Distances Between Sets Based on Set Commonality

Kathy Horadam and Michael Nyblom

RMIT University, Melbourne, Australia

Australian and New Zealand Mathematics Convention,
Melbourne
10 December, 2014
Outline

1. **Motivation**
   - Biometric Graph Matching
   - Scoring functions

2. **Minkowski-type Metrics for Sets**
   - Minkowski-type metrics for sets
   - Normalising the Minkowski-type metrics for sets
Outline

1 Motivation
   - Biometric Graph Matching
   - Scoring functions

2 Minkowski-type Metrics for Sets
   - Minkowski-type metrics for sets
   - Normalising the Minkowski-type metrics for sets
Background

- A novel area for me, probably a once-off!
- Work arises from a problem in biometric matching
- Biometric matching: Personal identification through “what you are" (fingerprint, face) not “what you know" (PIN, password) or “what you carry" (token, smartcard)
Biometric matching

- Biometric samples from an individual vary each time they are presented.
- Need error-tolerant or fuzzy matching of biometric features to authenticate.
- But not too error-tolerant or fuzzy, or an imposter can be authenticated as you!
Example: Features of Retina biometric

Extracting graphs from biometric images:
A retina image (a) and its retina graph (b)
Comparing biometric graphs

- Use error-correcting graph matching algorithm to compare retina graphs $G$ and $G^*$
- Optimal edit path defines Maximum Common Subgraph (mcs) $G \cap G^*$
- Can count many types of structural elements of each graph $G$, $G^*$, $G \cap G^*$
  - eg for vertices: we count $|V|$, $|V^*|$ and $|V \cap V^*|$.
  - Many other subgraphs could be counted: # edges, # paths length 2, # nodes degree 3, # simple cycles, etc etc
1 Motivation
   - Biometric Graph Matching
   - Scoring functions

2 Minkowski-type Metrics for Sets
   - Minkowski-type metrics for sets
   - Normalising the Minkowski-type metrics for sets
Example: use $\overline{d}_{sqrt}(G, G^*) = \frac{|V \cap V^*|}{\sqrt{|V||V^*|}}$ as the scoring function.

$\overline{d}_{sqrt}$ is a distance but NOT a metric (preferred)
What metrics are known for comparing sets?

**Distance** or dissimilarity $d : X \times X \rightarrow \mathbb{R}$ on data set $X$

- non-negative, symmetric and reflexive function.
- Normalised if $d(u, v) \leq 1 \ \forall \ u, v \in X$.
- And is a **metric** if $\forall \ u, v, w \in X$,
  - $d(u, v) = 0 \Leftrightarrow u = v$; and
  - the **triangle inequality** holds:
    
    $$d(u, v) \leq d(u, w) + d(w, v).$$
What metrics are known for comparing sets?

**Notation.** $U \neq \emptyset$, $X = \text{set of finite nonempty subsets of } U$. For $X_i, X_j, X_k \in X$, let $x_i = |X_i|$, $x_{ij} = |X_i \cap X_j|$, $x_{ijk} = |X_i \cap X_j \cap X_k|$. Let $m_{ij} = x_i - x_{ij}$, so $x_i > 0$ and $m_{ij} \geq 0$.

Put $x_i = x_i^* + y_{ij} + y_{ik} + x_{ijk}$, where $y_{ij} = y_{ji} = x_{ij} - x_{ijk}$, so $m_{ij} = x_i - x_{ij} = x_i^* + y_{ik}$.
What metrics are known for comparing sets?

The following are known normalised metrics on $X$.

**The Jaccard, or set-difference, metric**

$$d_{sd}(X_i, X_j) = (m_{ij} + m_{ji})/(x_i + x_j - x_{ij}) = 1 - x_{ij}/(x_i + x_j - x_{ij}).$$

**The normalised maximum metric**

$$d_{max}(X_i, X_j) = \max\{m_{ij}, m_{ji}\}/\max\{x_i, x_j\} = 1 - x_{ij}/\max\{x_i, x_j\}.$$
### Motivation

Minkowski-type Metrics for Sets

Biometric Graph Matching

Scoring functions

Which measures separate genuine from imposter retinas best?

<table>
<thead>
<tr>
<th>Score</th>
<th>Distance?</th>
<th>Metric?</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{\text{sqrt}}$ (good-ish)</td>
<td>✓</td>
<td>×</td>
<td>0.0486</td>
</tr>
<tr>
<td>$d_{sd}$ (good-ish)</td>
<td>✓</td>
<td>✓</td>
<td>0.0154</td>
</tr>
<tr>
<td>$d_{\text{max}}$ (ugly)</td>
<td>✓</td>
<td>✓</td>
<td>−0.0450</td>
</tr>
</tbody>
</table>

Comparison of nearest neighbour (NN) distances of vertex sets in VARIA retina database for 3 scoring functions

**ARE THERE BETTER SCORING METRICS THAN $d_{\text{max}}$?**
Which measures separate genuine from imposter retinas best?

<table>
<thead>
<tr>
<th>Score</th>
<th>Distance?</th>
<th>Metric?</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{\text{sqrt}}$ (good-ish)</td>
<td>✓</td>
<td>✗</td>
<td>0.0486</td>
</tr>
<tr>
<td>$d_{\text{sd}}$ (good-ish)</td>
<td>✓</td>
<td>✓</td>
<td>0.0154</td>
</tr>
<tr>
<td>$d_{\text{max}}$ (ugly)</td>
<td>✓</td>
<td>✓</td>
<td>−0.0450</td>
</tr>
</tbody>
</table>

Comparison of nearest neighbour (NN) distances of vertex sets in VARIA retina database for 3 scoring functions

Are there better scoring metrics than $d_{\text{max}}$?
Which measures separate genuine from imposter retinas best?

<table>
<thead>
<tr>
<th>Score</th>
<th>Distance?</th>
<th>Metric?</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{\text{sqrt}}$ (good-ish)</td>
<td>✓</td>
<td>✗</td>
<td>0.0486</td>
</tr>
<tr>
<td>$d_{sd}$ (good-ish)</td>
<td>✓</td>
<td>✓</td>
<td>0.0154</td>
</tr>
<tr>
<td>$d_{\text{max}}$ (ugly)</td>
<td>✓</td>
<td>✓</td>
<td>−0.0450</td>
</tr>
</tbody>
</table>

Comparison of nearest neighbour (NN) distances of vertex sets in VARIA retina database for 3 scoring functions

ARE THERE BETTER SCORING METRICS THAN $d_{\text{max}}$?
Motivation

Minkowski-type Metrics for Sets

Outline

1 Motivation
   - Biometric Graph Matching
   - Scoring functions

2 Minkowski-type Metrics for Sets
   - Minkowski-type metrics for sets
   - Normalising the Minkowski-type metrics for sets
The Minkowski or $p$-norm metrics $d_p$ on $\mathbb{R}^n$

These are defined for real $p \geq 1$ and dimension $n \geq 1$ to be

$$d_p((u_1, u_2, \ldots, u_n), (v_1, v_2, \ldots, v_n)) = \left( \sum_{i=1}^{n} |u_i - v_i|^p \right)^{\frac{1}{p}}.$$

- $p = 1$: absolute value distance, (taxicab, city-block or Manhattan distance).
- $p = 2$: usual Euclidean distance.
- $\lim_{p \to \infty} d_p = d_\infty$: infinity norm (Chebyshev) distance, max distance $d_\infty((u_1, u_2, \ldots, u_n), (v_1, v_2, \ldots, v_n)) = \max\{|u_1 - v_1|, |u_2 - v_2|, \ldots, |u_n - v_n|\}$.
- Varying $p$ changes the weight given to larger and smaller differences.
Eureka! Minkowski-type Metric Family for Sets

The following definition gives set-based metrics with analogous properties. For each $p \geq 1$, define $d_{2,p} : X \times X \rightarrow \mathbb{R}$ to be

$$d_{2,p}(X_i, X_j) = [m_{ij}^p + m_{ji}^p]^{rac{1}{p}}.$$ 

**Theorem**

1. $d_{2,p}$ is a metric;
2. $d_{2,1} = d_{sd}$, $\lim_{p \to \infty} d_{2,p} = d_{\text{max}}$ and if $p < p'$, $d_{2,p} \geq d_{2,p'}$.

For sets, the Minkowski-type metric is a modification of the 2D real Minkowski metric.

$$d_{2,p}(X_i, X_j) = d_p((x_i, x_j), (x_{ij}, x_{ij})).$$

$d_{2,2}$ is analogous to the Euclidean metric $d_2$ in the plane.
Outline

1. Motivation
   - Biometric Graph Matching
   - Scoring functions

2. Minkowski-type Metrics for Sets
   - Minkowski-type metrics for sets
   - Normalising the Minkowski-type metrics for sets
Normalising the Minkowski-type Metric Family

Theorem

For each \( p \geq 1 \), define \( \overline{d}_{2,p} \) to be

\[
\overline{d}_{2,p}(X_i, X_j) = \left[ m_{ij}^p + m_{ji}^p \right]^{1/p} / \left( x_{ij} + [m_{ij}^p + m_{ji}^p]^{1/p} \right).
\]

Then

1. \( \overline{d}_{2,p} \) is a normalised metric on \( X \);
2. \( \overline{d}_{2,1} = \overline{d}_{sd} \), (Jaccard metric)
3. \( \lim_{p \to \infty} \overline{d}_{2,p} = \overline{d}_{\text{max}} \) (maximum metric)
4. \( \overline{d}_{2,p}(X_i, X_j) = 1 \iff X_i \cap X_j = \emptyset \);
5. \( \overline{d}_{2,p} \) is monotone decreasing in \( p \).
Lemma

Reduction Lemma. The triangle inequality holds for $X_i, X_j, X_k$ if and only if it holds for the subsets $X'_i \subseteq X_i, X'_j \subseteq X_j$ and $X'_k \subseteq X_k$ where $X'_i = [X_i \setminus (X_i \cap X_j)] \cup (X_i \cap X_j \cap X_k)$, $X'_j = [X_j \setminus (X_i \cap X_j)] \cup (X_i \cap X_j \cap X_k)$ similarly, and $X'_k = (X_i \cap X_k) \cup (X_j \cap X_k)$.

(only if) $\bar{d}_{2,p}(X_i, X_j) \leq \bar{d}_{2,p}(X'_i, X'_j)$ since

$$[1 + x_{ij}/(m^p_{ij} + m^p_{ji})^{1/p}]^{-1} \leq [1 + x_{ijk}/(m^p_{ij} + m^p_{ji})^{1/p}]^{-1},$$

$\bar{d}_{2,p}(X'_i, X'_k) + \bar{d}_{2,p}(X'_j, X'_k) \leq \bar{d}_{2,p}(X_i, X_k) + \bar{d}_{2,p}(X_j, X_k)$ since

$$[1 + x_{ik}/((x^*_i)^p + y^p_{jk})^{1/p}]^{-1} \leq [1 + x_{ik}/(m^p_{ik} + m^p_{ki})^{1/p}]^{-1};$$

and by symmetry

$$[1 + x_{jk}/((x^*_j)^p + y^p_{ik})^{1/p}]^{-1} \leq [1 + x_{jk}/(m^p_{jk} + m^p_{kj})^{1/p}]^{-1}. \quad \square$$
Motivation

Minkowski-type Metrics for Sets

Normalising the Minkowski-type metrics for sets

Normalising the Minkowski-type Metric Family

Proof.

WTP $d_{2,p}(X'_i, X'_j) \leq d_{2,p}(X'_i, X'_k) + d_{2,p}(X'_j, X'_k)$.

Set $u = ((x_i^*)^p + y_{jk}^p)^{1/p}$, $v = ((x_j^*)^p + y_{ik}^p)^{1/p}$, $w = (m_{ij}^p + m_{ji}^p)^{1/p}$, so $w \leq u + v$. We need to show that

$$\frac{u}{u + y_{ik} + x_{ijk}} + \frac{v}{v + y_{jk} + x_{ijk}} \geq \frac{w}{w + x_{ijk}}.$$

Since $y_{ik} \leq v$ and $y_{jk} \leq u$, the LHS is at least

$$\frac{u}{u+v+x_{ijk}} + \frac{v}{v+u+x_{ij}} = \frac{u+v}{u+v+x_{ijk}}.$$

Since the function $t \mapsto \frac{t}{t+x_{ijk}}$ is increasing in $t$, LHS is $\geq \frac{w}{w+x_{ijk}}$ as required.
Reprise: Which measures separate genuine from imposter best?

<table>
<thead>
<tr>
<th>Score</th>
<th>Distance?</th>
<th>Metric?</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{d}_{sqrt}$</td>
<td>✓</td>
<td>✗</td>
<td>0.0486</td>
</tr>
<tr>
<td>$\overline{d}_{av}$</td>
<td>✓</td>
<td>✗</td>
<td>0.0232</td>
</tr>
<tr>
<td>$\overline{d}_{sd}$</td>
<td>✓</td>
<td>✓</td>
<td>0.0154</td>
</tr>
<tr>
<td>$\hat{d}_{22}$</td>
<td>✓</td>
<td>✗</td>
<td>−0.0007</td>
</tr>
<tr>
<td>$\overline{d}_{22}$</td>
<td>✓</td>
<td>✓</td>
<td>−0.0022</td>
</tr>
<tr>
<td>$\overline{d}_{min}$</td>
<td>✗</td>
<td>✗</td>
<td>−0.0139</td>
</tr>
<tr>
<td>$\overline{d}_{max}$</td>
<td>✓</td>
<td>✓</td>
<td>−0.0450</td>
</tr>
</tbody>
</table>

Comparison of nearest neighbour (NN) distances of vertex sets in VARIA retina database for 7 scoring functions
THANKYOU.....QUESTIONS?

References:
