Semimetric properties of Sørensen-Dice and Tversky Indices

Alonso GRAGERA

(joint work with Vorapong Suppakitpaisarn)

alonso@is.s.u-tokyo.ac.jp

The University of Tokyo

Short overview

Jaccard-Tanimoto index

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Short overview

Jaccard-Tanimoto index

$$S_J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Sørensen-Dice index

$$S_{SD}(X, Y) = \frac{2\ |X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

東京大学 THE UNIVERSITY OF TOKYO

Short overview

Jaccard-Tanimoto index

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Sørensen-Dice index

$$S_{SD}(X,Y) = \frac{2\,|X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha\,|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Short overview

Metric **(distance)**

1. d(x,y) ≥ 0
2. d(x,y) = 0 ⇔ x = y
3. d(x,y) = d(y,x)
4. d(x,z) ≤ d(x,y) + d(y, z)

METRIC

Jaccard-Tanimoto index

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Sørensen-Dice index

$$S_{SD}(X,Y) = \frac{2\,|X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha\,|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Short overview

Metric **(distance)**

1. d(x,y) ≥ 0
2. d(x,y) = 0 ⇔ x = y
3. d(x,y) = d(y,x)
4. d(x,z) ≤ d(x,y) + d(y, z)

↓

4. d(x,z) ≤ ρ(d(x,y) + d(y, z))

METRIC

Jaccard-Tanimoto index

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

NEAR-METRIC

Sørensen-Dice index

$$S_{SD}(X,Y) = \frac{2\,|X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha\,|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Short overview

Metric **(distance)**

1. d(x,y) ≥ 0
2. d(x,y) = 0 ⇔ x = y
3. d(x,y) = d(y,x)
4. d(x,z) ≤ d(x,y) + d(y, z)

↓

4. d(x,z) ≤ ρ(d(x,y) + d(y, z))

Can guarantee efficiency of several machine learning algorithms if ρ is small

METRIC

Jaccard-Tanimoto index

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

NEAR-METRIC

Sørensen-Dice index (ρ = 1.5)

$$S_{SD}(X,Y) = \frac{2\,|X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index (ρ = (α + 1)/2)

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha\,|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Outline

1. Similarity indexes
2. Near-metricness
3. Our results
4. Proof outline

Similarity indexes

Jaccard-Tanimoto index (α = 1)

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

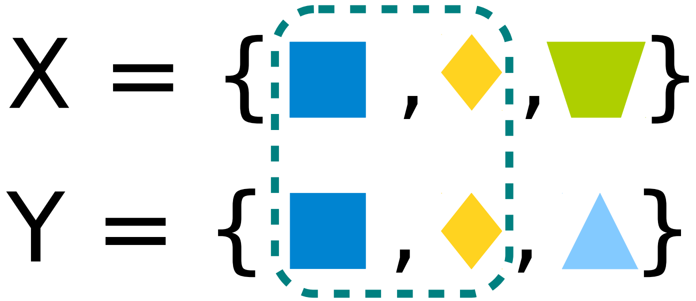
Sørensen-Dice index (α = 2)

$$S_{SD}(X,Y) = \frac{2\ |X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index **[Our proposal]**

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha |X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Similarity indexes



Jaccard-Tanimoto index (α = 1)

$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Sørensen-Dice index (α = 2)

$$S_{SD}(X,Y) = \frac{2\,|X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index **[Our proposal]**

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha\,|X \cap Y|}{|X \cup Y| + (\alpha - 1)\,|X \cap Y|}$$

Similarity indexes

Jaccard-Tanimoto index (α = 1)

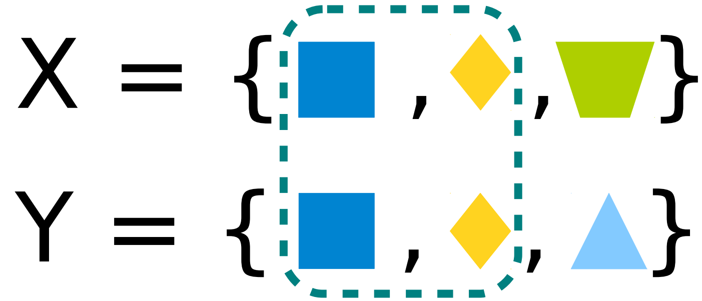
$$S_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Sørensen-Dice index (α = 2)

$$S_{SD}(X,Y) = \frac{2\ |X \cap Y|}{|X \cup Y| + |X \cap Y|}$$

Robust Jaccard index **[Our proposal]**

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha |X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$



$$|X \cap Y| = 2$$
$$|X \cup Y| = 4$$

$$S_J(X,Y) = \frac{2}{4} = 0.5$$

$$S_{SD}(X,Y) = \frac{2 \cdot 2}{4+2} = 0.66$$

$$S_{RJ,4}(X,Y) = \frac{4 \cdot 2}{4+3 \cdot 2} = 0.8$$

東京大学 THE UNIVERSITY OF TOKYO

Similarity indexes

Robust Jaccard index **[Our proposal]**

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Tversky index family

$$S_T(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \beta|X - Y| + \gamma|Y - X|}$$

with β,ɣ ∈ [0,1]

Similarity indexes

Robust Jaccard index **[Our proposal]**

$$S_{RJ,\alpha}(X,Y) = \frac{\alpha|X \cap Y|}{|X \cup Y| + (\alpha - 1)|X \cap Y|}$$

Tversky index family

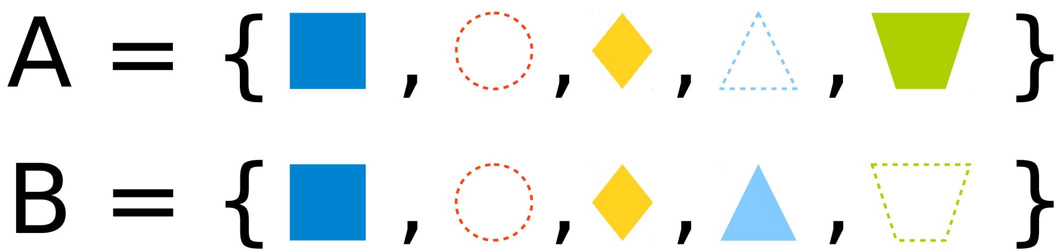
$$S_T(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \beta|X - Y| + \gamma|Y - X|}$$

with β,ɣ ∈ [0,1]

- Robust Jaccard index is equal to Jaccard-Tanimoto index when α = 1.
- Robust Jaccard index is equal to Sørensen-Dice index α = 2.
- Robust Jaccard index is a subset of the Tversky index family when β = ɣ.

Why is 'Robustness' needed?

Larger α values are the more robust it becomes.

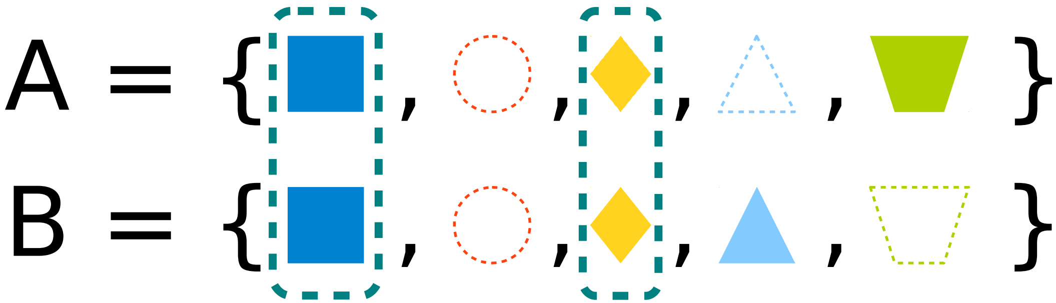


東京大学 THE UNIVERSITY OF TOKYO

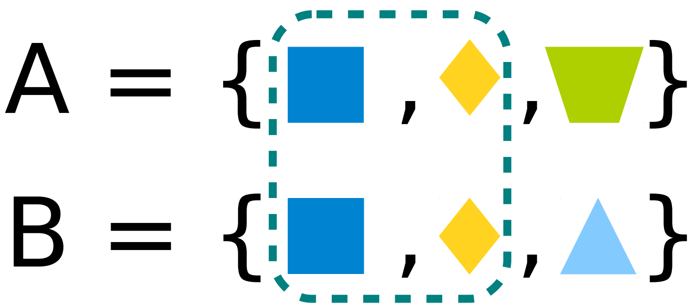
Why is 'Robustness' needed?

When measuring similarity, context (instrumental error, missing information, ...) matters.

It is not the same:



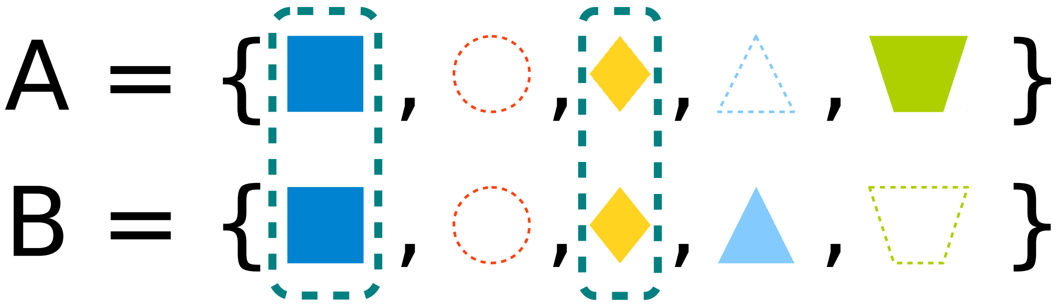
Than:



Why is 'Robustness' needed?

When measuring similarity, context (instrumental error, missing information, ...) matters.

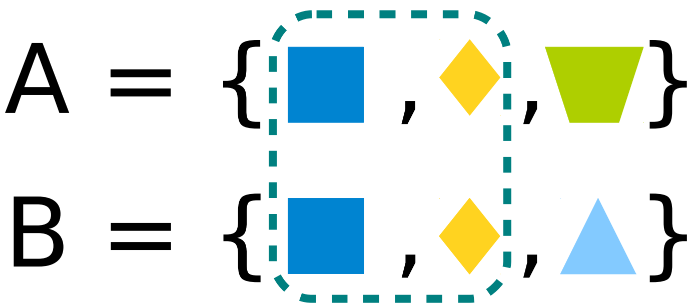
It is not the same:



$S_J = 0.5$

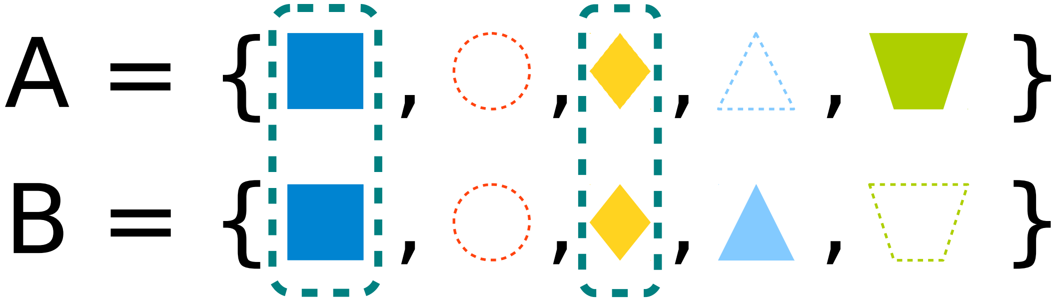
$S_{SD} = 0.66$

Than:

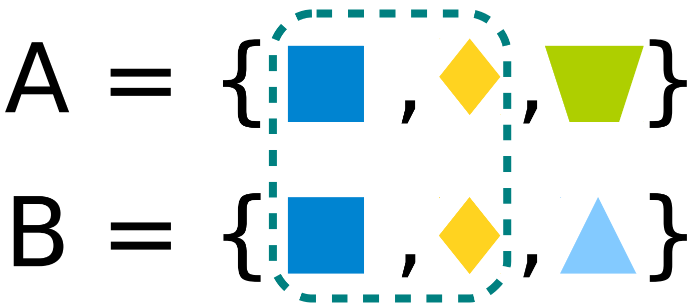


Why is 'Robustness' needed?

But if we have an insight of how good is the observation at each time, then we can adjust the granularity of our similarity measure.

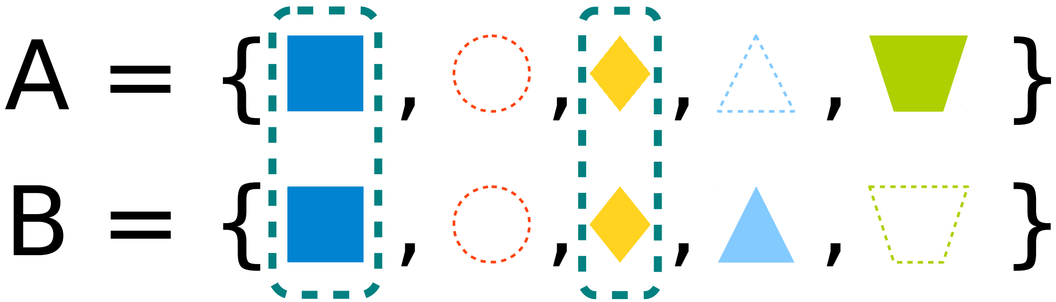


Or



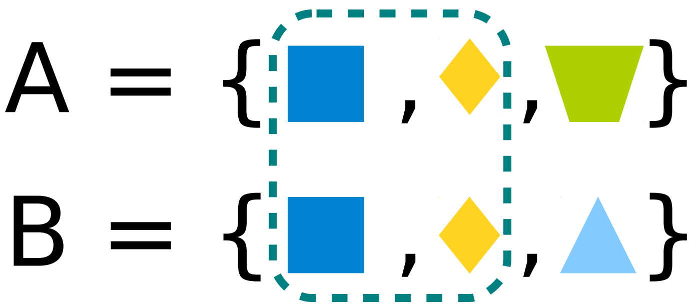
Why is 'Robustness' needed?

But if we have an insight of how good is the observation at each time, then we can adjust the granularity of our similarity measure.



$S_{RJ,4} = 0.8$

Or

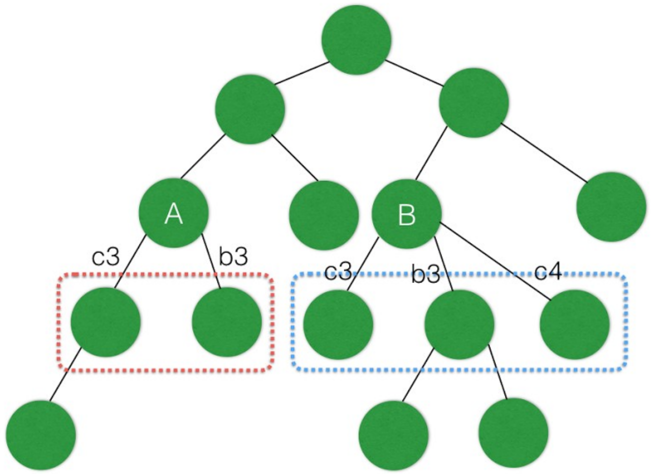


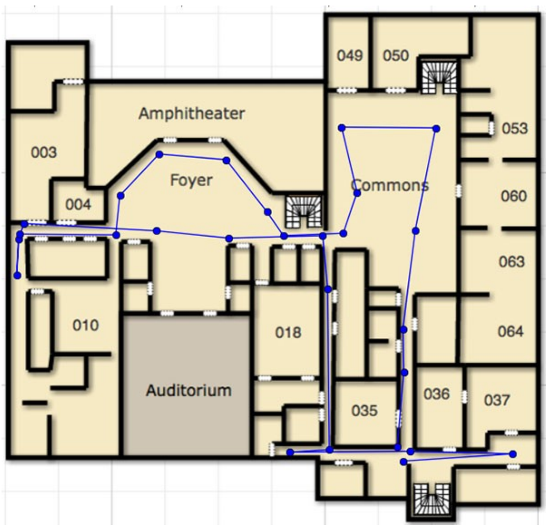
$S_{RJ,1} = 0.5$

Why is 'Robustness' needed?

Computer Go
In Monte-Carlo Tree Search algorithm the quality of the moves for each board position depends on the number of simulations performed so far.

[Gragera, 2015]





Self Localization and Mapping
In wifi-SLAM for robotics the signal presence and intensity depends on the noise of the environment.

[Miyagusuku, 2016]

Outline

1. Similarity indexes
2. Near-metricness
3. Our results
4. Proof outline

Metric, Semi-metric and Near-metric

Metric properties *(or distance)*

1. $d(x,y) \geq 0$ — N*on-negativity*
2. $d(x,y) = 0 \Leftrightarrow x = y$ — *Identity of indiscernibles*
3. $d(x,y) = d(y,x)$ — *Symmetry*
4. $d(x,z) \leq d(x,y) + d(y, z)$ — *Triangle inequality*

東京大学 THE UNIVERSITY OF TOKYO

Metric, Semi-metric and Near-metric

Metric properties *(or distance)*

1. $d(x,y) \geq 0$ — N*on-negativity*
2. $d(x,y) = 0 \Leftrightarrow x = y$ — *Identity of indiscernibles*
3. $d(x,y) = d(y,x)$ — *Symmetry*
4. $d(x,z) \leq d(x,y) + d(y, z)$ — *Triangle inequality*

Semi-metric properties

4. $\cancel{d(x,z) \leq d(x,y) + d(y, z)}$

東京大学 THE UNIVERSITY OF TOKYO

Metric, Semi-metric and Near-metric

Metric properties *(or distance)*

1. $d(x,y) \geq 0$ — N*on-negativity*
2. $d(x,y) = 0 \Leftrightarrow x = y$ — *Identity of indiscernibles*
3. $d(x,y) = d(y,x)$ — *Symmetry*
4. $d(x,z) \leq d(x,y) + d(y, z)$ — *Triangle inequality*

Semi-metric properties

4. ~~$d(x,z) \leq d(x,y) + d(y, z)$~~

Near-metric properties *(or ρ-relaxed semi-metric properties)*

4. $d(x,z) \leq \rho(d(x,y) + d(y, z))$ — Approximate triangle inequality

東京大学 THE UNIVERSITY OF TOKYO

Why is ρ so important?

Proving algorithms efficiency

For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011]

Why is ρ so important?

Proving algorithms efficiency

For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011] Streaming k-means

Online median

Why is ρ so important?

Proving algorithms efficiency
For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011] Streaming k-means

Online median

Proving algorithms approximation ratios
Approximation ratio depends on the value of ρ of the near-metric used.

[Krug, 2013][Jaiswal, Kumar and Yadav, 2015]

Why is ρ so important?

Proving algorithms efficiency

For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011] Streaming k-means

Online median

Proving algorithms approximation ratios

Approximation ratio depends on the value of ρ of the near-metric used.

[Krug, 2013][Jaiswal, Kumar and Yadav, 2015] (1+Ɛ) approximation for k-means

Δ_β-TSP

Why is ρ so important?

Proving algorithms efficiency
For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011] Streaming k-means

Online median

Proving algorithms approximation ratios
Approximation ratio depends on the value of ρ of the near-metric used.

[Krug, 2013][Jaiswal, Kumar and Yadav, 2015] (1+Ɛ) approximation for k-means

Δ_β-TSP

Proving properties of generalized problems
Several properties of well-studied metric problems can be generalized for the near-metric cases.

[Xia, 2008]

Why is ρ so important?

Proving algorithms efficiency
For many recent algorithms, computation time depends on the value of ρ.

[Mettu & Plaxton, 2006][Braverman et al., 2011] ← Streaming k-means

Online median

Proving algorithms approximation ratios
Approximation ratio depends on the value of ρ of the near-metric used.

[Krug, 2013][Jaiswal, Kumar and Yadav, 2015] ← (1+Ɛ) approximation for k-means

Δ_β-TSP

Proving properties of generalized problems
Several properties of well-studied metric problems can be generalized for the near-metric cases.

[Xia, 2008] ← Geodesic problem

東京大学 THE UNIVERSITY OF TOKYO

Outline

1. Similarity indexes
2. Near-metricness
3. Our results
4. Proof outline

Summary of our paper results

Dissimilarity		Result	
Jaccard-Tanimoto	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$	ρ* = 1	(metric) *[Lipkus, 1999]*
Sørensen-Dice	$\frac{2\,\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert + \lvert X \cap Y \rvert}$	ρ* = 1.5	(near-metric) *[Our Result]*
Robust-Jaccard	$\frac{\alpha \lvert X \cap Y \rvert}{\lvert X \cup Y \rvert + (\alpha - 1) \lvert X \cap Y \rvert}$	ρ* = (α + 1)/2	(near-metric) *[Our Result]*
Tversky	$\frac{\lvert X \cap Y \rvert}{\lvert X \cap Y \rvert + \beta \lvert X - Y \rvert + \gamma \lvert Y - X \rvert}$	*If β = γ = 1/α* ρ* = (α + 1)/2	(near-metric) *[Our Result]*
		If β ≠ γ	

Near-metricness is proven and tight ρ values are also given.

Summary of our paper results

Dissimilarity		Result	
Jaccard-Tanimoto	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$	$\rho^* = 1$	(metric) *[Lipkus, 1999]*
Sørensen-Dice	$\frac{2\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert + \lvert X \cap Y \rvert}$	$\rho^* = 1.5$	(near-metric) *[Our Result]*
Robust-Jaccard	$\frac{\alpha \lvert X \cap Y \rvert}{\lvert X \cup Y \rvert + (\alpha - 1)\lvert X \cap Y \rvert}$	$\rho^* = (\alpha + 1)/2$	(near-metric) *[Our Result]*
Tversky	$\frac{\lvert X \cap Y \rvert}{\lvert X \cap Y \rvert + \beta \lvert X - Y \rvert + \gamma \lvert Y - X \rvert}$	*If $\beta = \gamma = 1/\alpha$* $\rho^* = (\alpha + 1)/2$	(near-metric) *[Our Result]*
		If $\beta \neq \gamma$ Near-quasimetric	

Near-metricness is proven and tight ρ values are also given.

東京大学 THE UNIVERSITY OF TOKYO

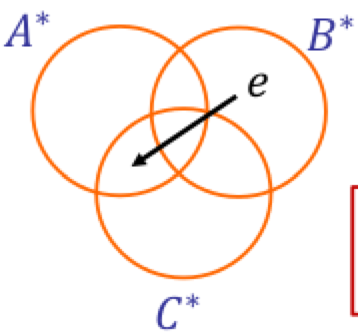
Outline

1. Similarity indexes
2. Near-metricness
3. Our results
4. Proof outline

Proof Outline

$$\rho^* = \max_{A,B,C} \frac{d(A,C)}{d(A,B)+d(B,C)}$$

Find $A^*, B^*, C^* = \max_{A,B,C} \frac{d(A,C)}{d(A,B)+d(B,C)}$



$$d(A', B') \leq d(A^*, B^*)$$
$$d(B', C') \leq d(B^*, C^*)$$
$$d(A', C') \geq d(A^*, C^*)$$

A', B', C' can also be the optimization results.

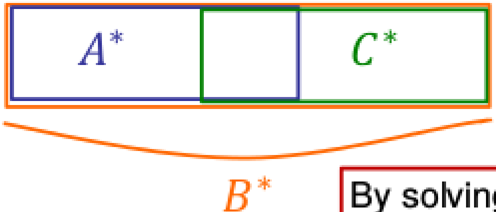
$$d(A', C') = \frac{\alpha(|A^* \cap C^*| + 1)}{|A^* \cup C^*| + 1 + (\alpha - 1)(|A^* \cap C^*| + 1)}$$
$$= \frac{\alpha|A^* \cap C^*| + \alpha}{|A^* \cup C^*| + (\alpha - 1)|A^* \cap C^*| + \alpha}$$
$$\geq \frac{\alpha|A^* \cap C^*|}{|A^* \cup C^*| + (\alpha - 1)|A^* \cap C^*|} = d(A^*, C^*)$$

To find ρ is equivalent to find A^*,B^*,C^*

Adding certain elements doesn't improve to A^*,B^* and C^*

A^* and C^* should of the same size and disjoint

B^* is the union of A^* and C^*



By solving optimization problem, we have $A^* = \{1\}$, $B^* = \{1,2\}$, $C^* = \{2\}$

Concluding remarks

- Robustness is very interesting property of Robust Jaccard index for dynamic problems.
- Sørensen-Dice, Robust Jaccard and Tversky are now included the indexes that can be safely used as near-metrics.
- We now know their value of ρ, that is critical to guarantee the performance of several algorithms.

References

- P. Jaccard, *"Lois de distribution florale dans la zone alpine"*, Corbaz, 1902.
- T. Sørensen, *"A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons"*, Biologiske Skrifte vol. 5, pages 1–34, 1948.
- A. Tversky and I. Gati, *"Similarity, separability, and the triangle inequality"*. Psychological review, vol.89, no. 2, page 123, 1982.
- A. Lipkus, *"A proof of the triangle inequality for the Tanimoto distance"*, Journal of Mathematical Chemistry, vol. 26, no. 1-3, pages 263-265, 1999.
- Q. Xia, "The geodesic problem in nearmetric spaces", Technical report 2008.
- R.R Mettu and C.G. Plaxton, *"The online median problem"*, SIAM Journal on Computing, vol. 32, no. 3, pages 816–832, 2003.
- V. Braverman et al., *"Streaming k-means on well-clusterable data"*, Proceedings of SODA'11, pages 26-40, 2011.
- R. Jaiswal, M. Kumar and P. Yadav, *"Improved analysis of D2-sampling based PTAS for k-means and other clustering problems"*, Information Processing Letters, vol. 115, no. 2, pages 100-103, 2015.

References

- S. Krug, *“Analysis of a near-metric TSP approximation algorithm”*, RAIRO - Theoretical Informatics and Applications - Informatique Théorique et Applications, vol. 47, no. 3, pages 293-314, 2013.
- A. Gragera, “Approximate matching for Go board positions”, Proceedings of GPW’15, 2015.
- R. Miyagusuku, Personal communication, 2016.


Thank you!

Q&A