Efficient Partial Order Preserving Unsupervised Feature Selection on Networks

Xiaokai Wei, Sihong Xie, Philip S. Yu

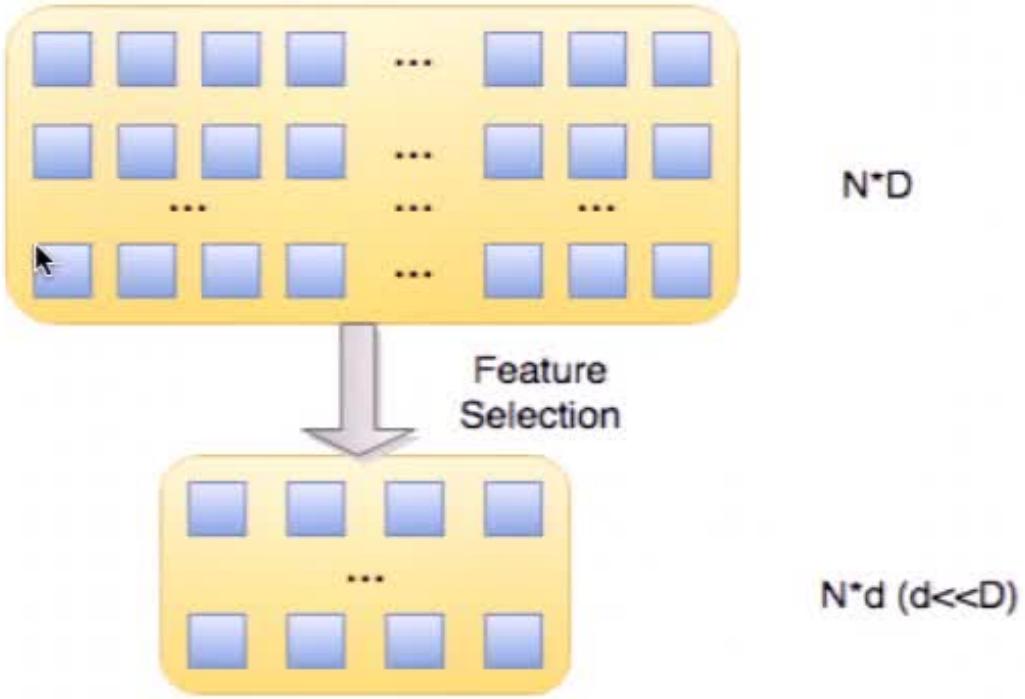
Department of Computer Science,
University of Illinois at Chicago

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

SDM 2015

Feature Selection

- Effective Way for Dimension Reduction
  - Faster learning time
  - Better prediction accuracy
  - Better Interpretability



Supervised v.s Unsupervised

- Supervised Feature Selection
  - Relatively easy
  - Class label can provide clear guidance
- Unsupervised Feature Selection
  - More difficult due to lack of labels
  - Various criteria have been proposed (e.g., max variance, Laplacian Score [1])
- Pseudo-label Based Approaches
  - Generate cluster labels from attributes and perform sparse regression

Feature Selection on Networks

- More and More Network Data

- Traditional Feature Selection
  - i.i.d assumption

- Feature Selection on Networks
  - Data are not i.i.d (homophily effect)

Feature Selection on Networks

- New Challenges and Opportunities
- How to effectively exploit network links?
  - Can help select better features?
- Efficiency
  - Real-world social/information networks can be huge
  - DBLP has more than 2 million CS articles
  - Facebook: > 1 billion users
  - Linkedin: > 300 million users

Limitations of Existing Work

- Most existing work cannot exploit link structure
  - LUFS[3] is the only work that attempts to use links for unsupervised feature selection
- Pseudo-labels can be unreliable
- Not Efficient
  - Rely on intensive matrix computation to converge to local optima

Partial Order Preserving (POP)

- A New Way to Exploit Homophily Effect
  - Effective and efficient

Figure 1: An example network with 9 nodes

- Key Intuition
  - Neighbors are more likely to be from the same class than two random non-neighbors
  - E.g., friends in social network, cited paper/citing paper, co-authors
  - Selected features should make neighbors similar and non-neighbors not so similar

Partial Order Preserving (POP)

- Formulate the Intuition into Partial Order
  - Features that preserve such partial orders are likely to be high-quality ones

DEFINITION 4. **Link-based Partial Order** *We formulate such property as partial order $j >_i k$, where node v_j and node v_k are in the linked set and unlinked set of node v_i, respectively. Node v_i is referred to as the pivot of this partial order. Such partial order is denoted as a triplet (i, j, k) or $j >_i k$.*

$$sim(v_i, v_j) > sim(v_i, v_k), v_j \in \mathcal{L}(v_i), v_k \in \mathcal{U}(v_i) \quad (3.1)$$

Partial Order Preserving (POP)

- Example: Paper Citation Network
  - Papers in Machine Learning, Database, OS...
  - Features are the terms in the paper
- Indiscriminative Terms
  - E.g., *propose, compare,* which does not help preserve partial order
- Discriminative Terms
  - E.g., *SVM, classification, database*
  - Neighbors are more likely to share these terms than non-neighbors.

Partial Order Preserving (POP)

- Formulations
  - Selection indicator: $\mathbf{w} = (w_1, w_2, \ldots, w_D)^T$
  - Similarity with selected features

$$s_{ij} = sim(diag(\mathbf{w})\mathbf{x}_i, diag(\mathbf{w})\mathbf{x}_j)$$

$$s_{ijk} = s_{ij} - s_{ik}$$

$$= \mathbf{x}_i^T diag(\mathbf{w})\mathbf{x}_j - \mathbf{x}_i^T diag(\mathbf{w})\mathbf{x}_k$$

  - Large s_{ijk} indicates the partial order is well preserved

Partial Order Preserving (POP)

- Link function
  - Transform s_{ijk} to loss
  - Should be monotonically non-decreasing

$$l(j >_i k \mid \mathbf{w}) = f(s_{ijk} \mid \mathbf{w})$$

- Objective function:

$$\max_{\mathbf{w}} L(>) = \sum_{(i,j,k)\in\Omega} l(j >_i k \mid \mathbf{w})$$

$$= \sum_{i\in V} \sum_{j\in\mathcal{L}_i} \sum_{k\in\mathcal{U}_i} f(s_{ijk} \mid \mathbf{w})$$

$$\text{s.t. } w_i \in \{0,1\}, \sum_{i=1}^{D} w_i = d$$

Simple POP

- Most simple instantiation
  - Identity function as link function
- Linked Score & Unlinked Score

$$score(a) = \sum_{(i,j,k)\in\Omega} I(i,j,a) - \sum_{(i,j,k)\in\Omega} I(i,k,a)$$

Linked Score

Unlinked Score

- Selection Criterion
  - The difference between linked score and unlinked score

Simple POP

- Example: Paper Citation Network
  - Papers in Machine Learning, Database, OS...
  - Features are the terms in the paper
- Indiscriminative Terms
  - E.g., *propose, compare*
  - High linked score and high unlinked score
- Discriminative Terms
  - E.g., *SVM, classification, database*
  - High linked score and low unlinked score

Probabilistic POP

- Assumption:
  - Partial orders are generated by s_{ijk} in a probabilistic way

- Approach:
  - Probability partial order $j >_i k$ is preserved

$$P(j >_i k \mid \mathbf{w}) = \sigma(s_{ijk})$$

$$\sigma(x) = 1/(1 + e^{-x})$$

Probabilistic POP

$$P(j >_i k \mid \mathbf{w}) = \sigma(s_{ijk})$$

$$\sigma(x) = 1/(1 + e^{-x})$$

- Maximize the Log-likelihood:

$$\max_{\mathbf{w}} \quad \log P(> |\mathbf{w}) = \sum_{(i,j,k)\in\Omega} \log P(j >_i k|\mathbf{w})$$

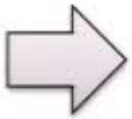
$$= \sum_{(i,j,k)\in\Omega} \log \sigma(s_{ijk})$$

$$\text{s.t. } w_i \in \{0,1\}, \ \sum_{i=1}^{D} w_i = d$$

Max Margin POP

- Max Margin POP
  - Inspired by Structural SVM
  - Aim to make neighbors/non-neighbors well separated

$$\min_{\mathbf{w}} \sum_{(i,j,k)\in\Omega} \mu_{ijk}$$
$$\text{s.t. } s_{ijk} \geq 1 - \mu_{ijk}, \forall (i,j,k) \in \Omega$$
$$w_i \in \{0,1\}, \sum_{i=1}^{D} w_i = d$$



$$\max_{\mathbf{w}} \sum_{(i,j,k)\in\Omega} -\max(0, 1 - s_{ijk})$$
$$\text{s.t. } w_i \in \{0,1\}, \sum_{i=1}^{D} w_i = d$$

Summary

Instantiation	SPOP	PPOP	MMPOP
Link function	Identity	Log of Sigmoid	Negative hinge
Evaluate features jointly	No	Yes	Yes

Connection to AUC

- AUC (Area Under ROC Curve)
  - Metric for evaluating binary prediction such as recommender system and link prediction

$$AUC(v_i) = \frac{1}{|\mathcal{L}_i||\mathcal{U}_i|} \sum_{j \in \mathcal{L}_i} \sum_{k \in \mathcal{U}_i} I(s_{ijk} > 0)$$

- Partial Order Preserving Principle
  - PPOP and MMPOP are continuous approximation of AUC (with logistic loss and hinge loss, respectively)

Optimization

- '0/1' Integer Programming
  - Relax '0/1' constraints on w_i
- Large Number of Partial Order Triplets:
  - O(|V||E|)
  - Stochastic (sub)gradient descent
  - Sample a small portion of triplets
- (Sub)Gradient Update
  - Each iteration takes O(m) (m: avg. number of non-zero features)

Stochastic (Sub)Gradient Descent

- Simple POP:
  - Gradient $$\frac{\partial l(j >_i k)}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} s_{ijk}$$
- Probabilistic POP:
  - Gradient $$\frac{\partial l(j >_i k)}{\partial \mathbf{w}} = \frac{e^{-s_{ijk}}}{1 + e^{-s_{ijk}}} \cdot \frac{\partial}{\partial \mathbf{w}} s_{ijk}$$
- Max Margin POP:
  - Subgradient $$\frac{\partial l(j >_i k)}{\partial \mathbf{w}} = \begin{cases} \frac{\partial}{\partial \mathbf{w}} s_{ijk} & \text{if } s_{ijk} < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial w_p} s_{ijk} = \begin{cases} 1 & \text{if } x_{ip} = 1 \;\&\; x_{jp} = 1 \;\&\; x_{kp} = 0 \\ -1 & \text{if } x_{ip} = 1 \;\&\; x_{jp} = 0 \;\&\; x_{kp} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Evaluation

- Datasets
  - Citeseer (citation network)
  - Cora (citation network)
  - Wikipedia (wiki articles)

Table 2: Statistics of three datasets

Statistics	Citeseer	Cora	Wiki
# of instances	3312	2708	3363
# of links	4598	5429	33219
# of features	3703	1433	4973
avg. # of non-zero features per instance	31.75	18.17	630.57
# of classes	6	7	19

Evaluation

- Baselines
  - All Features
  - Link Only
  - Laplacian Score [1]
  - UDFS [2] (Unsupervised Discriminative Feature Selection)
  - LUFS [3] (Linked Unsupervised Feature Selection)

Efficiency

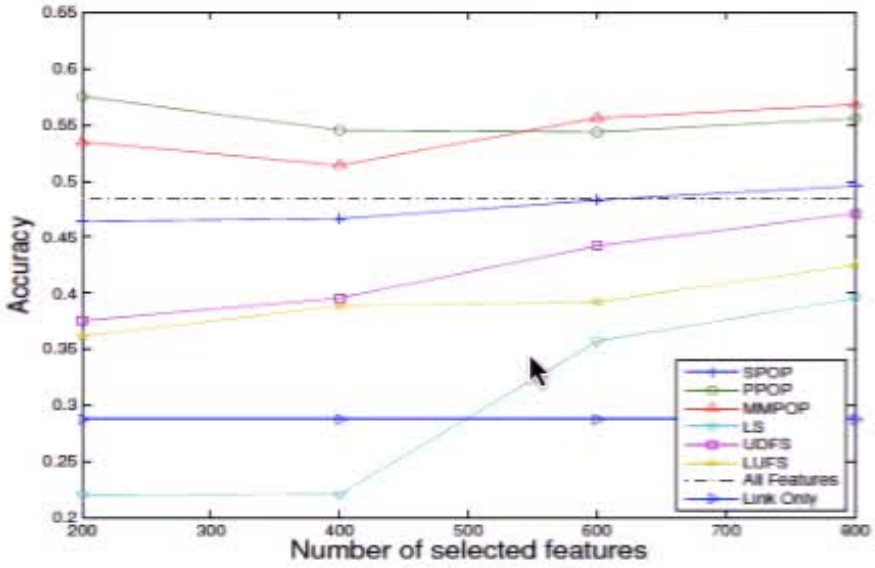
- POP
  - Much faster and able to perform online update by SGD
- UDFS/LUFS
  - Rely on intensive matrix factorization to converge to a local optima

Table 3: Running time (seconds) of different feature selection algorithms

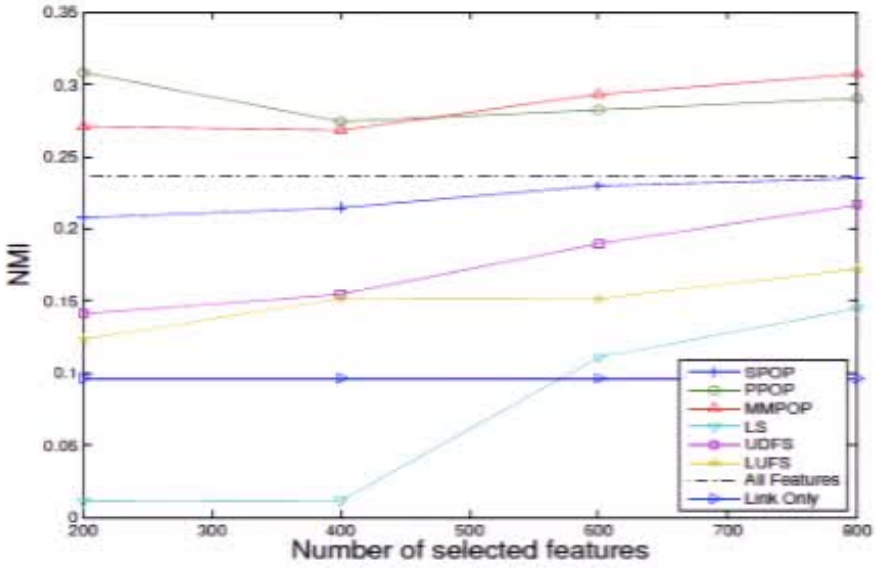
Dataset	LS	UDFS	LUFS	SPOP	PPOP	MMPOP
Citeseer	10	1234	1420	1	2	2
Cora	5	161	113	1	1	1
Wiki	23	2536	2788	19	22	19

Evaluation

- Clustering Performance
  - KMeans on selected features
  - Accuracy and NMI reported

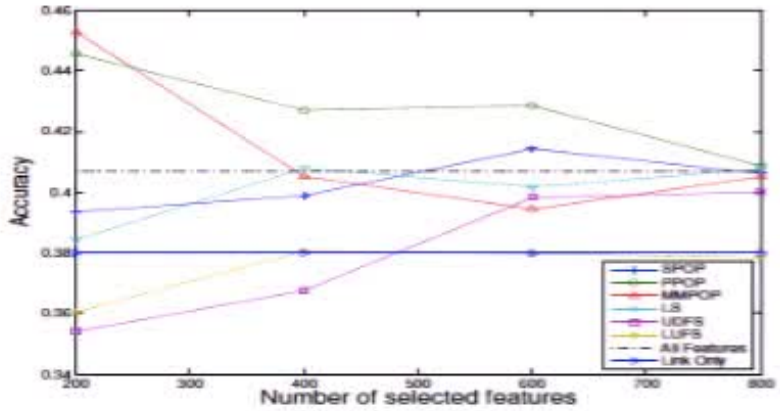


(a) Accuracy on Citeseer

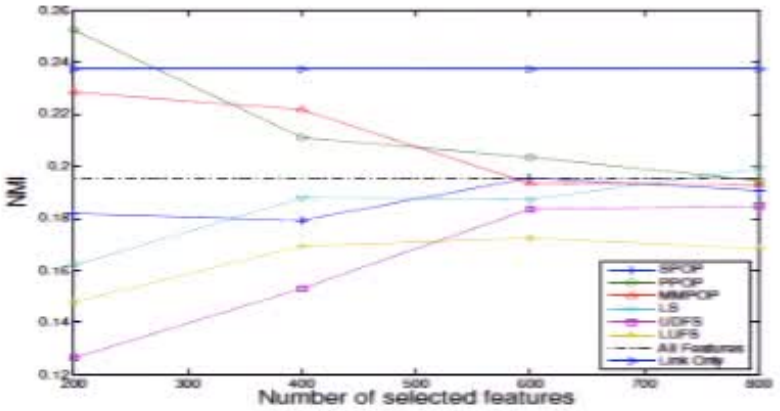


(b) NMI on Citeseer

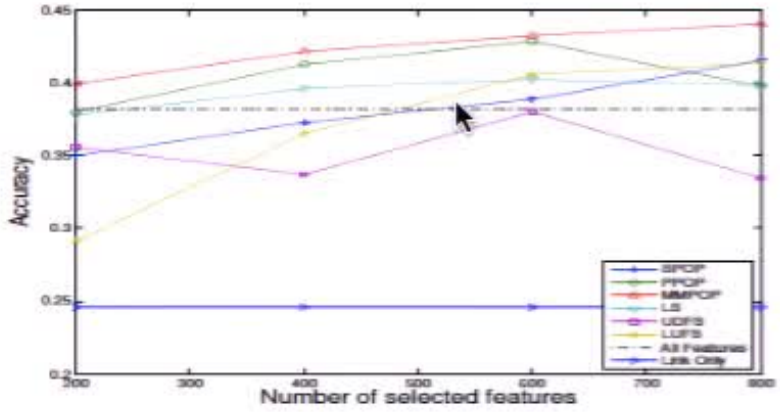
Evaluation



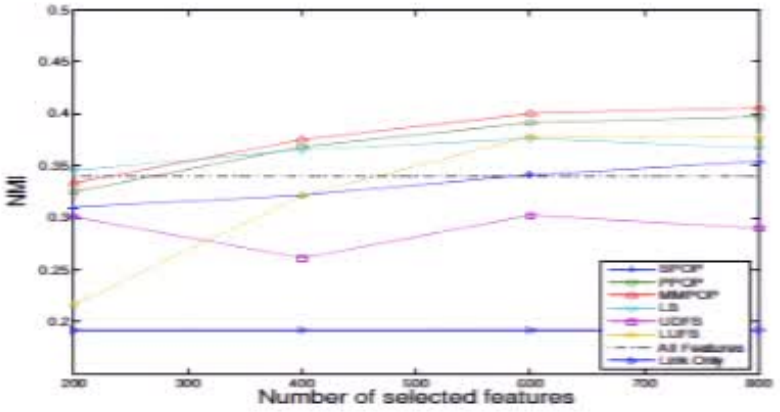
(c) Accuracy on Cora



(d) NMI on Cora



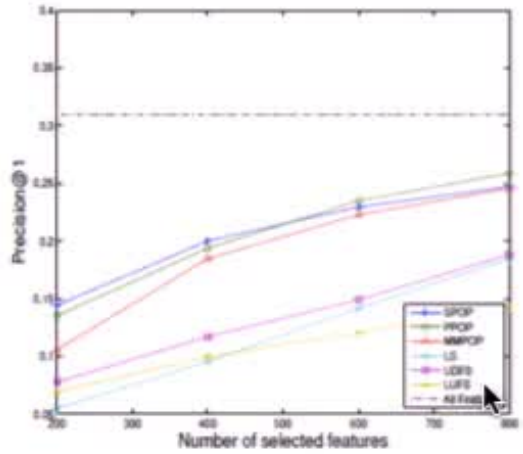
(e) Accuracy on Wiki



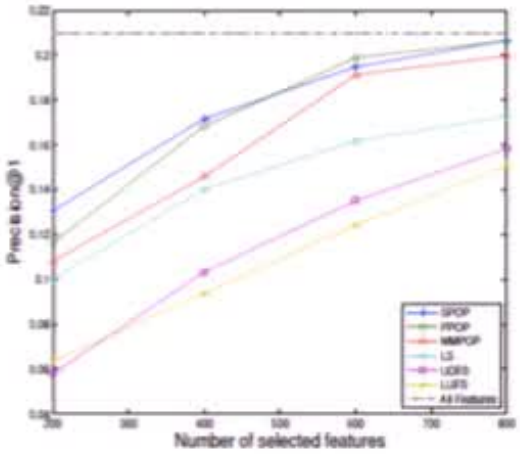
(f) NMI on Wiki

Evaluation

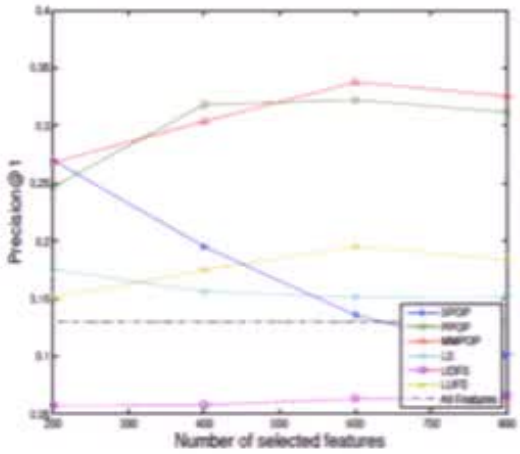
- Partial Order Preserving Property
  - Potential for link prediction



(a) Precision@1 on Citeseer



(b) Precision@1 on Cora



(c) Precision@1 on Wiki

Figure 3: 1NN Results on Three Datasets

Conclusion

- Conclusion
  - New criterion for unsupervised feature selection
  - More discriminative features
  - Much less running time
  - Experimental results verified the effectiveness and efficiency of the proposed approach