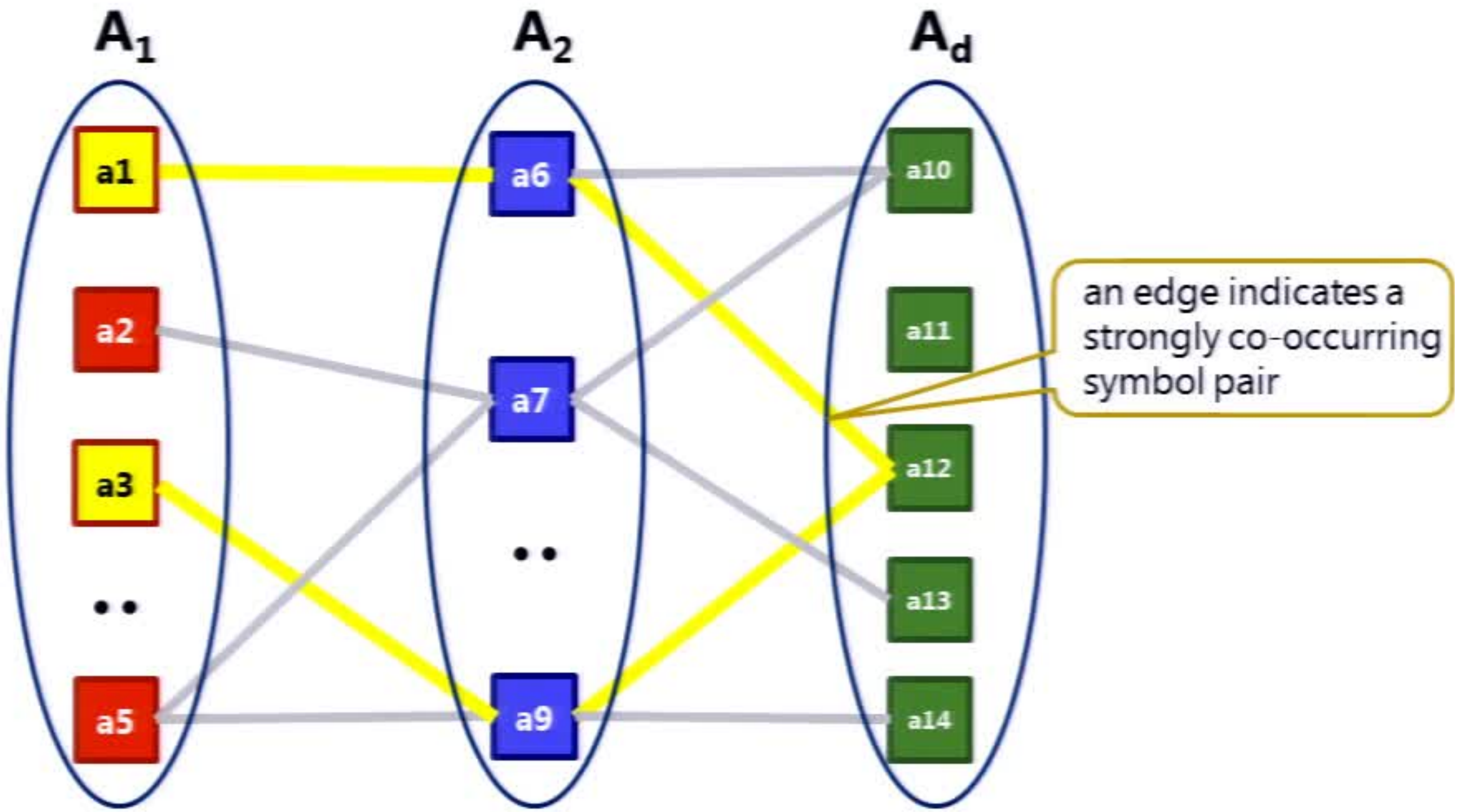
sample X_j



$$X_{ij} \in A$$

Multiple Transitive Distance Learning

- Transitive distances: shortest path



Initial d-partite graph of symbols

Choice of d-partite graphs

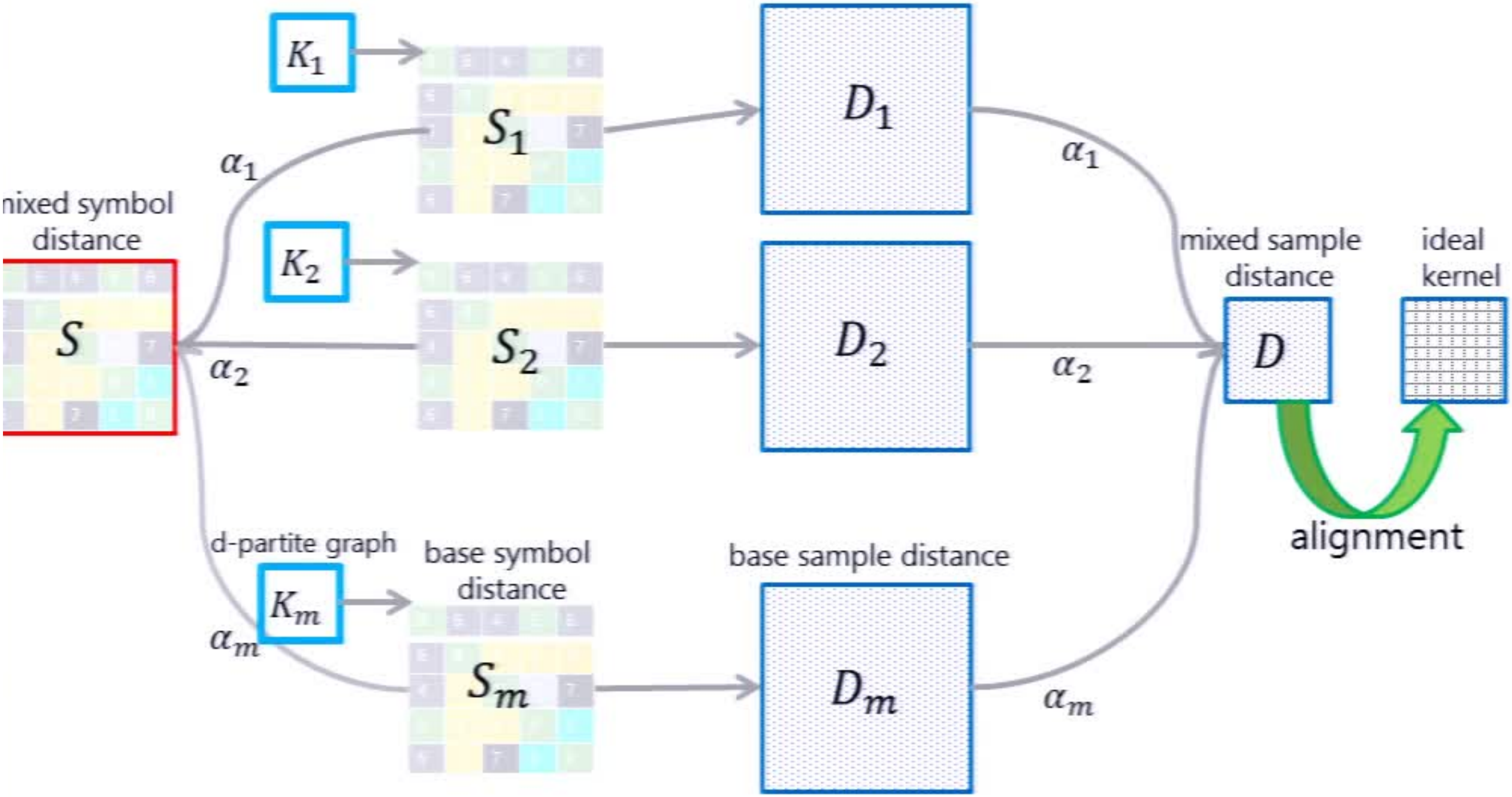
k-partite graph ***K*** using different proximities

- step 1: sparsification
- step 2: turn similarity to distance ($\frac{1}{x}, -\log(x)$)

	Proximity	Connected edge	Non-conn.
Similarities	co-occurrence	$\lvert a_p \cap a_q \rvert$	0
	norm. co-occr.	$\lvert a_p \cap a_q \rvert / \lvert a_p \cup a_q \rvert$	
	mutual info.	$\sum p(a_p = e, a_q = \hat{e}) \log\left(\frac{p(a_p = e, a_q = \hat{e})}{p(a_p = e) p(a_q = \hat{e})}\right)$	
Distances	hamming dist.	$\lvert a_p - a_q \rvert$	∞
	Euclidian dist.	$\lVert a_p - a_q \rVert_2$	
	cosine dist.	$arccos\left(a_p' a_q / \sqrt{\lVert a_p \rVert \lVert a_q \rVert} \right)$	

Multiple Distance Learning

- Align multiple base distances to class labels



Multiple distance learning

Given constraints: must-link $\boldsymbol{S}$; cannot-link $\boldsymbol{D}$

- sum of within-class distance pairs: $\mu_{ij} = [D_{1(i,j)}, D_{2(i,j)}, \ldots, D_{m(i,j)}]$

$$\boldsymbol{J}_S^{\epsilon} = \sum_{(i,j)\in S} \left(\sum_{m=1}^{L} \alpha_m D_{m(i,j)} \right)^{\epsilon} = \sum_{(i,j)\in S} (\langle \mu_{ij}, \boldsymbol{\alpha} \rangle)^{\epsilon}$$

- sum of inter-class distance pairs $\boldsymbol{J}_D = \sum_{(i,j)\in D} (\langle \mu_{ij}, \boldsymbol{\alpha} \rangle)^{\epsilon}$

Discriminative embedding

- linear program

$$\max_{\boldsymbol{\alpha}} \boldsymbol{J}_D^1 - \boldsymbol{J}_S^1 = \left(\sum_{(i,j)\in D} \mu_{ij} - \sum_{(i,j)\in S} \mu_{ij} \right)^T \boldsymbol{\alpha}$$

$$s.t. \ \ \boldsymbol{\alpha} \geq 0, \boldsymbol{\alpha}^T 1 = 1$$

- quadratic program

$$\min_{\boldsymbol{\alpha}} \frac{\boldsymbol{J}_S^2}{\boldsymbol{J}_D^2} = \frac{\boldsymbol{\alpha}^T \left(\sum_{(i,j)\in S} \mu_{ij} {\mu_{ij}}^T \right) \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \left(\sum_{(i,j)\in D} \mu_{ij} {\mu_{ij}}^T \right) \boldsymbol{\alpha}}$$

Experimental Results

Competing Methods

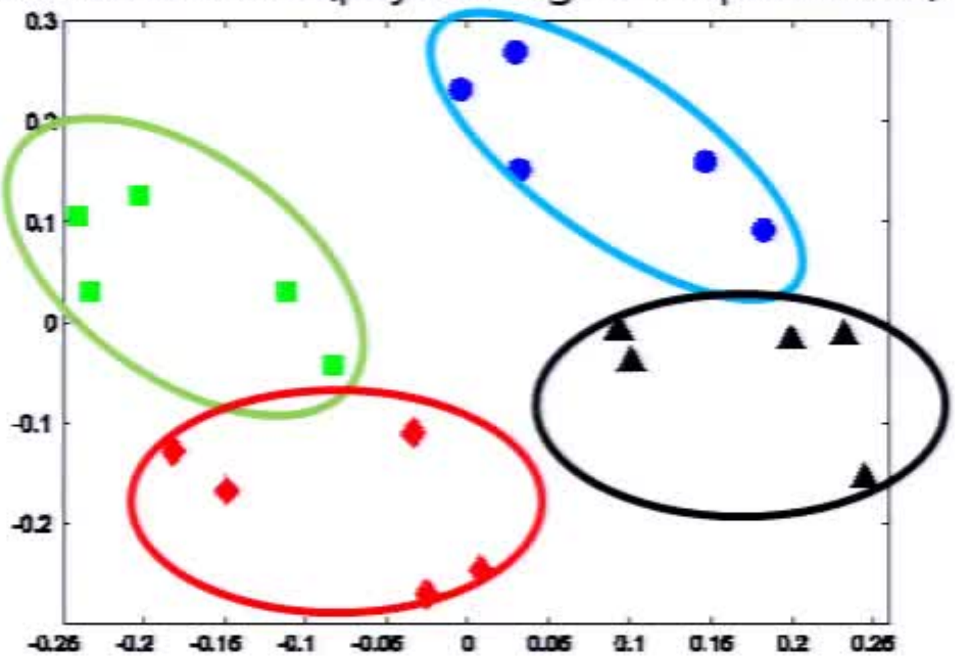
1. decision tree
2. dummy coding
3. density-based logistic regression [Chen et al 2013]
4. our method

Mean and variance of the classification accuracy on UCI datasets.
(random splits 50/50 as training/test data repeated 30 times)

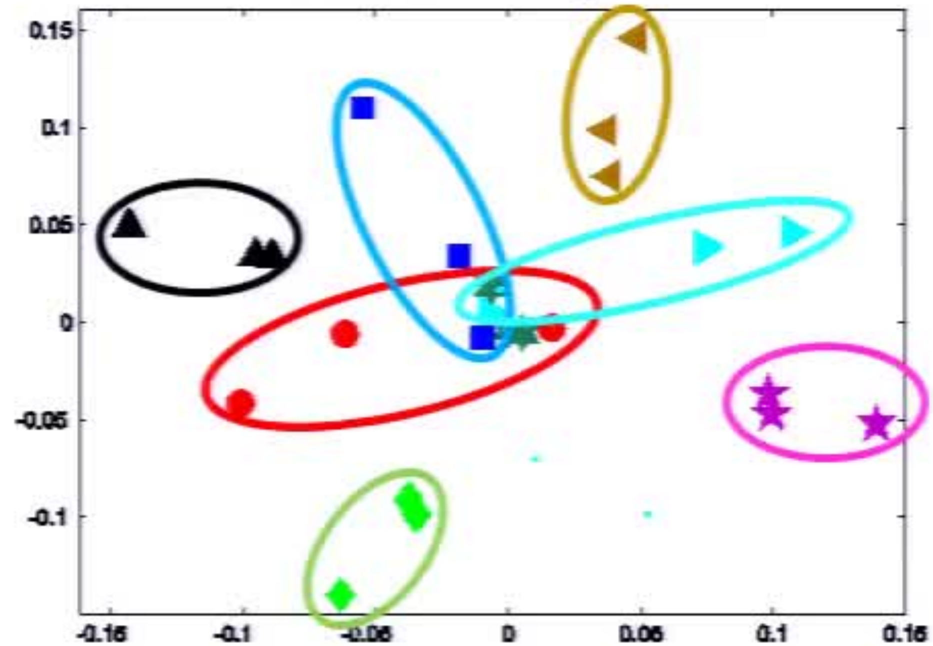
	Balance	Mushroom	Tic-Tac-Toe	Splice	Cancer	Hayes-Roth	Monk
Dummy	98.50±0.75	99.04±0.73	95.25±0.42	91.18±0.94	95.81±1.23	77.42±6.82	96.24±1.13
Density	96.59±1.58	98.43±0.57	70.60±2.34	**91.29±0.98**	**96.7±0.76**	57.7±8.94	96.37±0.85
Dec. Tree	88.49±1.63	**99.40±0.54**	87.48±2.18	88.68±1.73	94.25±1.38	73.93±7.74	97.13±0.72
Ours	**99.42±0.58**	**99.54±0.46**	**98.25±0.44**	**91.51±1.20**	95.94±0.93	**83.91±3.64**	**97.81±1.23**

Embedding Case Study

Balance-scale (psychological experiment)



Tic-Tac-Toe Endgame



1. Left-Weight: (1, 2, 3, 4, 5)
2. Left-Distance: (1, 2, 3, 4, 5)
3. Right-Weight: (1, 2, 3, 4, 5)
4. Right-Distance: (1, 2, 3, 4, 5)

1. top-left-square: {x,o,b}
2. top-middle-square: {x,o,b}
3. top-right-square: {x,o,b}
4. middle-left-square: {x,o,b}
5. middle-middle-square: {x,o,b}
6. middle-right-square: {x,o,b}
7. bottom-left-square: {x,o,b}
8. bottom-middle-square: {x,o,b}
9. bottom-right-square: {x,o,b}