Mining Networked Data

Michelangelo Ceci
Outline

• We will
  • Motivate the current shift of interest towards learning with networked data
  • Formalize the problem of learning from networked data
  • Clarify the role of various forms of within-network correlation
  • Establish links with related areas of research (transductive and semi-supervised learning, collective inference)

• We will
  • Point to important results (Theoretical / Experimental) of this paradigm shift in machine learning
  • Applications (Biological domain, energy prediction)
Mining network data: different dimensions

Type

- Inductive
- Transductive
- Semi-supervised
- Learning setting

Subject: Nodes/edges
- Ranking
- Link prediction
- Network reconstruction

Type
- Predictive
- Descriptive
- Clusters
- Patterns
- Sub-networks
- Cycles

Mining network data
- Time dimension
  - Data mining task
    - Type of network
      - Type of network
        - Information (attributes)
          - Nodes
            - Edges
          - Directed/undirected
        - Homogeneous
        - Heretogeneous
        - Complex structures
        - Anomalies
      - Type of network
        - Information (attributes)
          - Nodes
            - Edges
          - Directed/undirected
        - Homogeneous
        - Heretogeneous
        - Complex structures
        - Anomalies
  - Theoretical issues
    - @work
  - Applications
    - autocorrelation
    - i.i.d. assumption
    - Streaming
    - Evolving
    - Non evolving
Networked Data

• A collection of interconnected entities

• Entities can be
  • homogeneous/heterogeneous
  • Labelled/unlabelled
  • Described by a single / multiple attribute(s) / structured representations
  • Defined at various levels of abstractions

• Connections can be
  • Homogeneous / heterogeneous
  • Labelled / unlabelled
  • Binary / n-ary
  • Defined at various levels of abstraction
Ubiquity of Networked Data

- **Entities**: Web pages
- **Connections**: Hyperlinks

- **Entities**: Scientific research papers
- **Connections**: References
Ubiquity of Networked Data

• **Entities**: authors of papers on DBLP
• **Connections**: co-authorship

[Rattigan & Jenssen, 2005]
Ubiquity of Networked Data

• **Entities**: movies, studios and actors.
• **Connections**: movies are produced by some studio, actors are starring in a movie

[Neville & Jensen 2003]
Ubiquity of Networked Data

- **Entities**: blogs, web pages.
- **Connections**: a link from the blog to another blog or website. Links can be categorized according to the destination and where they appear within the blog pages:
  - a link appearing in a blog entry,
  - a link appearing in a comment posted as a response to a blog entry,
  - a link in the “friends” category.

[Bhagat, Rozenbaum, Cormode 2007]
Ubiquity of Networked Data

• **Entities**: hijackers

• **Connections**:
  • contacts,
  • meetings,
  • activities,
  • training,
  • financial transactions,
  • authority relationships,
  • …

The 9/11 Hijacker Extended Network [Krebs, 2002]
Ubiquity of Networked Data

• **Entities:** proteins
• **Connections:**
  • interactions

http://www.cdsb.dk/about.php
Ubiquity of Networked Data

- **Entities**: map cells
- **Connections**: adjacency
Ubiquity of Networked Data

• **Entities**: spatial objects (rivers, roads, …)

• **Connections**:
  - intersection,
  - inclusion,
  - adjacency,
  - overlapping,
  - distance,
  - …
Prediction Tasks in Networked Data

• Predicting the label of an entity
  • What’s the topic of a web page?
Structured Output Prediction Tasks in Networked Data

• Predicting the label of an entity
  • What’s the function of a gene? (Gene Ontology)

[Stojanova et al, 2013]

A. METABOLISM
A.1 amino acid metabolism
A.2 nitrogen, sulfur, selenium met.
A.1.3 assimilation of ammonia
A.1.3.1 metabolism of glutamine
A.1.3.1.1 biosynthesis of glutamine
A.1.3.1.2 degradation of glutamine
...
B. ENERGY
B.1 glycolysis and gluconeogenesis
C. CELL CYCLE and DNA PROCESSING
D. TRANSCRIPTION
D.1 RNA synthesis
D.2 RNA processing
D.3 transcriptional control
Prediction Tasks in Networked Data

• Predicting the existence of a new connection
  • possible co-authorships
Prediction Tasks in Networked Data

• Predicting the label of a connection
  • Is a reference to the paper correct?
  • Is the co-author relationship an advisor-advisee relationship?
Prediction Tasks in Networked Data

• Predicting when two entities are the same (entity resolution)
  • Are two citations referring to the same paper?
Ranking: recommend friends (social networks)

Goal: Given a user $s$, recommend friends
Positive: Nodes to which $s$ links to in the future
Negative: Nodes to which $s$ does not link

Supervised ranking problem: Assign higher scores to positive nodes than to negative nodes

Combine PageRank with Supervised learning to combine network structure and node and edge features

Idea: Use node and edge features to “guide” the random walk

[Leskovec, 2012]
Ranking: influencers in social networks

Given a graph $G$, a database of propagation $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$ (composed of DAGs)

Find:
Summaries $S$ (set of DAGs) that:
- Have homogeneity population
- Have a good hierarchical structure (good ranking)
Description Tasks in Networked Data

• Detect groups which explain
  • Why movies make more than $2 million in opening weekend box-office returns

[Neville & Jensen 2005]
Description Tasks in Networked Data

• Detect anomalous links

[Top 10 Most Unlikely Papers of 2003-2004]

1. Jean-Francois Boulicaut, Celine Robardet. Constraint-Based Mining Of Formal Concepts In Transactional Data. 2004
3. Jiawei Han, Michael Welge, David Clutter. MAIDS: Mining Alarming Incidents From Data Streams. 2004.
4. Jiawei Han, Michael Garland. Mining Scale-Free Networks Using Geodesic Clustering. 2004.
Description Tasks in Networked Data

• Detect anomalous links

Top 10 Most Likely Papers of 2003–2004

1. Jiawei Han, Jian Pei, Jiayong Wang. CLOSET+: Searching For The Best Strategies For Mining Frequent Closed Itemsets. 2003.
2. Takashi Washio, Hiroshi Motoda. State Of The Art Of Graph-Based Data Mining. 2003.
3. Jiawei Han, Joyce M W Lam, Guozhu Dong, Ke Wang, Jian Pei. Mining Constrained Gradients In Large Databases. 2004.
8. Jiawei Han, Ling Feng, Anthony K H Tung, Hongjun Lu. Efficient Mining Of Intertransaction Association Rules. 2003.
Description Tasks in Networked Data

- Cluster spatially distributed sensor networks with same trend of measurement

[Appice et al, 2015]
Relational mining for Discovering Changes in Evolving Networks

• Application to real-world heterogeneous networks (KEDS, DBLP,...)
• with a background hierarchy on the nodes: multi-level change chains

\[ P(c)^1 = network(N), is(X, africa), is(Y, america), \]
\[ (consult_{June.2008}(X, Y) \rightarrow express\_intent\_to\_cooperate_{December.2008}(X, Y)) \]
\[ ([June.2008, December.2008], \gamma = 0.287) \]

\[ P(c)^2 = network(N), is(X, africa), is(Y, america), \]
\[ (express\_intent\_to\_cooperate_{December.2008}(X, Y) \rightarrow make\_public\_statement_{June.2009}(X, Y)) \]
\[ ([December.2008, June.2009], \gamma = 0.287) \]
\[ (supp = 0.0519) \]
Within-Network Inference

• Training entities are connected directly to those entities whose labels are to be estimated
Across-Network Inference

- Learning from one network and applying the learned model to a separate, presumably similar, network.
Formalization: Graph-based

Given

• An undirected graph $G=(V,E)$,

• A vector of features $X_1, \ldots, X_n$ (explanatory variables) is defined on each node $v \in V$.

In predictive tasks, a response (target) variable $Y$ is defined for a subset $V^K \subseteq V$.

Main assumptions:

• homogeneity of entities and connections;

• symmetry of connections.
Formalization: Graph-based

• $N(v) \subseteq V$ neighborhood of a node $v \in V$
  • Set of nodes in $V$ which are “close” to $v$. Henceforth, assume: $v \in N(v)$ and $|N(v)| > 1$

• Closeness:
  • Directly linked to $v$
  • Within a certain distance from $v$
Correlations

Within network

1. Between the response variable $Y$ of a node $v \in V$ and the explanatory variables $X_i$ of nodes in $N(v)$.

Cross-correlation

Exploited in within network and across network inference

*The cost of an apartment in a location depends on the services both in that location and in its neighborhood.*
Correlations

Within network

2. Between the response variable $Y$ of a node $v \in V$ and the response variable of nodes in $N(v)$.

Auto-correlation

Exploited in within network inference, not in across network inference

*The price level for a good at a retail outlet in a city depends on the price for the same good in the nearby.*
Causes of Autocorrelation

• Tobler’s first law of geography [Legendre 1993]
  Everything is related to everything else, but nearby things are more related than distant things.

• Homophily [McPherson et al. 2001]
  The tendency of nodes with similar values to be linked with each other.
Relational Correlation

- **A**: a set of objects described by attribute \( f \)
- **B**: a set of objects described by attribute \( g \)
- **\( P_R \)**: a set of pairs of A and B related by paths in the network

**Relational correlation**: correlation between all pairs \((f(a),g(b))\), where \(a \in A\), \(b \in B\), \(p(a,b) \in P_R\)

[Jensen & Neville, 2006]
Relational Correlation

Independent instances

Dependent instances

(Class Movie Attr. value Studio Other object)

(Jensen & Neville, 2002)
Estimating Relational Correlation

- Given a set of related pairs $P^R$, the correlation between two discrete variables can be computed as Pearson’s Contingency* Coefficient:

$$r_{ac} = \sqrt{\frac{\chi^2}{(N + \chi^2)}}$$

where $\chi^2$ is the Chi-square statistics and $N$ is the sample size.

[Neville, 2006]

* defined for contingency tables
Estimating Correlation

\[ r_{ac} = \sqrt{\frac{\chi^2}{(N + \chi^2)}} \]

• \( r_{ac} \in [0,1] \)

• 1 the value of the attribute \( f \) for a node \( v_i \) is always equal to all other nodes \( v_j \) reachable by a path in the network, with a given value of attribute \( g \).

• 0 the values of \( f \) and \( g \) in related pairs of nodes are independent.
Estimating Autocorrelation

• \( X \): attribute
• \( A, B \): two sets of objects described by \( X \)
• \( P_R \): a set of pairs of \( A \) and \( B \) related by paths in the network

Relational autocorrelation: correlation between all pairs \((X(a), X(b))\), where \( a \in A, b \in B, p(a,b) \in P_R \)

(Jensen & Neville, 2006)
Relational AutoCorrelation

Autocorrelation on the Movie database

\[ r_{ac} = \sqrt{\frac{\chi^2}{N + \chi^2}} \]

<table>
<thead>
<tr>
<th>Autocorrelation Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C'(Movie,Receipts,Made</td>
<td>Studio</td>
</tr>
<tr>
<td>C'(Movie,Receipts,Directed</td>
<td>Director</td>
</tr>
<tr>
<td>C'(Movie,Receipts,Produced</td>
<td>Producer</td>
</tr>
<tr>
<td>C'(Movie,Receipts,ActedIn</td>
<td>Actor</td>
</tr>
</tbody>
</table>

*Note:* The notation \( a|x|b \) to denote paths with links of type \( a \) and \( b \) and intervening objects of type \( x \).

(Jensen & Neville, 2002)
Relational AutoCorrelation
Data Relationship $\neq$ Statistical Independence

Low autocorrelation between *Movies* through *Actor*, but the two pieces of data are anyway linked.

In general, the network of *statistical dependencies* does not necessarily correspond to the *data network*.
Estimating Autocorrelation (continuous attrs.)

• Given a set of related pairs $P_R$, we can measure the autocorrelation of a continuous variable $X$ as the correlation between:

  • $X_i$: value of $X$ for $v_i$
  • $X_j$: value of $X$ for $v_j$

$$r_{ac} = \frac{\sum_{i,j \text{ s.t.}(v_i,v_j) \in P_R} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i \text{ s.t.}(v_i) \in P_R} (X_i - \bar{X})^2}$$

[Neville, 2006] [Stojanova et al., 2012]
Learning with Networked Data

• The main issue in learning with networked data is the presence of within-network correlation.

• Standard learning approaches:
  • Appropriate when there is no correlation within network (i.i.d. assumption holds)
  • Efficient
  • Theoretically well-founded
  • Well-investigated properties

• A different class of methods is required to account for within-network correlation.

Network-based learning / mining
Learning Network-based Predictive Clustering Trees

Predictive models

Measures of Network Autocorrelation

- Global Moran’s I (spatial AC)

\[ I_X = \frac{N \sum_i \sum_j w_{ij} (Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_i \sum_j w_{ij} \sum_i (Y_i - \overline{Y})^2} \]

where \( W = [w_{ij}] \) is the weighting matrix

- Randić Connectivity Index (CI)

\[ \chi = \sum_{edges \ ij} \frac{1}{\sqrt{D(i)D(j)}} \]

where \( D(i) \) and \( D(j) \) represent the weighted node degree vector

- Relational AC Coefficient (P)

\[ P_Y = \frac{\sum_{ij \ s.t. \ (u_i,u_j) \in P_R} (Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_{ij \ s.t. \ (u_i,u_j) \in P_R} (Y_i - \overline{Y})^2} \]

where \( P_r \) represents the set of related pairs
Moran’s I

- Moran’s I ranges between –1.0 and + 1.0
  - When autocorrelation is high, the coefficient is high
  - A high positive value indicates positive autocorrelation

Moran’s I: 0.66

Moran’s I: 0.12
Top down induction of NPCTs

Algorithm 1. Top-down induction of NetworkPCTs

1: procedure NetworkPCT(A) returns tree
2: if stop(A) then
3: return leaf(Prototype(A))
4: else
5: \((c^*, h^*, \mathcal{P}^*) = (null, 0, \emptyset)\)
6: for each possible test \(c\) do
7: \(\mathcal{P} = \text{partition induced by \(c\) on \(A\)}\)
8: \(h = \frac{\alpha}{|Y|} \sum_{Y} \Delta_Y(A, \mathcal{P}) + \frac{1-\alpha}{|Y|} \sum_{Y} S_Y(A, \mathcal{P})\)
9: if \((h > h^*)\) then
10: \((c^*, h^*, \mathcal{P}^*) = (c, h, \mathcal{P})\)
11: end if
12: end for
13: for each \(A_k \in \mathcal{P}^*\) do
14: \(\text{tree}_k = \text{NetworkPCT}(A_k)\)
15: end for
16: return node\((c^*, \bigcup_k \{\text{tree}_k\})\)
17: end if

\(\Delta_Y\) variance reduction
(N)PCTs

\(S_Y\) network autocorrelation
NPCTs
<table>
<thead>
<tr>
<th>Method/ Network dataset</th>
<th>VideoL</th>
<th>MOVIES1</th>
<th>MOVIES2</th>
<th>BOOKS</th>
<th>TWITTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCLUS_I</td>
<td>α = 0.0</td>
<td>686.32</td>
<td>1.08</td>
<td>1.17</td>
<td>2.53</td>
</tr>
<tr>
<td>NCLUS_I</td>
<td>α = 0.5</td>
<td>653.2</td>
<td>1.06</td>
<td>1.02</td>
<td>2.53</td>
</tr>
<tr>
<td>NCLUS_RA</td>
<td>α = 0.0</td>
<td>686.32</td>
<td>1.08</td>
<td>1.17</td>
<td>2.53</td>
</tr>
<tr>
<td>NCLUS_RA</td>
<td>α = 0.5</td>
<td>653.2</td>
<td>1.06</td>
<td>1.02</td>
<td>2.53</td>
</tr>
<tr>
<td>NCLUS_CI</td>
<td>α = 0.0</td>
<td>686.32</td>
<td>1.62</td>
<td>2.11</td>
<td>2.53</td>
</tr>
<tr>
<td>NCLUS_CI</td>
<td>α = 0.5</td>
<td>653.2</td>
<td>1.51</td>
<td>1.31</td>
<td>2.53</td>
</tr>
<tr>
<td>CLUS</td>
<td></td>
<td>660.69</td>
<td>1.53</td>
<td>2.42</td>
<td>2.83</td>
</tr>
<tr>
<td>SVR</td>
<td></td>
<td>721.43</td>
<td>1.77</td>
<td>2.52</td>
<td>2.65</td>
</tr>
<tr>
<td>k-NN</td>
<td></td>
<td>937.68</td>
<td>1.70</td>
<td>2.58</td>
<td>2.65</td>
</tr>
<tr>
<td>M5’</td>
<td></td>
<td>574.17</td>
<td>2.09</td>
<td>2.2</td>
<td>2.67</td>
</tr>
<tr>
<td>Base</td>
<td></td>
<td>722.39</td>
<td>2.10</td>
<td>2.70</td>
<td>2.53</td>
</tr>
</tbody>
</table>

The RMSEs (estimated by 10-fold CV) of the models obtained with NCLUS_I, NCLUS_RA, NCLUS_CI, CLUS, SVR, k-NN and M5’, as well as the default Base model. For each network dataset, the best results are highlighted in bold.
Table 3 Comparison of error made by different learning approaches on spatial network data

<table>
<thead>
<tr>
<th>Method /Spatial Dataset</th>
<th>FF</th>
<th>NWE</th>
<th>FOIXA</th>
<th>GASD</th>
<th>MS</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCLUS_I $\alpha = 0.0$</td>
<td>42.82</td>
<td>2.16</td>
<td>2.53</td>
<td>0.18</td>
<td>2.47</td>
<td>5.44</td>
</tr>
<tr>
<td>NCLUS_I $\alpha = 0.5$</td>
<td>56.55</td>
<td>2.48</td>
<td>2.34</td>
<td>0.17</td>
<td>2.29</td>
<td>5.81</td>
</tr>
<tr>
<td>NCLUS_RA $\alpha = 0.0$</td>
<td>42.82</td>
<td>2.45</td>
<td>2.65</td>
<td>0.18</td>
<td>2.47</td>
<td>6.60</td>
</tr>
<tr>
<td>NCLUS_RA $\alpha = 0.5$</td>
<td>53.27</td>
<td>2.46</td>
<td>2.66</td>
<td>0.17</td>
<td>2.35</td>
<td>5.92</td>
</tr>
<tr>
<td>NCLUS_CI $\alpha = 0.0$</td>
<td>42.82</td>
<td>2.47</td>
<td>2.61</td>
<td>0.17</td>
<td>2.49</td>
<td>6.72</td>
</tr>
<tr>
<td>NCLUS_CI $\alpha = 0.5$</td>
<td>52.79</td>
<td>2.45</td>
<td>2.66</td>
<td>0.16</td>
<td>2.35</td>
<td>5.93</td>
</tr>
<tr>
<td>CLUS</td>
<td>49.21</td>
<td>2.46</td>
<td>2.65</td>
<td>0.16</td>
<td>2.35</td>
<td>5.64</td>
</tr>
<tr>
<td>CLUS*</td>
<td>47.22</td>
<td>2.47</td>
<td>2.52</td>
<td>0.16</td>
<td>2.54</td>
<td>6.68</td>
</tr>
<tr>
<td>ITL</td>
<td>58.25</td>
<td>2.54</td>
<td>3.55</td>
<td>0.14</td>
<td>1.92</td>
<td>3.52</td>
</tr>
<tr>
<td>SVR</td>
<td>64.58</td>
<td>2.50</td>
<td>2.95</td>
<td>0.14</td>
<td>2.80</td>
<td>6.60</td>
</tr>
<tr>
<td>kNN</td>
<td>65.44</td>
<td>2.40</td>
<td>2.73</td>
<td>0.16</td>
<td>2.37</td>
<td>4.56</td>
</tr>
<tr>
<td>M5'</td>
<td>47.22</td>
<td>2.47</td>
<td>2.66</td>
<td>0.16</td>
<td>2.47</td>
<td>5.92</td>
</tr>
<tr>
<td>Base</td>
<td>63.66</td>
<td>2.50</td>
<td>2.93</td>
<td>0.20</td>
<td>3.23</td>
<td>8.58</td>
</tr>
</tbody>
</table>

The RMSEs (estimated by 10-fold CV) of the models obtained with NCLUS_I, NCLUS_RA, NCLUS_CI, CLUS, ITL, SVR, k-NN and M5', as well as the default Base model. For each network dataset, the best results are highlighted in bold. In the case ITL outperforms other approaches, we report values in italic. This because comparison to ITL is not fair. Results for NWE are multiplied by $10^3$. 

| 47 |
Learning with Networked Data

Network linkage poses two additional issues:

**TRAINING**: The distinction between training and testing is not always possible.
PREDICTIONS: Predictions cannot be performed independently
Basic approaches for prediction

**Problem:** *simultaneous inference* of $Y$ for remaining nodes in $V^U = V - V^K$.

*Naïve approach:* estimate the full joint probability distribution

$$P(Y^U \mid G, Y^K)$$

*Approximation:* Markov assumption

$$P(Y_j \mid G, Y^K) = P(Y_j \mid N(v_i))$$
Basic approaches for prediction

Weighted-vote Relational Neighbor Classifier (wvRC)

\[ P(Y_j = c \mid N(v_j)) = \frac{1}{Z} \sum_{v_k \in N(v_j)} w_{j,k} P(Y_k = c \mid N(v_k)) \]

Z: a normalizing constant.

Class-membership probability is estimated by assuming positive autocorrelation.

Other classifiers have been proposed [Macskassi & Provost, 2007]
Collective Inference

\[ N(v_j) \cap V^U \text{ may not be empty.} \rightarrow \]

**Collective Inference:** Simultaneous judgments regarding the values of response variables for multiple linked entities for which some attribute values are not known.

- Gibbs sampling
- Relaxation labeling
- Iterative classification

[Macskassy & Provost, 2007]
Transductive Inference

• An inference mechanism “from particular to particular”.
• Uses both labeled & unlabeled data to build classifier, whose goal is to classify only unlabeled data as accurately as possible.
• No general rule valid for all possible instances is generated.
Semi-Supervised Smoothness Assumption

• If two points $x_1$ and $x_2$ in a high-density region are close, then so should be the corresponding outputs $y_1$, $y_2$.
• The label function is smoother in high-density regions than in low-density regions.
• This assumption entails that if two points are separated by a low-density region, then their outputs need not be close.
• It is also called label smoothness assumption.
Transductive Inference

Labeled data only

SVM

Transductive SVM
Analogy

- In within-network classification both labeled and unlabeled nodes are used and only unlabelled nodes have to be classified.
- Homophily justifies transductive learning.

Learning with Networked Data + Transductive Learning?

[Macskassy, 2007]
Multi-type Classification from Heterogeneous Networks

• Exploiting clustering of heterogeneous networks for classification purposes
  • Extraction of multi-type clusters
  • Hierarchical organization of clusters

• Simultaneous classification of objects of different types
  • Catching possible dependencies among labels
  • Exploiting the hierarchical structure to take into account dependencies at different levels of granularity

G. Pio, F. Serafino, D. Malerba, M. Ceci “Multi-Type Classification from Heterogeneous Networks”, Data Mining and Knowledge Discovery (submitted).
Questions?
Thank you

Questions?

References


