GRAPHLET KERNELS FOR VERTEX CLASSIFICATION

Presenter: José Lugo-Martínez
Phd Candidate
jlugomar@indiana.edu

Jun 12, 2015
Overview of classification problems on graphs

Graphlet kernels for vertex classification

Case study: Structure-based functional residue prediction
  – Inferring molecular mechanisms of disease (if time permits)
Graph Classification

Vertex Classification

Edge Classification

Link Prediction

Task: Classify graph as +1 or -1

Human lymphocyte kinase
**Classification Problems on Graphs**

**Vertex (or Edge) Classification**

**Task:** Classify node (or edge) as +1 or -1

Y394 of human lymphocyte kinase

Depth-3 graph neighborhood for Y394

\[ C_{\alpha} - C_{\alpha} \leq 6\text{Å} \]
**Semi-Supervised Learning Scenario**

Objective: Predict class label for each unlabeled node

\[ G = (V, E, \Sigma) \]
\[ \Sigma = \{A, B\} \]
\[ f : V \rightarrow \Sigma \]

Training Data

Neighborhood graph

Research Question

How to measure similarity between rooted neighborhoods?
**Problem Statement**

Given two neighborhood graphs \( N(u), N(v) \) from a space of graphs \( \mathcal{G} \). The problem of rooted neighborhood comparison is to find a mapping

\[
s : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}
\]

s.t. \( s(N(u), N(v)) \) quantifies the similarity of \( N(u) \) & \( N(v) \)

**Task:** Design meaningful similarity measures between vertex neighborhoods
Graph Kernels

- Define kernel functions on pair of graphs $G$ and $G'$
  \[ k(G, G') = \langle \phi(G), \phi(G') \rangle \]

- Kernel matrix $K$ such that
  \[ K_{i,j} = k(G_i, G_j), 1 \leq i, j \leq D \]

- Properties of $K$
  
  I. Symmetric
  
  II. Positive semi-definite
**Methodology Overview**

If $Pr(+1|v) > 0.5$ then $t(v) = +1$

else $t(v) = -1$
**Graph Kernels Research in a Nutshell**

- **Diffusion kernels**
  - Kondor & Lafferty (2002)

- **Focus on counting graph substructures**

- **Three categories based on**
  - walks and paths
    - Kashima et al. (2003), Borgwardt & Kriegel (2005)
  - subtree patterns
    - Hido & Kashima (2009), Shervashidze et al. (2011)
  - subgraphs
    - Shervashidze et al. (2009), Vacic et al. (2010)

How about other factors?
Graphlet Kernels
An *n-graphlet* is a small \((n \leq 5)\) connected rooted subgraph.
BASE GRAPHLETS

Undirected:

Directed:

Labeled Graphlets

vertex labels alphabet

\[ \Sigma = \{A, B\} \]

\[ \kappa(n, \Sigma) = \kappa(3, \Sigma) = 20 \]

\[ |\Sigma|^n = 2^3 = 8 \]

\[ |\Sigma| \cdot (\frac{|\Sigma|+1}{|\Sigma|-1}) = 6 \]

same symmetry class

\[ |\Sigma| \cdot (\frac{|\Sigma|+1}{|\Sigma|-1}) = 6 \]
**Graphlet Kernel Example**

Graphlet kernel, N=4

\[
k(u, v) = \sum_{n=1}^{N} k_n(u, v)
\]

\[
k(3, \Sigma) = 20
\]

\[
\phi_3(u) = \begin{bmatrix} AAA & AAB & ABA & ABB & BAA & BAB & BBA & BBB & AAA & AAB & ABB & BAA & BAB & BBA & BBB \end{bmatrix}^T
\]

\[
\phi_3(v) = \begin{bmatrix} 1 & 1 \end{bmatrix}^T
\]

\[
k_3(u, v) = \langle \phi_3(u), \phi_3(v) \rangle = 3
\]
### How Many Labeled Graphlets?

\[
\kappa(n, \Sigma) = \sum_{i=1}^{|S(n)|} m_i(n, \Sigma) \cdot |S_i(n)|
\]

| \( n \) | \( |\Sigma| = 1 \) | \( |\Sigma| = 20 \) | \( |\Sigma| = 1 \) | \( |\Sigma| = 20 \) |
|---|---|---|---|---|
| 1 | 1 | 20 | 1 | 20 |
| 2 | 1 | 400 | 3 | 1,200 |
| 3 | 3 | 16,400 | 30 | 217,200 |
| 4 | 11 | 1,045,600 | 697 | 102,673,600 |
| 5 | 58 | 100,168,400 | 44,907 | 137,252,234,400 |

*base graphlets*

*labeled graphlets*
LIMITATIONS OF GRAPHLET KERNEL

• Exact matches less likely as alphabet size increases

• Can’t handle misannotated labels or missing edges
  – e.g. protein 3D structures can be noisy and incomplete

• Ineffective for evolving graph neighborhoods
  – e.g. closely relate protein structures

Goal: Design robust kernels in the presence of noisy and incomplete data
**Edit Distance Graphlet Kernels**

Generalize the concept of counting graphlets

Incorporate flexibility in counting via edit distance

**Definition (Graph Edit Distance)** Given two vertex- and/or edge-labeled graphs G and H. The edit distance between these graphs corresponds to the minimum number of edit operations necessary to transform G into H.

- Allowed edit operations include insertion or deletion of vertices and edges, or in the case of labeled graphs, substitutions of vertex and edge labels
- Any sequence of edit operations that transforms G into H is referred to as an edit path
- Thus, the graph edit distance between G and H corresponds to the length of the shortest edit path between them

EDIT DISTANCE OPERATIONS

Incorporate flexibility in counting via edit distance

**Vertex label substitutions**

**Edge insertions or deletions**

symmetric
Example Revisited

\[ \phi_{(3,1)}(u) : \]

\[ \phi_{(3,1)}(u) : \]

\[ \phi_{(3,1)}(v) : \]

1-label substitution
**Label Substitution Kernel**

\[
\phi^l_{(3,1)}(u) : \begin{array}{cccccccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
2 & 3_1 & 3_2 & 3_3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

\[
\phi^l_{(3,1)}(v) : \begin{array}{cccccccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

1-label substitution
**Label Substitution Kernel**

\[ \phi^l_{(3,1)}(u) : \]

\[
\begin{array}{cccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\[ \phi^l_{(3,1)}(v) : \]

\[
\begin{array}{cccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\[ k^l_{(3,1)}(u, v) = \left\langle \phi^l_{(3,1)}(u), \phi^l_{(3,1)}(v) \right\rangle = 15 \]
**Edge Indels Kernel**

\[
\phi^e_{(3,1)}(u) : \begin{array}{cccccccccccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
\end{array}
\]

\[
\phi^e_{(3,1)}(v) : \begin{array}{cccccccccccccccccccc}
\text{AAA} & \text{AAB} & \text{ABA} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} & \text{AAA} & \text{AAB} & \text{ABB} & \text{BAA} & \text{BAB} & \text{BBA} & \text{BBB} \\
1 & 1 & & & & & & & & & & & & & & 1 & & & & & & & & 1
\end{array}
\]
**Edge Indels Kernel**

\[
\phi^e_{(3,1)}(u) : \begin{array}{cccccccccccccccc}
AAA & AAB & ABA & ABB & BAA & BAB & BBA & BBB & AAA & AAB & ABB & BAA & BAB & BBA & BBB \\
3 &  &  &  &  &  &  &  & 1 &  &  &  &  &  & 1 \\
\end{array}
\]

\[
\phi^e_{(3,1)}(v) : \begin{array}{cccccccccccccccc}
AAA & AAB & ABA & ABB & BAA & BAB & BBA & BBB & AAA & AAB & ABB & BAA & BAB & BBA & BBB \\
1 & 1 &  &  &  &  &  &  & 1 &  &  &  &  &  &  \\
\end{array}
\]
**EDGE INDELS KERNEL**

\[ \phi_{(3,1)}^e(u) : \begin{array}{cccccccccccccccc}
3 & 1 & & & 1 & & & 1 & & & & & & & & & & & & \\
\end{array} \]

\[ \phi_{(3,1)}^e(v) : \begin{array}{cccccccccccccccc}
2 & 1 & & & 1 & & & 2 & 1 & & & & & & & & & & & & \\
\end{array} \]

\[ k_{(3,1)}^e(x, y) = \left\langle \phi_{(3,1)}^e(x), \phi_{(3,1)}^e(y) \right\rangle = 9 \]
Edit Distance Kernels

\[ k_m(u, v) = \sum_{n=1}^{N} k_{(n,m)}(u, v) \]

\[ k'(u, v) = \frac{k_m(u, v)}{\sqrt{k_m(u,u) \cdot k_m(v,v)}} \]

# of edit distance operations

Edit distance graphlet kernel

Normalized edit distance kernel

Case Study: Structure-based functional residue prediction

Joint work with Vikas Pejaver, Matthew Mort, David N. Cooper, Sean D. Mooney and Predrag Radivojac
PREDICTION OF FUNCTIONAL SITES FROM PROTEIN STRUCTURES

A. Zinc-binding site
B. Phosphorylation site
C. DNA-binding site

RESULTS: METAL BINDING RESIDUES

Iron (Fe)

Copper (Cu)

1 - sp
## Multiple Functional Residue Predictors

<table>
<thead>
<tr>
<th>Category</th>
<th>Site type</th>
<th># of chains</th>
<th>Positives</th>
<th>Negatives</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metal binding</strong></td>
<td>Calcium (Ca)</td>
<td>1092</td>
<td>4860</td>
<td>4860</td>
<td>0.830</td>
</tr>
<tr>
<td></td>
<td>Cadmium (Cd)</td>
<td>199</td>
<td>1003</td>
<td>1003</td>
<td>0.713</td>
</tr>
<tr>
<td></td>
<td>Cobalt (Co)</td>
<td>156</td>
<td>532</td>
<td>532</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>Copper (Cu)</td>
<td>105</td>
<td>440</td>
<td>440</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>Iron (Fe)</td>
<td>187</td>
<td>785</td>
<td>785</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>Potassium (K)</td>
<td>287</td>
<td>978</td>
<td>978</td>
<td>0.647</td>
</tr>
<tr>
<td></td>
<td>Magnesium (Mg)</td>
<td>1348</td>
<td>3282</td>
<td>3282</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>Manganese (Mn)</td>
<td>366</td>
<td>1344</td>
<td>1344</td>
<td>0.870</td>
</tr>
<tr>
<td></td>
<td>Sodium (Na)</td>
<td>961</td>
<td>2753</td>
<td>2753</td>
<td>0.620</td>
</tr>
<tr>
<td></td>
<td>Nickel (Ni)</td>
<td>254</td>
<td>680</td>
<td>680</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td>Zinc (Zn)</td>
<td>1307</td>
<td>5778</td>
<td>5778</td>
<td>0.882</td>
</tr>
<tr>
<td><strong>PTMs</strong></td>
<td>N-glycosylation (Nglyco)</td>
<td>339</td>
<td>736</td>
<td>736</td>
<td>0.801</td>
</tr>
<tr>
<td></td>
<td>Phosphorylation (Phos)</td>
<td>655</td>
<td>1157</td>
<td>1157</td>
<td>0.754</td>
</tr>
<tr>
<td><strong>Catalytic activity</strong></td>
<td>Catalytic residues (Cat)</td>
<td>721</td>
<td>2224</td>
<td>13356</td>
<td>0.881</td>
</tr>
<tr>
<td><strong>Macromolecule binding</strong></td>
<td>DNA</td>
<td>139</td>
<td>3791</td>
<td>6448</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>RNA</td>
<td>83</td>
<td>3436</td>
<td>3229</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td>PPI</td>
<td>112</td>
<td>4350</td>
<td>11790</td>
<td>0.654</td>
</tr>
<tr>
<td></td>
<td>PPI Hotspots</td>
<td>36</td>
<td>165</td>
<td>184</td>
<td>0.655</td>
</tr>
<tr>
<td><strong>Ligand binding</strong></td>
<td>ADP</td>
<td>163</td>
<td>2589</td>
<td>2589</td>
<td>0.786</td>
</tr>
<tr>
<td></td>
<td>ATP</td>
<td>105</td>
<td>1733</td>
<td>1733</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>BME</td>
<td>107</td>
<td>828</td>
<td>828</td>
<td>0.646</td>
</tr>
<tr>
<td></td>
<td>FAD</td>
<td>81</td>
<td>2248</td>
<td>2248</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td>FMN</td>
<td>43</td>
<td>788</td>
<td>788</td>
<td>0.804</td>
</tr>
<tr>
<td></td>
<td>GDP</td>
<td>46</td>
<td>593</td>
<td>593</td>
<td>0.838</td>
</tr>
<tr>
<td></td>
<td>GTP</td>
<td>23</td>
<td>366</td>
<td>366</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>HEM</td>
<td>83</td>
<td>2246</td>
<td>2235</td>
<td>0.730</td>
</tr>
<tr>
<td></td>
<td>NAD</td>
<td>74</td>
<td>1663</td>
<td>1663</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>NAG</td>
<td>237</td>
<td>2520</td>
<td>2520</td>
<td>0.684</td>
</tr>
<tr>
<td></td>
<td>PLP</td>
<td>34</td>
<td>477</td>
<td>477</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td>PO4</td>
<td>516</td>
<td>3957</td>
<td>3957</td>
<td>0.698</td>
</tr>
<tr>
<td></td>
<td>SO4</td>
<td>1355</td>
<td>12005</td>
<td>12005</td>
<td>0.638</td>
</tr>
<tr>
<td></td>
<td>UDP</td>
<td>28</td>
<td>398</td>
<td>398</td>
<td>0.708</td>
</tr>
<tr>
<td><strong>Allosteric regulation</strong></td>
<td>Allosteric (Allo)</td>
<td>682</td>
<td>682</td>
<td>109</td>
<td>0.638</td>
</tr>
</tbody>
</table>

AUC measured via per chain 10-fold cross-validation
**Quick Digression**

- Unprecedented growth of human genetic variant data
  - e.g. HGMD, dbSNP

- In particular, amino acid substitutions (AAS)

- Focus on tools that predict effects of AAS (deleterious vs neutral)
  - e.g. MutPred, SIFT, PolyPhen, SNPs3D, SNAP
**Motivation: Molecular Mechanisms of Disease**

**Sickle Cell Disease**
- Autosomal recessive disorder
- E6V in HBB causes interaction w/ F85 and L88
- Formation of amyloid fibrils
- Abnormally shaped red blood cells, leads to sickle cell anemia
- Manifestation of disease vastly different over patients

[Link to HBS2](http://gingi.uchicago.edu/hbs2.html)
INFERRING MOLECULAR MECHANISMS OF DISEASE

• Most of these tools do not predict biochemical cause of disease
  – In particular, molecular function alterations
• Lack of comprehensive studies using protein 3D structure data

Goal: Exploit the structural environment of a residue of interest to hypothesize specific molecular effects of AAS and to statistically attribute these effects to genetic disease

Idea:

• Develop methods to predict specific function
  • e.g. zinc-binding site or phosphorylation site
• Apply to amino acid substitution data
• Provide probabilistic estimates of molecular mechanisms of disease
**APPRAOCH**

**Consider:**
- phosphorylation in structure $s$ occurs at position $i$
- residue $x$ is mutated to $y$, at position $j$ ($xjy$)

Phosphorylation site: variant position (C46W)

\[
i = 45 \quad j = 46
\]

\[
\text{s: } \text{...LAGDKMGMGQSCVGALFNDVQ...}
\]

**Loss of phosphorylation:**

\[
\Pr (\text{loss of } p \text{ at } s_i | xjy) = \Pr (s_i = s_i^p | s) \cdot (1 - \Pr (s_i = s_i^p | s_{xjy}))
\]

**Gain of phosphorylation:**

\[
\Pr (\text{gain of } p \text{ at } s_i | xjy) = (1 - \Pr (s_i = s_i^p | s)) \cdot \Pr (s_i = s_i^p | s_{xjy})
\]

---

Identifying Active Mechanisms of Disease

<table>
<thead>
<tr>
<th>Data set</th>
<th>Total # of AAS</th>
<th># of AAS mapped to PDB</th>
<th># of genes</th>
<th># of PDB entries</th>
<th># of chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>282,625</td>
<td>8,049</td>
<td>2,095</td>
<td>3,047</td>
<td>3,500</td>
</tr>
<tr>
<td>Disease</td>
<td>52,406</td>
<td>10,629</td>
<td>583</td>
<td>1,177</td>
<td>1,387</td>
</tr>
</tbody>
</table>
Loss and gain of functional sites is an active mechanism of disease.
**Validation of Loss of Function Predictions**

- Mutagenesis experimental data (UniProt)
  - 3,356 amino acid substitutions mapped to PDB (880 distinct proteins)

<table>
<thead>
<tr>
<th>Category</th>
<th>Site type</th>
<th>$n_i$</th>
<th>AUC</th>
<th>$n_i^*$</th>
<th>AUC*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metal binding</td>
<td>Ca</td>
<td>13</td>
<td>0.750</td>
<td>2</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>Fe</td>
<td>42</td>
<td>0.759</td>
<td>30</td>
<td>0.913</td>
</tr>
<tr>
<td></td>
<td>Mg</td>
<td>35</td>
<td>0.761</td>
<td>11</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>Zn</td>
<td>47</td>
<td>0.765</td>
<td>26</td>
<td>0.960</td>
</tr>
<tr>
<td>PTMs</td>
<td>Nglyco</td>
<td>11</td>
<td>0.825</td>
<td>2</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td>Phos</td>
<td>59</td>
<td>0.652</td>
<td>12</td>
<td>0.948</td>
</tr>
<tr>
<td>Catalytic sites</td>
<td>Cat</td>
<td>1226</td>
<td>0.666</td>
<td>608</td>
<td>0.926</td>
</tr>
<tr>
<td>Macro-molecule binding</td>
<td>DNA</td>
<td>37</td>
<td>0.682</td>
<td>20</td>
<td>0.895</td>
</tr>
<tr>
<td></td>
<td>RNA</td>
<td>52</td>
<td>0.614</td>
<td>15</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>PPI</td>
<td>616</td>
<td>0.575</td>
<td>87</td>
<td>0.934</td>
</tr>
<tr>
<td>Ligand binding</td>
<td>ADP</td>
<td>12</td>
<td>0.771</td>
<td>4</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>ATP</td>
<td>121</td>
<td>0.782</td>
<td>81</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>GDP</td>
<td>17</td>
<td>0.585</td>
<td>3</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>GTP</td>
<td>39</td>
<td>0.646</td>
<td>11</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>NAD</td>
<td>16</td>
<td>0.637</td>
<td>4</td>
<td>0.987</td>
</tr>
</tbody>
</table>

- Feasibility of computationally predicting loss of functional sites
Amyotrophic lateral sclerosis (ALS)

- D83G in superoxide dismutase (SOD1) causes:
  - Loss of zinc-binding that destabilizes native structure
  - Leads to protein aggregation that forms amyloid-like fibrils
Multiple graphlet kernels for vertex classification
  – Available at http://sourceforge.net/projects/graphletkernels/

Successfully applied to prediction of many types of functional residues

Useful for predicting impact of mutations and understanding molecular mechanism of disease

Implications on the ways we study disease

Implications on precision medicine
ACKNOWLEDGEMENTS

Indiana University
- Predrag Radivojac
- Vikas Pejaver
- Radivojac’s group
- Esfan Haghverdi

University of Washington
- Sean Mooney’s group

Cardiff University
- David Cooper’s group

Thank you!