MA-INF 4306 Lab Development and Application of Data Mining and Learning Systems: Big Data
A measure for (weak) convexity

Second Progress Meeting

Author: Timon Oerder, Mat.-Nr. 2679722

Institut für Informatik Uni Bonn

s6tioerd@uni-bonn.de

Table of contents

The top-level road-map

1. Define a measure for (weak) convexity

The top-level road-map

1. Define a measure for (weak) convexity
2. Test that measure on artificial data

The top-level road-map

1. Define a measure for (weak) convexity
2. Test that measure on artificial data
3. Test the measure on real-world data

Questions addressed since the last meeting

1. Sanity check for the convex hull Algorithm

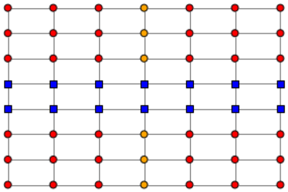
Questions addressed since the last meeting

1. Sanity check for the convex hull Algorithm
2. Can ϵ and θ be modified as a measure?

Questions addressed since the last meeting

1. Sanity check for the convex hull Algorithm
2. Can ϵ and θ be modified as a measure?
3. Idea for measures.

Sanity check for the convex hull Algorithm



(a) Testcase grid

(b) Testcase line

Figure: Application with $\theta = 3$, $\epsilon = 0.0$ and step distance;
class 0 is red; class 1 is blue; class 1 assigned as the convex hull of class 0 is orange

Can ϵ and θ be modified as a measure? (I)

In theory we can combine the notion of (θ, ϵ)-convexity as follows:
$A \subset X, z \in X, \epsilon \geq 0, \theta \geq 0.$
$z \in A \Leftrightarrow \exists x, y \in A:$
$d(x,z) + d(z,y) \leq d(x,y) + \epsilon \wedge d(x,z) + d(z,y) \geq d(x,y) - \epsilon \wedge d(x,y) \leq \theta$

Can ϵ and θ be modified as a measure? (I)

In theory we can combine the notion of (θ, ϵ)-convexity as follows:
$A \subset X, z \in X, \epsilon \geq 0, \theta \geq 0.$
$z \in A \Leftrightarrow \exists x, y \in A:$
$d(x,z) + d(z,y) \leq d(x,y) + \epsilon \wedge d(x,z) + d(z,y) \geq d(x,y) - \epsilon \wedge d(x,y) \leq \theta$

1. For $\epsilon = 0$ and $\theta \geq \underset{x,y \in A}{argmax}\ d(x,y)$ we get the standard notion for convexity.
   $(z \in A \Leftrightarrow \exists x, y \in A : d(x,z) + d(z,y) = d(x,y))$

Can ϵ and θ be modified as a measure? (I)

In theory we can combine the notion of (θ, ϵ)-convexity as follows:

$A \subset X, z \in X, \epsilon \geq 0, \theta \geq 0.$

$z \in A \Leftrightarrow \exists x, y \in A:$

$d(x,z) + d(z,y) \leq d(x,y) + \epsilon \wedge d(x,z) + d(z,y) \geq d(x,y) - \epsilon \wedge d(x,y) \leq \theta$

1. For $\epsilon = 0$ and $\theta \geq \underset{x,y \in A}{argmax}\ d(x,y)$ we get the standard notion for convexity.
   $(z \in A \Leftrightarrow \exists x, y \in A : d(x,z) + d(z,y) = d(x,y))$
2. For $\epsilon = 0$ and $\theta \in [0, \ldots \underset{x,y \in A}{argmax}\ d(x,y))$ we get θ-convexity [SHW21].
   $(z \in A \Leftrightarrow \exists x, y \in A : d(x,z) + d(z,y) = d(x,y) \wedge d(x,y) \leq \theta)$

Can ϵ and θ be modified as a measure? (I)

In theory we can combine the notion of (θ, ϵ)-convexity as follows:
$A \subset X, z \in X, \epsilon \geq 0, \theta \geq 0.$
$z \in A \Leftrightarrow \exists x, y \in A :$
$d(x,z) + d(z,y) \leq d(x,y) + \epsilon \wedge d(x,z) + d(z,y) \geq d(x,y) - \epsilon \wedge d(x,y) \leq \theta$

1. For $\epsilon = 0$ and $\theta \geq \underset{x,y \in A}{argmax}\ d(x,y)$ we get the standard notion for convexity.
   $(z \in A \Leftrightarrow \exists x, y \in A : d(x,z) + d(z,y) = d(x,y))$
2. For $\epsilon = 0$ and $\theta \in [0, \ldots \underset{x,y \in A}{argmax}\ d(x,y))$ we get θ-convexity [SHW21].
   $(z \in A \Leftrightarrow \exists x, y \in A : d(x,z) + d(z,y) = d(x,y) \wedge d(x,y) \leq \theta)$
3. ϵ "smears" the boundary of the convex hull. For a big enough ϵ every point $z \in X$ is contained in the convex hull.

Can ϵ and θ be modified as a measure? (I)

In theory we can combine the notion of (θ, ϵ)-convexity as follows:
$A \subset X, z \in X, \epsilon \geq 0, \theta \geq 0$.
$z \in A \Leftrightarrow \exists x, y \in A:$
$d(x,z) + d(z,y) \leq d(x,y) + \epsilon \wedge d(x,z) + d(z,y) \geq d(x,y) - \epsilon \wedge d(x,y) \leq \theta$

1. For $\epsilon = 0$ and $\theta \geq \underset{x,y \in A}{argmax}\ d(x,y)$ we get the standard notion for convexity.
   ($z \in A \Leftrightarrow \exists x, y \in A: d(x,z) + d(z,y) = d(x,y)$)
2. For $\epsilon = 0$ and $\theta \in [0, \dots \underset{x,y \in A}{argmax}\ d(x,y))$ we get θ-convexity [SHW21].
   ($z \in A \Leftrightarrow \exists x, y \in A: d(x,z) + d(z,y) = d(x,y) \wedge d(x,y) \leq \theta$)
3. ϵ "smears" the boundary of the convex hull. For a big enough ϵ every point $z \in X$ is contained in the convex hull.
4. If we find a "core" object $C \subset A$ which is the biggest convex subset, for what parameter ϵ and θ as inputs to a convex hull algorithm do we get A as result?

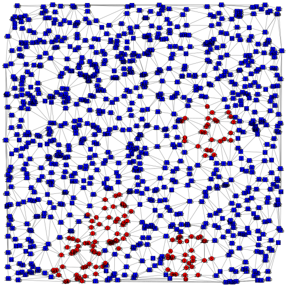
Can ϵ and θ be modified as a measure? (II)

How does the number of components and the number of vertices in the convex hull develop for different θs?

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;
class 0 is red; class 1 is blue; class 1 assigned as the convex hull of class 0 is orange

Can ϵ and θ be modified as a measure? (II)

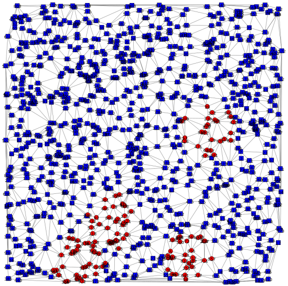
How does the number of components and the number of vertices in the convex hull develop for different θs?



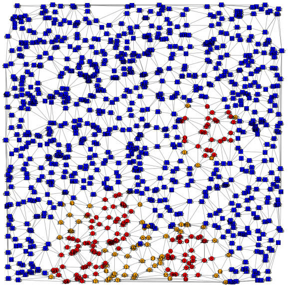
(a) $\theta = 0.1$

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;
class 0 is red; class 1 is blue; class 1 assigned as the convex hull of class 0 is orange

How does the number of components and the number of vertices in the convex hull develop for different θs?



(a) $\theta = 0.1$

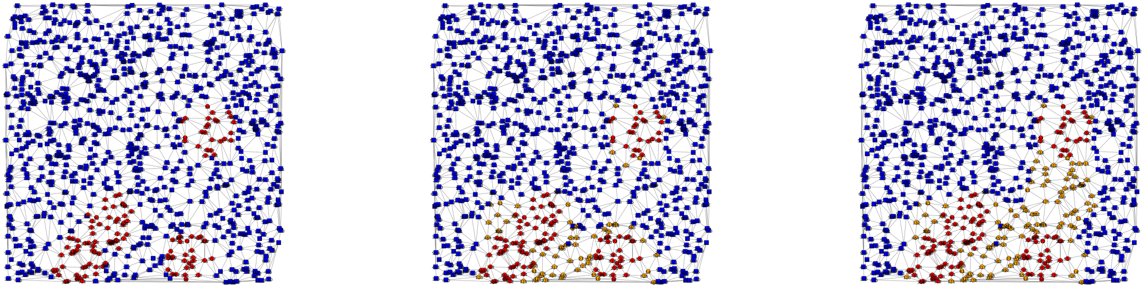


(b) $\theta = 2.32$

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;
class 0 is red; class 1 is blue; class 1 assigned as the convex hull of class 0 is orange

Can ϵ and θ be modified as a measure? (II)

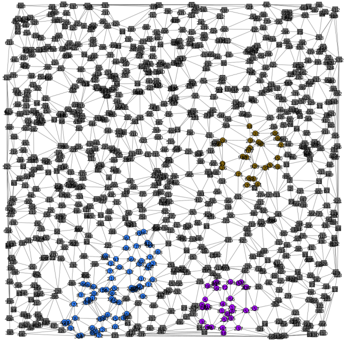
How does the number of components and the number of vertices in the convex hull develop for different θs?



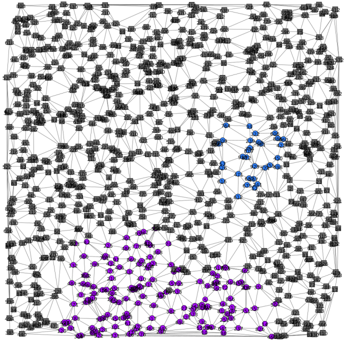
(a) $\theta = 0.1$ (b) $\theta = 2.32$ (c) $\theta = 4.54$

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;
class 0 is red; class 1 is blue; class 1 assigned as the convex hull of class 0 is orange

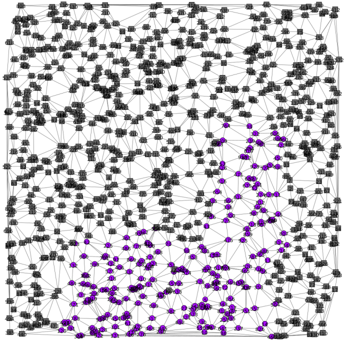
Can ϵ and θ be modified as a measure? (III)



(a) $\theta = 0.1$



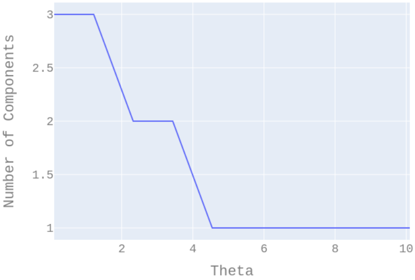
(b) $\theta = 2.32$



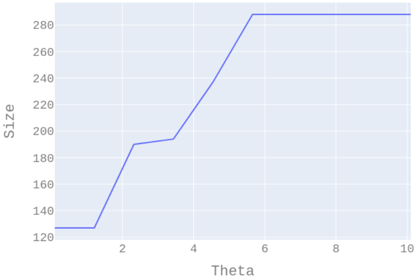
(c) $\theta = 4.54$

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;
vertices not part of a component and not part of the convex hull are grey, all other vertices are brown, blue or purple depending on their component.

Can ϵ and θ be modified as a measure? (IV)



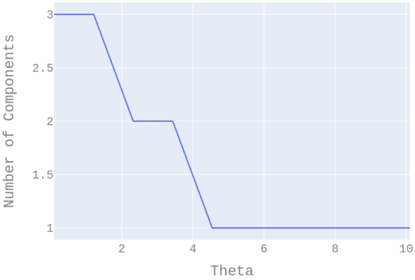
(a) Number of Components



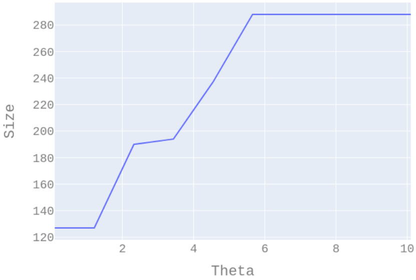
(b) Number of Vertices in the convex hull

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;

Can ϵ and θ be modified as a measure? (IV)



(a) Number of Components



(b) Number of Vertices in the convex hull

Figure: Application of the Algorithm with different θ, $\epsilon = 0.0$, $n = 1000$ vertices, 4 clusters and step distance;

No answers for now. Questions for the future: Does a core object exists? Is it unique?

Convexity Measure Proposal 1 (I)

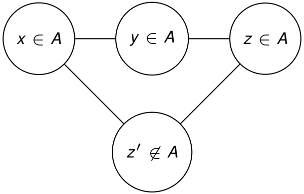
We can calculate a measure for convexity by looking at all the shortest paths between nodes and see how often we need to take a vertex outside of our set on such a path.

Convexity Measure Proposal 1 (I)

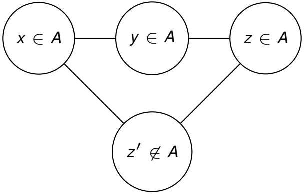
We can calculate a measure for convexity by looking at all the shortest paths between nodes and see how often we need to take a vertex outside of our set on such a path.
For $x, y \in V : sp(x, y) = \{\{k_1..k_n\} | \{k_1..k_n\}$are verticies on a shortest path from x to y$\}$.
$A \subset V : C_1(A) = \sum_{x,y \in A} (\frac{1}{|sp(x,y)|} \cdot \sum_{p \in sp(x,y)} |p \setminus A|)$

Convexity Measure Proposal 1 (I)

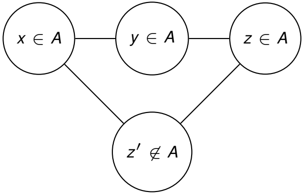
We can calculate a measure for convexity by looking at all the shortest paths between nodes and see how often we need to take a vertex outside of our set on such a path.
For $x, y \in V : sp(x, y) = \{\{k_1..k_n\}|\{k_1..k_n\}$are verticies on a shortest path from x to y$\}$.
$A \subset V : C_1(A) = \sum_{x,y \in A} (\frac{1}{|sp(x,y)|} \cdot \sum_{p \in sp(x,y)} |p \setminus A|)$



Convexity Measure Proposal 1 (II)

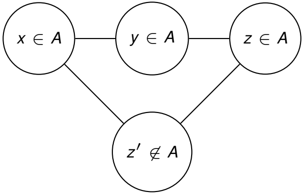


Convexity Measure Proposal 1 (II)



$$A = \{x, y, z\} \quad sp(x, y) = \{\{x, y\}\},\ sp(y, z) = \{\{y, z\}\},\ sp(x, z) = \{\{x, y, z\}, \{x, y, z'\}\}$$

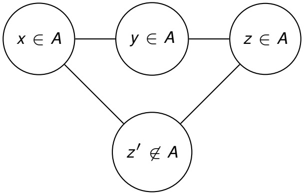
Convexity Measure Proposal 1 (II)



$$A = \{x, y, z\} \quad sp(x, y) = \{\{x, y\}\},\ sp(y, z) = \{\{y, z\}\},\ sp(x, z) = \{\{x, y, z\}, \{x, y, z'\}\}$$

$$C_1(A) = \frac{1}{|sp(x,y)|} \cdot \sum_{p \in sp(x,y)} |p \setminus A| + \frac{1}{|sp(y,z)|} \cdot \sum_{p \in sp(y,z)} |p \setminus A| + \frac{1}{|sp(x,z)|} \cdot \sum_{p \in sp(x,z)} |p \setminus A|$$

Convexity Measure Proposal 1 (II)



$A = \{x, y, z\} \quad sp(x, y) = \{\{x, y\}\},\ sp(y, z) = \{\{y, z\}\},\ sp(x, z) = \{\{x, y, z\}, \{x, y, z'\}\}$

$$C_1(A) = \frac{1}{|sp(x,y)|} \cdot \sum_{p \in sp(x,y)} |p \setminus A| + \frac{1}{|sp(y,z)|} \cdot \sum_{p \in sp(y,z)} |p \setminus A| + \frac{1}{|sp(x,z)|} \cdot \sum_{p \in sp(x,z)} |p \setminus A|$$

$$C_1(A) = \frac{1}{\underbrace{|\{\{x, y\}\}|}_{=1}} \cdot |\underbrace{\{x, y\} \setminus A}_{=\emptyset}| + \frac{1}{\underbrace{|\{\{y, z\}\}|}_{=1}} \cdot |\underbrace{\{y, z\} \setminus A}_{=\emptyset}| + \frac{1}{\underbrace{|\{\{x, y, z\}, \{x, y, z'\}\}|}_{=2}} \cdot$$

$$|\underbrace{\{x, y, z\} \setminus A}_{=\emptyset}| + \frac{1}{\underbrace{|\{\{x, y, z\}, \{x, y, z'\}\}|}_{=2}} \cdot |\underbrace{\{x, y, z'\} \setminus A}_{=\{z'\}}| = \frac{1}{2}$$

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Let P be the set of all points of a polygon on it‘s boundary and interior.
For a point $p \in P$: $V(p, P)$ defines the visible polygon as viewed from p and $A(V(p, P))$ the visible area. The area of the polygon will be denoted $A(P)$.

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Let P be the set of all points of a polygon on it's boundary and interior.
For a point $p \in P$: $V(p, P)$ defines the visible polygon as viewed from p and $A(V(p, P))$ the visible area. The area of the polygon will be denoted $A(P)$.
The sum of all visible areas for every point in the polygon: $AT(P) = \int\limits_{p \in P} A(V(p, P))dp$

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Let P be the set of all points of a polygon on it's boundary and interior.
For a point $p \in P$: $V(p, P)$ defines the visible polygon as viewed from p and $A(V(p, P))$ the visible area. The area of the polygon will be denoted $A(P)$.
The sum of all visible areas for every point in the polygon: $AT(P) = \int\limits_{p \in P} A(V(p, P))dp$

Probability density function: $f(p) = A(V(p, P))/AT(P)$

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Let P be the set of all points of a polygon on it's boundary and interior.
For a point $p \in P$: $V(p, P)$ defines the visible polygon as viewed from p and $A(V(p, P))$ the visible area. The area of the polygon will be denoted $A(P)$.
The sum of all visible areas for every point in the polygon: $AT(P) = \int_{p \in P} A(V(p, P))dp$

Probability density function: $f(p) = A(V(p, P))/AT(P)$
Polygonal Entropy $E(P) = -\int_{p \in P} f(p) ln(f(p))dp$

Convexity Measure Proposal 2 (I)

Polygonal Entropy: a convexity measure [Ste89]

How much of the polygon is visible from any point in it?
In a convex polygon, every point can see every other point.
$\Rightarrow$ What is the information this point gains/has about the rest of the polygon?

Let P be the set of all points of a polygon on it's boundary and interior.
For a point $p \in P : V(p, P)$ defines the visible polygon as viewed from p and $A(V(p, P))$ the visible area. The area of the polygon will be denoted $A(P)$.
The sum of all visible areas for every point in the polygon: $AT(P) = \int\limits_{p \in P} A(V(p, P))dp$

Probability density function: $f(p) = A(V(p, P))/AT(P)$
Polygonal Entropy $E(P) = -\int\limits_{p \in P} f(p)ln(f(p))dp$

Convexity of a polygon P is $C(P) = E(P)/E_{max}(P)$ with $E_{max}(P) = -ln(1/A(P))$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.
$$AT(P) = \sum_{p \in P} A(V_{all}(p, P))$$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.
$AT(P) = \sum_{p \in P} A(V_{all}(p, P)) = \sum_{p \in P} |\{p' \in P\}| = |P| \cdot |P|$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.
$AT(P) = \sum_{p \in P} A(V_{all}(p, P)) = \sum_{p \in P} |\{p' \in P\}| = |P| \cdot |P|$
$f(p) = A(V_{in}(p, P))/AT(P) = \frac{A(V_{in}(p,P))}{|P|^2}$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.
$AT(P) = \sum_{p \in P} A(V_{all}(p, P)) = \sum_{p \in P} |\{p' \in P\}| = |P| \cdot |P|$
$f(p) = A(V_{in}(p, P))/AT(P) = \frac{A(V_{in}(p,P))}{|P|^2}$
Entropy of a vertex set $E(P) = -\sum_{p \in P} f(p) ln(f(p))$

Convexity Measure Proposal 2 (II)

We can adapt that definition to graphs $G = (V, E)$:
Let p be a point in the set $P \subset V$.
For any $p \in P$ the Area of visible points of P in P is defined as
$A(V_{in}(p, P)) = |\{p' \in P | \exists path \in sp(p, p') : path \setminus P = \emptyset\}|$,
the Area of visible points of P in V (in the convex hull) is defined as
$A(V_{all}(p, P)) = |\{p' \in P\}|$
$A(P) = |P|$ is analog defined as the cardinality of the set.
$AT(P) = \sum_{p \in P} A(V_{all}(p, P)) = \sum_{p \in P} |\{p' \in P\}| = |P| \cdot |P|$
$f(p) = A(V_{in}(p, P))/AT(P) = \frac{A(V_{in}(p,P))}{|P|^2}$
Entropy of a vertex set $E(P) = -\sum_{p \in P} f(p) ln(f(p))$
Convexity of a set of vertices is $C_2(P) = E(P)/E_{max}(P)$ with $E_{max}(P) = -ln(1/A(P))$

First Results (I)

(a) $C_1(P) = 0$, $C_2(P) = 1$

(b) $C_1(P) \approx 1.3$, $C_2(P) \approx 0.65$

Figure: Step distance

First Results (II)

(a) $\theta = 2$, $C_1(P) \approx 1.090$, $C_2(P) \approx 0.65$, convex

(b) $\theta = 3$, $C_1(P) = 0$, $C_2(P) = 1$

Figure: Step distance; Results for measures after application of the convex hull algorithm

Next steps

- Can ϵ and θ be modified as a measure?

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?
- Any other measures to test?

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?
- Any other measures to test?
- How do the measures behave relate to the notion of weak convexity?

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?
- Any other measures to test?
- How do the measures behave relate to the notion of weak convexity?
- Find repository data sets to test the measure(s) on

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?
- Any other measures to test?
- How do the measures behave relate to the notion of weak convexity?
- Find repository data sets to test the measure(s) on
- Test the measure(s)

Next steps

- Can ϵ and θ be modified as a measure?
- Can we find a unique core?
- Any other measures to test?
- How do the measures behave relate to the notion of weak convexity?
- Find repository data sets to test the measure(s) on
- Test the measure(s)

Thank you for your attention!

What are your questions?

References

Eike Stadtländer, Tamás Horváth, and Stefan Wrobel.
Learning weakly convex sets in metric spaces.
In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*, page 200–216, Berlin, Heidelberg, 2021. Springer-Verlag.

Helman I. Stern.
Polygonal entropy: A convexity measure.
Pattern Recognition Letters, 10(4):229–235, 1989.