Matching Node Embeddings using Valid Assignment Kernels

Giannis Nikolentzos, Changmin Wu and Michalis Vazirgiannis

DaSciM, Lix, Ecole Polytechnique

May 31, 2018

Overview

Introduction

Graph Comparison

Graph

$G := (V, E)$ an ordered pair compromising a set V of vertices and a set E of edges.

Problem: Graph Comparison

Given two graphs $G, G' \in \mathcal{G}$, find a mapping s

$$s := \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}$$

where $s(G, G')$ measures the similarity between G, G'

Applications: Graph Classification / Clustering

Computational Biology, Information Retrieval, Cybersecurity...

Graph Kernel

Graph Kernel

K is a positive definite (pd) function $\mathcal{G} \times \mathcal{G} \to \mathbb{R}$ defined on $\mathcal{G}$ with a corresponding Hilbert space $\mathcal{H}$, inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a map $\phi : \mathcal{G} \to \mathcal{H}$ such that:

$$k(G, G') = \langle \phi(G), \phi(G') \rangle_{\mathcal{H}} \quad \forall G, G' \in \mathcal{G}$$

Graph Kernel

R-Convolution kernel

$$K_{convolution}(G, G') = \sum_{(x,G)\in\mathcal{R}} \sum_{(x',G')\in\mathcal{R}} k_{part}(x, x')$$

- $\mathcal{R}$ is the decomposition of graph
- k_{part} is usually a simple function, *i.e.*

$$\begin{aligned} k_{part}(x, x') &= 1 \quad \text{if } x, x' \text{ isomorphic} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Graph Kernel

Optimal Assignment Kernel

$K_{assignment} : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ is defined for every $G, G' \in \mathcal{G}$ as

$$K_{assignment}(G, G') = \begin{cases} \max_{\pi \in S_{|x|}} \sum_{i=1}^{|x|} k_{base}(x_i, x'_{\pi(i)}) & \text{if } |g'| > |g|, \\ \max_{\pi \in S_{|x'|}} \sum_{i=1}^{|x'|} k_{base}(x_{\pi(i)}, x'_i) & \text{otherwise.} \end{cases}$$

- $(x_1, x_2, \ldots, x_n)$ decomposition of G, n denoted as $|x|$
- $S_{|x|}$ a permutation of $|x|$ elements

Graph Kernel

Advantages

- Reveal structural correspondence between two graphs
- Do not suffer from diagonal dominance problem

However...

Theorem: [Vert, 2008] *The optimal assignment kernel is not always positive definite.*

Valid Optimal Assignment Kernel [Kriege, 2016]

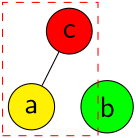
Hierarchy

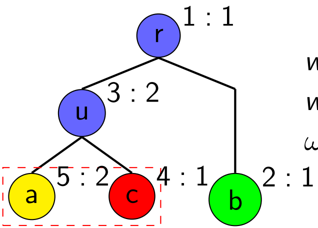
Let T be a rooted tree such that the leaves of T are the elements of $\mathcal{X}$. Each inner vertex v in T will correspond to a subset of $\mathcal{X}$ compromising all the leaves of the subtrees rooted at v. Let $w : V(T) \to \mathbb{R}_0^+$ a weight function such that $w(v) \geq w(parent(v))$ for all $v \in T$. (T, w) is referred as a hierarchy on $\mathcal{X}$.

Hierarchy-induced Kernel

Let $H = (T, w)$ be a hierarchy on $\mathcal{X}$, then the function defined as $k(x, y) = w(LCA(x, y))$ for all $x, y \in \mathcal{X}$ is the kernel on $\mathcal{X}$ induced by H. $LCA(\cdot, \cdot)$ refers to Least Common Ancestor.

Valid Optimal Assignment Kernel





$w(v) : \omega(v)$

$w(v) \geq w(parent(v))$

$\omega(v) = w(v) - w(parent(v))$

Feature map

A map $\phi : \mathcal{X} \to \mathbb{R}^t$, $t := |V(T)|$ defined as

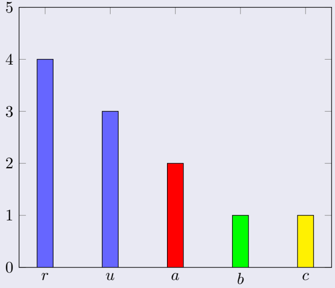
$$[\phi(x)]_v = \begin{cases} \sqrt{\omega(v)} & \text{if } v \in p(x) \\ 0 & \text{otherwise.} \end{cases}$$

	r	u	a	b	c
$\phi(a)$	$\sqrt{1}$	$\sqrt{2}$	$\sqrt{2}$	0	0
$\phi(b)$	$\sqrt{1}$	0	0	$\sqrt{1}$	0
$\phi(c)$	$\sqrt{1}$	$\sqrt{2}$	0	0	$\sqrt{1}$

Valid Optimal Assignment Kernel

Histogram

$$H^k(X) = \sum_{x \in \mathcal{X}} \phi(x) \circ \phi(x)$$



Strong Kernel

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is called strong kernel if

$$k(x, y) \geq \min\{k(x, z), k(z, y)\} \quad \forall x, y, z \in \mathcal{X}$$

Valid Optimal Assignment Kernel

Theorem 1

A kernel k on $\mathcal{X}$ is strong if and only if it is induced by a hierarchy on $\mathcal{X}$.

Theorem 2

Let k be a strong base kernel and histogram H^k defined as previous, then the optimal assignment kernel $K_{\mathcal{B}}^k(X, Y) = K_{\sqcap}(H^k(X), H^k(Y))$ for all $X, Y \in [\mathcal{X}]^n$.
where $K_{\sqcap}$ is the histogram intersection kernel defined as

$$K_{\sqcap}(g, h) = \sum_{i=1}^{t} \min([g]_i, [h]_i)$$

Collary

If the base kernel k is strong, than $K_{\mathcal{B}}^k$ is valid.

Pyramid Matching Kernel [Nikolentzos, 2017]

Basic idea

Matching vector representations of the vertices of two graphs:

- Bag-of-Vectors representation of graph
- map these vectors to multi-resolution histograms, and compare with a weighted histogram intersection measure
- histogram construction: partitioning the embedding space into grid regions of increasingly larger size

$$K_{\triangle}(G, G') = I(H_G^L, H_{G'}^L) + \sum_{l=0}^{L-1} \frac{1}{2^{l-1}} (I(H_G^l, H_{G'}^l) - I(H_G^{l+1}, H_{G'}^{l+1}))$$

Hierarchy Construction

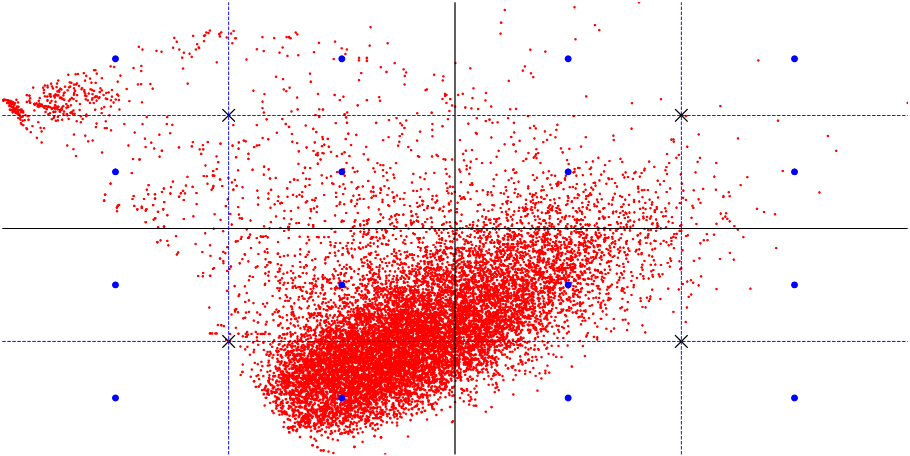


Figure: Illustration of Grid Partition. Data points from IMDB-MULTI(Node embeddings projected to 2D space)

Embeddings Optimal Assignment Kernel

Embeddings Optimal Assignment Kernel

Adjacency matrix

For graph $G = (V, E)$, its adjacency matrix $A_{|V|\times|V|}$ is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

Embedding of nodes

Given a graph $G = (V, E)$, its node embeddings are generated by the eigenvectors of adjacency matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ with row vectors of $\mathbf{U}$ as representations of nodes.

Kernel function

$$K_{\mathfrak{B}}^{k}(\mathcal{X}, \mathcal{X}') = \max_{B \in \mathfrak{B}(\mathcal{X}, \mathcal{X}')} \sum_{(\mathbf{x}, \mathbf{x}') \in B} k(\mathbf{x}, \mathbf{x}')$$

Hierarchy Construction

- Hierarchical clustering to create irregular multi-resolution partition
- Spherical K-Means
  - K-Means operates on an unit sphere (embeddings are normalized to unit norm)
  - Objective function: $\arg\max_{\mathcal{C}} Q(\{\mathcal{C}_i\}_{i=1}^{k}) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \langle \mathbf{x}, \mathbf{c}_i \rangle$
  - Advantage: directly optimize the similarity (inner product) between nodes
  - Advantage: faster than hierarchical clustering (agglomerative) with reasonable memory requirements

Hierarchy Construction

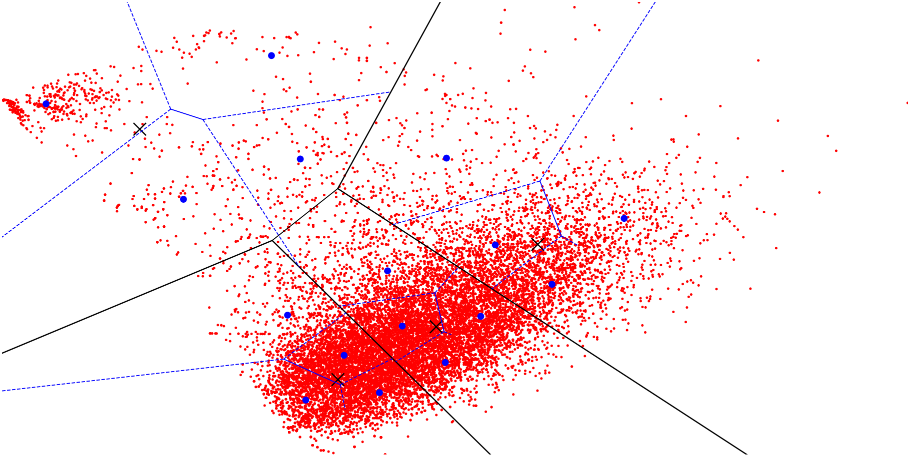


Figure: Illustration of Hierarchical Clustering. Data points from IMDB-MULTI(Node embeddings projected to 2D space)

Hierarchy Construction

Algorithm 1: Spherical KMeans for Hierarchy Construction

Data: $\mathbf{X}, K, L$

Result: Adjacency List of Nodes

initialization;

while $i <= L$ **do**

  if $i{=}{=}0$ **then**

    Apply S-KMeans(K) on $\mathbf{X}$;

    Note clusters as $C_j^0, j = 1, \ldots, K$;

    Note centroids as c_j^0;

  else

    for *every* C_j^{L-1} **do**

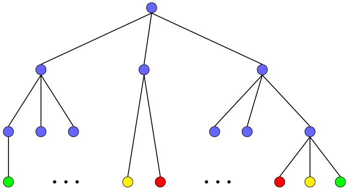
      Apply S-KMeans(K);

      $\forall x \in C_k^L, parent(x) = c_k^L$;

    end

  end

end

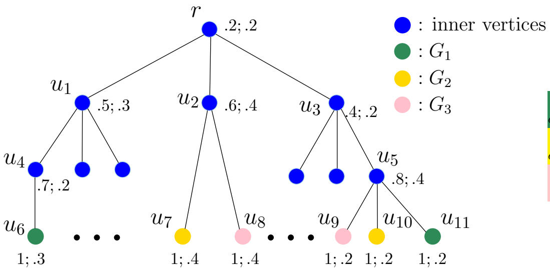


Weight function

For inner node v that corresponds to a cluster $\mathcal{C}$ of data points, its weight is set equal to:

$$w(v) = \min_{x \in \mathcal{C}} \langle \mathbf{x}, \mathbf{c} \rangle$$

Its corresponding feature value $\omega(v) = w(v) - w(\mathbf{c})$

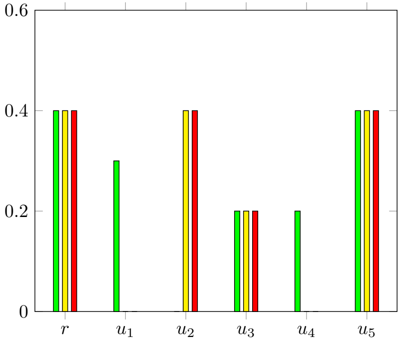


	r	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}
$\phi(u_6)$	$\sqrt{.2}$	$\sqrt{.3}$	0	0	$\sqrt{.2}$	0	$\sqrt{.3}$	0	0	0	0	0
$\phi(u_{11})$	$\sqrt{.2}$	0	0	$\sqrt{.2}$	0	$\sqrt{.4}$	0	0	0	0	0	$\sqrt{.4}$
$\phi(u_7)$	$\sqrt{.2}$	0	$\sqrt{.4}$	0	0	0	0	$\sqrt{.4}$	0	0	0	0
$\phi(u_{10})$	$\sqrt{.2}$	0	0	$\sqrt{.2}$	0	$\sqrt{.4}$	0	0	0	0	$\sqrt{.4}$	0
$\phi(u_8)$	$\sqrt{.2}$	0	$\sqrt{.4}$	0	0	0	0	0	$\sqrt{.4}$	0	0	0
$\phi(u_9)$	$\sqrt{.2}$	0	0	$\sqrt{.2}$	0	$\sqrt{.4}$	0	0	0	$\sqrt{.4}$	0	0

Figure: An example of a hierarchy where each vertex v is annotated by its weights $w(v) : \omega(v)$ and its color indicates the graph to which it belongs (left), and the derived feature vectors (right).

Weight function

	r	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}
G_1	.4	.3	0	.2	.2	.4	.3	0	0	0	0	.4
G_2	.4	0	.4	.2	0	.4	0	.4	0	0	.4	0
G_3	.4	0	.4	.2	0	.4	0	0	.4	.4	0	0



	G_1	G_2	G_3
G_1	1.5	1.0	1.0
G_2	1.0	1.4	1.4
G_3	1.0	1.4	1.4

Analysis

Theorem

Let $\mathcal{C}$ be the set of points of a cluster and $\mathbf{c}$ its centroid. Let also $\mathbf{x}, \mathbf{y}$ be any two points of $\mathcal{C}$. Then, it holds that

$$\langle \mathbf{x}, \mathbf{y} \rangle \geq 4 \min_{\mathbf{z} \in \mathcal{C}} \langle \mathbf{z}, \mathbf{c} \rangle - 3$$

For clusters at low levels (where inner products between datapoints are high), the bound become tight and as we aim to maximize the similarity $k(\mathbf{x}, \mathbf{y})$, Our method offers good approximation to the objective function: $\max_{B \in \mathfrak{B}(\mathcal{X}, \mathcal{X}')} \sum_{(\mathbf{x}, \mathbf{x}') \in B} k(\mathbf{x}, \mathbf{x}')$

An Variant of EOA: EOA-SP

- Using K-Means instead of Spherical K-Means
- Weight function w set as the depth of the node: $w(v) = path\_length(v, root)$ for all $v \in V(T)$
- Feature value ω computed as

$$\omega(v) = \frac{w(parent(v))}{w(v)}$$

which assures the weights of children are always greater than these of their parents.

Experimental Evaluation

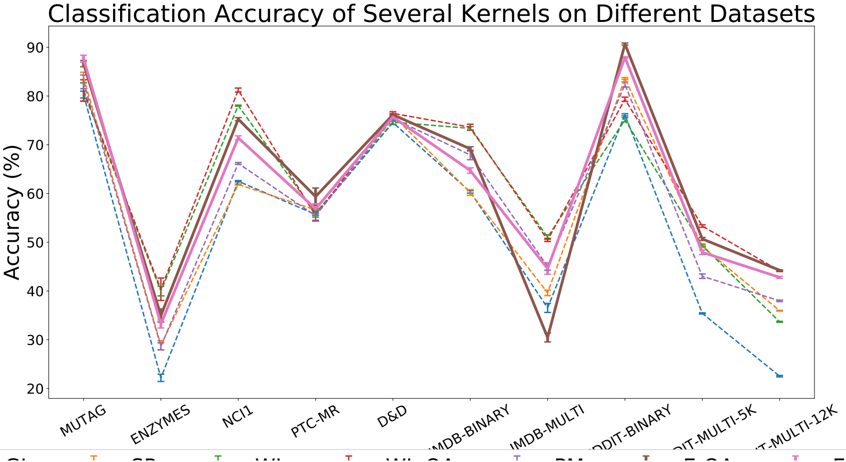
Graph Classification

Method \ Datasets	MUTAG	ENZYMES	NCI1	PTC-MR	D&D
GL	80.29 (± 0.70)	22.18 (± 0.74)	62.52 (± 0.14)	55.71 (± 0.19)	74.55 (± 0.36)
SP	83.79 (± 1.09)	28.86 (± 0.94)	61.85 (± 0.11)	56.63 (± 0.59)	76.02 (± 0.37)
WL	80.84 (± 1.87)	39.98 (± 0.98)	78.03 (± 0.10)	55.99 (± 0.84)	74.65 (± 0.47)
WL-OA	81.13 (± 2.20)	**40.36** (± 2.30)	**81.22** (± 0.41)	55.47 (± 0.98)	**76.44** (± 0.33)
PM	82.90 (± 1.40)	28.65 (± 0.72)	66.17 (± 0.19)	55.44 (± 1.12)	75.40 (± 0.60)
E-OA-SP	86.64 (± 0.64)	34.98 (± 1.34)	75.25 (± 0.32)	**59.37** (± 1.76)	76.15 (± 0.22)
E-OA	**87.64** (± 0.73)	33.23 (± 0.82)	71.41 (± 0.43)	56.85 (± 1.05)	75.69 (± 0.21)

Method \ Datasets	IMDB BINARY	IMDB MULTI	REDDIT BINARY	REDDIT MULTI-5K	REDDIT MULTI-12K
GL	60.33 (± 0.25)	36.53 (± 0.93)	76.15 (± 0.21)	35.41 (± 0.12)	22.52 (± 0.15)
SP	60.21 (± 0.58)	39.62 (± 0.57)	83.60 (± 0.18)	49.13 (± 0.14)	35.96 (± 0.08)
WL	73.36 (± 0.38)	**51.06** (± 0.47)	75.12 (± 0.44)	49.33 (± 0.28)	33.68 (± 0.10)
WL-OA	**73.61** (± 0.60)	50.48 (± 0.33)	79.34 (± 0.43)	**53.33** (± 0.25)	44.12 (± 0.13)
PM	67.91 (± 0.98)	45.03 (± 0.77)	82.35 (± 0.52)	43.04 (± 0.46)	37.98 (± 0.16)
E-OA-SP	69.16 (± 0.43)	30.47 (± 0.92)	**90.67** (± 0.21)	50.68 (± 0.31)	**44.26** (± 0.08)
E-OA	64.71 (± 0.56)	44.58 (± 1.16)	87.92 (± 0.12)	47.94 (± 0.47)	42.80 (± 0.22)

Table: Classification accuracy (± standard deviation), averaged on 10 iterations. Model is optimized using 10-fold cross validation.

Graph Classification



Text Categorization

Method	BBCSport	Subjectivity	Polarity	TREC	Twitter
BOW TF-IDF	98.38	90.67	77.14	97.00	75.12
CR	**99.59**	90.90	77.79	96.60	72.65
RAND-OA	96.08	89.89	75.72	97.00	75.25
E-OA-SP	99.05	91.25	76.96	97.00	75.41
E-OA	99.45	**91.92**	**77.87**	**97.80**	**76.34**

Table: Classification accuracy of the 3 variants of the proposed kernel (using pre-trained and randomly initialized embeddings), the bag-of-words representation with tf-idf weights (BOW TF-IDF) and the centroid representation (CR) on the 5 text categorization datasets.

Conclusion

Conclusion and future works

- What we did?
  - A kernel comparing sets of vectors (node embeddings)
  - Achieve good performance on graph classification and text categorization tasks with respect to state-of-the-art methods
- What could be next?
  - apply method on labeled graphs
  - find more stable hierarchical clustering method
  - find better parameters(hierarchy tree depth, branching width, ...)
  - find better node embeddings

References

Jean-Philippe Vert (2008)
The optimal assignment kernel is not positive definite
arXiv preprint arXiv:0801.4061

Nils.M.Kriege, Pierre-Louis Giscard and Richard Wilson (2016)
On Valid Optimal Assignment Kernels and Applications to Graph Classification
Advances in Neural Information Processing Systems, p1623–1631

G. Nikolentzos, P.Meladianos and M.Vazirgiannis (2017)
Matching Node Embeddings for Graph Similarity
Proceedings of the 31st AAAI Conference in Artificial Intelligence, p1891–1901

Thank you!