Adaptive Unsupervised Feature Selection on Attributed Networks

Jundong Li[†], Ruocheng Guo[†], Chenghao Liu[‡], Huan Liu[†]
[†]Computer Science and Engineering, Arizona State University
[‡]School of Information Systems, Singapore Management University
{jundongl,rguo12,huanliu}@asu.edu,chliu@smu.edu.sg

ABSTRACT

Attributed networks are pervasive in numerous of high-impact domains. As opposed to conventional plain networks where only pairwise node dependencies are observed, both the network topology and node attribute information are readily available on attributed networks. More often than not, the nodal attributes are depicted in a high-dimensional feature space and are therefore notoriously difficult to tackle due to the curse of dimensionality. Additionally, features that are irrelevant to the network structure could hinder the discovery of actionable patterns from attributed networks. Hence, it is important to leverage feature selection to find a high-quality feature subset that is tightly correlated to the network structure. Few of the existing efforts either model the network structure at a macro-level by community analysis or directly make use of the binary relations. Consequently, they fail to exploit the finer-grained tie strength information for feature selection and may lead to suboptimal results. Motivated by the sociology findings, in this work, we investigate how to harness the tie strength information embedded on the network structure to facilitate the selection of relevant nodal attributes. Methodologically, we propose a principled unsupervised feature selection framework ADAPT to find informative features that can be used to regenerate the observed links and further characterize the adaptive neighborhood structure of the network. Meanwhile, an effective optimization algorithm for the proposed ADAPT framework is also presented. Extensive experimental studies on various real-world attributed networks validate the superiority of the proposed ADAPT framework.

KEYWORDS

Unsupervised Feature Selection; Attributed Networks; Adaptive Neighborhood Structure; Tie Strength

ACM Reference Format:
Jundong Li, Ruocheng Guo, Chenghao Liu, Huan Liu. 2019. Adaptive Unsupervised Feature Selection on Attributed Networks. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330856

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330856

1 INTRODUCTION

It is a booming trend to model myriad of real-world information networks (e.g., social media networks, academic collaboration networks, and cellular networks) as *attributed networks*. The reason is ascribed to the fact that such networks often have rich accompanying features delineating the properties or characteristics of nodes, in addition to the raw network topology. For example, in parallel with the widely observed friendship relations, user profile information such as interests and demographic information are also readily available in social media platforms. Recent studies [18, 29] verified the existence of statistical dependencies (a.k.a. *autocorrelations*) between the attributes of linked nodes, and the presence of *autocorrelations* offers a unique opportunity to improve the learning performance of various graph mining tasks, ranging from node classification, community detection to link prediction.

Most, if not all, of the current studies on attributed networks, make use of all the given attributes [3, 15, 31]. However, the illusion that all attributes are complementary for the network structure information will break when the node attributes are in a high-dimensional space, as a number of noisy, irrelevant, and redundant attributes might be included. For example, social media platforms are swarming with low-quality user-generated content, and apparently, not all attributes could be leveraged for learning. Furthermore, such high-dimensional features pose additional challenges to the underlying learning algorithms due to the curse of dimensionality [7, 13]. Feature selection [21] has long been regarded as a very powerful tool to alleviate the curse of dimensionality. Given a specific target, it attempts to select a subset of high-quality features from the original high-dimensional feature space before performing further analytical or predictive tasks. In a supervised setting, the target can be specified as finding relevant features to discriminate instances from different classes [21]. As in the case of attributed networks, label information of nodes is both time and labor consuming to collect. In this regard, it is more desired to study *unsupervised feature selection on attributed networks*.

Without the guidance of label information, the task is shifted to find relevant features that are tightly correlated to the network topology. For example, one family of existing methods [24, 34] model the network structure at a macro-level through community analysis and then regard the node-community assignment as constraints to enable the probe of relevant features. However, it is often argued that these algorithms fail to exploit the fine-grained link information, and the involved matrix operations hinder their practical usage on large-scale networks. The other family of methods [37, 38] model the network information at a micro-level by preserving the partial-order node relations. A fundamental assumption behind

these algorithms is that connected nodes are more similar than non-connected nodes and an informative feature should be able to differentiate these two groups. These attempts, however, overwhelmingly focus on binary relations (e.g., coauthors or not) among nodes which only yield a coarse indication of the heterogeneity of relations. As indicated by tie strength theories [10, 11, 39], the strength of links could vary remarkably over the full spectrum (e.g., from close friends to acquaintances), thus treating all links equally may result in the selection of a suboptimal feature set. For example, in academic collaboration networks, informative features such as research interests should not only be able to make a distinction between coauthors and non-coauthors but also should distinguish collaborators with strong ties from the ones with weak ties.

Despite the fundamental importance of quantifying the heterogeneity of tie strength, the development of a sophisticated learning framework which can adaptively infer the tie strength for unsupervised feature selection is still in its infancy. On account of this, we explore the link information at a finer granularity and investigate how to exploit the adaptive neighborhood structure around each node[1] to obtain more relevant features on attributed networks. However, it remains a daunting task, mainly because of the following challenges: (1) According to the *dyadic hypothesis* in sociology [11], the strength of a tie is largely determined by the similarity of the two end nodes, which can be estimated from the node attribute information. Hence, the natural question is how to leverage node attribute information to find the adaptive neighborhood structure around each node? (2) As mentioned previously, informative features should differentiate node pairs with different levels of tie strength over the whole spectrum, then the second question is how to seamlessly incorporate the adaptive neighborhood characterization into feature selection? To tackle these challenges, we propose a novel adaptive unsupervised feature selection framework ADAPT. The main contributions of this paper are listed as follows:

- We systematically examine the fundamental significance of performing unsupervised feature selection on attributed networks and analyze the limitations of existing efforts.
- We propose a principled way to characterize the optimal adaptive neighborhood structure around each node for unsupervised feature selection on attributed networks.
- We present an effective optimization algorithm for the proposed unsupervised feature selection framework ADAPT.
- We empirically show the effectiveness of the proposed framework by carrying out extensive empirical studies on various real-world attributed networks.

2 PROBLEM DEFINITION

We use bold uppercase characters for matrices (e.g., $\mathbf{A}$), bold lowercase characters for vectors (e.g., $\mathbf{a}$), and normal lowercase characters for scalars (e.g., a). Also, we represent the i-th element of vector $\mathbf{a}$ as a_i, the i-the row of matrix $\mathbf{A}$ as $\mathbf{A}_{i*}$, the j-the column of matrix $\mathbf{A}$ as $\mathbf{A}_{*j}$, the (i,j)-th entry of matrix $\mathbf{A}$ as $\mathbf{A}_{ij}$, the transpose of $\mathbf{A}$ as $\mathbf{A}^T$. We use diag($\mathbf{a}$) to denote the diagonalization of vector $\mathbf{a}$. Meanwhile, $\mathbf{1}$ denotes a vector whose elements are all 1. The

[1]In this work, adaptive neighborhood structure around a node v implies that the links initiated from the node v to other nodes are of different levels of tie strength.

Notations	Definitions or Descriptions
n	number of nodes
d	number of original features
k	number of selected features
c	number of pseudo labels
$\mathcal{F}$	original feature set
$\mathcal{S}$	selected feature set
v_i	the i-th node on the attributed network
$\mathbf{x}_i \in \mathbb{R}^d$	feature vector of node v_i
$\mathbf{X} \in \mathbb{R}^{n\times d}$	feature matrix of all n nodes
$\mathbf{A} \in \{0,1\}^{n\times n}$	adjacency matrix of the attributed network
$\mathbf{a}_i \in \mathbb{R}^n$	tie strength vector for node v_i
Ω_i	constraint on $\mathbf{a}_i$
$\mathbf{w} \in \{0,1\}^d$	the selection indicator vector
$\mathcal{N}(v_i)$	neighbors of v_i

Table 1: Symbols.

ℓ_2-norm of a vector $\mathbf{a} \in \mathbb{R}^d$ is $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^T\mathbf{a}}$. The ℓ_1-norm of $\mathbf{a} \in \mathbb{R}^d$ is $\|\mathbf{a}\|_1 = \sum_{i=1}^{d} |\mathbf{a}_i|$. The main symbols are summarized in Table 1.

Definition 2.1. (**Attributed Networks**): An attributed network $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ consists of - (1) $\mathcal{V}$: the set of nodes ($n = |\mathcal{V}|$); (2) $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$: the set of edges ($e = |\mathcal{E}|$); and (3) the node attributes $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$, where $\mathbf{x}_i \in \mathbb{R}^d$ ($i = 1, ..., n$) is the attribute information of the i-th node.

With the definition of attributed networks, we now formally define the studied problem as follows.

Problem 1. ***(Unsupervised Feature Selection on Attributed Networks)**: Given an attributed network G, the problem of unsupervised feature selection on attributed networks aims to select a subset of k features $\mathcal{S}$ from the original d-dimensional feature space $\mathcal{F}$ ($k \ll d$) which are the most tightly correlated to the network structure. In particular, we assume that the selected features are encoded in a feature indicator vector $\mathbf{w} \in \{0,1\}^d$ such that $\mathbf{1}^T\mathbf{w} = k$, where $w_i = 1$ if the i-th feature is included in $\mathcal{S}$; otherwise $w_i = 0$.*

Next, we will introduce the proposed framework which finds the optimal adaptive neighborhood structure around each node for unsupervised feature selection on attributed networks. We use features and attributes interchangeably in the remaining parts.

3 THE PROPOSED FRAMEWORK

In this section, we present the proposed framework of adaptive unsupervised feature selection on attributed networks - ADAPT.

As the main focus on this work is to make the nodal features interact synergistically with the network topology, the observed features can be employed to obtain the adaptive neighborhood structure around each node such that the heterogeneity of tie strength can be well characterized. To find a subset of the most informative features, we assume that the network structure can be regenerated (i.e., network reconstruction) through these informative features via a probabilistic framework. At last, to capture the inherent correlation between network structure and node attributes, we impose a constraint on the network reconstruction process to ensure that it preserves the adaptive neighborhood structure measured by the tie strength. An illustration of these three aforementioned steps of ADAPT is shown in Figure 1. In the following parts of this paper, we will elaborate on these three phases in details.

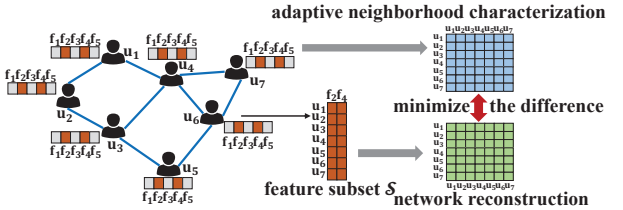


Figure 1: An illustration of the proposed framework - ADAPT. In the above depicted attributed network, we make use of the observed features to characterize the adaptive neighborhood structure around each node. The features f_2 and f_4 are selected as relevant features as they can be exploited to maximize the probability of observed links (a.k.a. network reconstruction). To capture the correlation between node attributes and network structure, we impose a constraint to minimize the difference between the probabilistic distributions from the aforementioned two phases.

3.1 Characterizing the Adaptive Neighborhood

As mentioned previously, existing methods either exploit the links at the community level or treat all links equally as binary relations. Hence, they fail to capture the finer-grained tie strength for feature selection. First, we embark on the definition of the adaptive neighborhood structure around each node, which is used to represent the heterogeneity of tie strength over its full spectrum.

Definition 3.1. (**Adaptive Neighborhood Structure**): Given an attributed network G, the adaptive neighborhood structure around a particular node v_i is encoded in the tie strength vector $\mathbf{a}_i = [a_{i1}, ...a_{in}]^T \in \mathbb{R}^n$ with the following constraints: (i) $\mathbf{1}^T\mathbf{a}_i = 1$; (ii) $a_{ij} \geq 0, \forall v_j \in \mathcal{N}(v_i)$; (iii) $a_{ij} = 0, \forall v_j \notin \mathcal{N}(v_i)$. We denote the constraint that is imposed on $\mathbf{a}_i$ as Ω_i.

The above definition implies that if node v_j is not directly connected with node v_i, then the tie strength between them is specified as 0. Otherwise, we adaptively quantify the tie strength of existing links from node v_i to ensure the summation of their tie strengths equals to 1. Next, we will investigate how to characterize the tie strength vector $\mathbf{a}_i$ for each node v_i on the attributed network.

The proposed adaptive neighborhood characterization framework is motivated by the findings in sociology, especially the *dyadic hypothesis* [11]. Specifically, it insinuates that the strength of a tie is largely determined by how similar the characteristics of two end nodes are, which in fact, can be further modeled as a hidden effect of the nodal attribute similarity [39]. However, characterizing the adaptive neighborhood structure from the attribute information has its unique challenges. Firstly, as label information in attributed networks is both time and labor intensive to acquire, we are in short of reliable ground truths to measure whether a node pair is indeed similar or not. Second, for each individual in the network, the number of its strongly connected nodes and the number of weakly connected nodes could differ remarkably. Hence, it necessitates a principled solution that can find the optimal number of neighbors for each node in the adaptive neighborhood characterization phase.

To tackle this problem, we introduce the concept of pseudo class labels to portray the characteristics of nodes. For example, pseudo labels can imply the interests or political polarizations of users in a social network or the research areas of scholars in an academic collaboration network. For each node v_i, we use $\mathbf{y}_i \in \{0,1\}^c$ to denote its pseudo class label vector (c is the number of pseudo classes) and we assume it can be obtained by applying a mapping function $f(.) : \mathbb{R}^d \rightarrow \mathbb{R}^c$ on its feature vector $\mathbf{x}_i$:

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon_i \quad (\forall i = 1, ..., n), \tag{1}$$

where $\epsilon_i \in \mathbb{R}$ $(i = 1, ..., n)$ are assumed to be the independent noisy terms. We also assume $f(\mathbf{x}_i)$ is a Lipschitz continuous function [19, 42] such that for a given node pair v_i and v_j, it holds that $|f(\mathbf{x}_i)_k - f(\mathbf{x}_j)_k| \leq L.d(\mathbf{x}_i, \mathbf{x}_j)$ for any pseudo class label k, where $L \geq 0$ is referred to as a Lipschitz constant, $d(.,.)$ is a predefined distance function, and $f(\mathbf{x})_k$ denotes the k-th entry of $f(\mathbf{x})$.

With the concept of pseudo labels, we can characterize the adaptive neighborhood structure around each node. Specifically, we assume that the pseudo label of each node v_i $(i = 1, ..., n)$ can be estimated via a weighted average of the noisy pseudo labels of its neighbors on the attributed network. Let the estimated pseudo class label vector of node v_i be $\hat{f}(\mathbf{x}_i)$, then it holds that $\hat{f}(\mathbf{x}_i) \approx \sum_{j=1}^n a_{ij}\mathbf{y}_j$. More concretely, to obtain the optimal weight vector $\mathbf{a}_i$, we can minimize the Manhattan distance[2] between the estimator and the ground truth pseudo class label (without noise) as:

$$\min \| \sum_{j=1}^{n} a_{ij}\mathbf{y}_j - f(\mathbf{x}_i)\|_1 \ \text{ s.t. } \mathbf{a}_i \in \Omega_i, \quad (\forall i = 1, ..., n). \tag{2}$$

where Ω_i denotes the constraint of $\mathbf{a}_i$ as mentioned before. The above formulation enables to adaptively find the adaptive neighborhood structure around node v_i. In particular, the node v_j has a stronger tie with node v_i (w.r.t. the pseudo label estimation) if the corresponding value a_{ij} is higher, and vice versa. It is also in line with the *dyadic hypothesis* [11] as higher nodal attribute similarity implies similar pseudo labels, which leads to stronger tie strength.

3.2 Reconstruct the Network

Our target is to find a subset of features $\mathcal{S}$ that are the most tightly correlated to the network structure, thus we assume that the observed links $\mathcal{E}$ on the network can be reconstructed from the feature subset $\mathcal{S}$. We first define how to measure the node similarity w.r.t. a subset of informative features $\mathcal{S}$.

Definition 3.2. (**Node Similarity w.r.t. $\mathcal{S}$**): Given a feature subset $\mathcal{S}$ and the corresponding feature indicator vector $\mathbf{w} \in \{0,1\}^d$, the node similarity between two nodes v_i and v_j on the attributed network G is defined as $s_{ij} = \mathbf{x}_i^T \text{diag}(\mathbf{w})\mathbf{x}_j$.

From a generative point of view, given the node v_i, we assume that the probability of an edge (e.g., $(v_i, v_j) \in \mathcal{E}$) can be decided by quantifying how similar the two end nodes v_i and v_j are in the feature space $\mathcal{S}$. Specifically, we refer the node v_i as the *source node* and the node v_j as the *target node*, and the *target node* v_j is treated as in the context of the *source node* v_i. Then for each observed link $(v_i, v_j) \in \mathcal{E}$, the probability that v_j is in the context of v_i is determined by the softmax function as follows:

$$p(v_j|v_i) = \frac{\exp(s_{ij})}{\sum_{m=1}^n \exp(s_{im})} = \frac{\exp(\mathbf{x}_i^T \text{diag}(\mathbf{w})\mathbf{x}_j)}{\sum_{m=1}^n \exp(\mathbf{x}_i^T \text{diag}(\mathbf{w})\mathbf{x}_m)}. \tag{3}$$

[2]We choose to use the Manhattan distance for the sake of simplicity, but it can be extended to other distance measures.

From the above formulation we can observe that the more similar node v_j and node v_i is in the feature space $\mathcal{S}$, the more likely we can reconstruct the observed link $(v_i, v_j) \in \mathcal{E}$. In other words, given the *source node* v_i, the above softmax function actually defines the conditional distribution vector $\mathbf{p}_i = [p(v_1|v_i), ..., p(v_n|v_i)]^T$, which measures the probability of observed links that are initiated from v_i on the attributed network.

3.3 Capturing the Correlation between Network Structure and Node Attributes

We have shown that the role of the observed features $\mathcal{F}$ is to quantify the tie strength of observed links to obtain the optimal adaptive neighborhood structure, while the informative feature subset $\mathcal{S}$ can be used to reconstruct the original network structure. To further capture the inherent correlation between node attributes and network structure, for each node v_i, we enforce its conditional distribution vector $\mathbf{p}_i = [p(v_1|v_i), ..., p(v_n|v_i)]^T$ to preserve the optimal adaptive neighbor structure specified by the tie strength vector $\mathbf{a}_i$. This target can be achieved by minimizing the KL divergence [17] between the distributions $\mathbf{p}_i$ and $\mathbf{a}_i$. By summing up the KL divergence for all nodes, we obtain the following problem:

$$\begin{aligned} &\min \sum_{i=1}^{n} \gamma_i D_{KL}(\mathbf{a}_i || \mathbf{p}_i) \\ &\text{s.t. } \mathbf{1}^T \mathbf{w} = k; \ \ w_m \in \{0, 1\}, \ \ (\forall m = 1, ..., d). \end{aligned} \tag{4}$$

In the above equation, γ_i is introduced to show the prestige of node v_i, and its value can be determined by various node centrality measures such as degree centrality and PageRank [43]. We can further expand the above objective function value as follows:

$$\sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \log(a_{ij}) - \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \log p(v_i|v_j). \tag{5}$$

In the above equation, the computation of the conditional probability $p(v_j|v_i)$ is very expensive due to the summation of all possible terms $\sum_{m=1}^{n} \exp(\mathbf{x}_i^T \text{diag}(\mathbf{w}) \mathbf{x}_m)$ in the denominator of the softmax function, especially when the number of nodes n is large. To address this issue, we make use of the negative sampling approach proposed in [28] to reformulate the optimization problem in Eq. (4):

$$\begin{aligned} &\min \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \log(a_{ij}) \\ &- \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \underbrace{\Big(\log\sigma(s_{ij}) + \sum_{m=1}^{K} \mathbb{E}_{v_m \sim P_v} [\log\sigma(-s_{im})] \Big)}_{\Theta_{ij}} \\ &\text{s.t. } \mathbf{1}^T \mathbf{w} = k; \ \ w_m \in \{0, 1\}, \ \ (\forall m = 1, ..., d). \end{aligned} \tag{6}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function, and K is the number of negative samples. Given the node v_i, the task now is to distinguish its neighborhood nodes v_j from other K nodes randomly drawn from the noisy distribution P_v. In other words, we attempt to upweight the similarity of connected nodes v_i and v_j while downweight the similarity between v_i and a randomly selected node v_m. Specifically, we follow [28, 36] to set P_v proportional to the node degree distribution raised to the power of 3/4.

The optimization problem in Eq. (6) is NP-hard due to the discrete nature of $\mathbf{w}$. To address this issue, we relax the discrete constraint on $\mathbf{w}$ by reformulating it as a real-valued vector in the range of $[0, 1]$ [38]. Furthermore, we rewrite the constraint $\mathbf{1}^T \mathbf{w} = k$ in the Lagrangian, resulting in the following optimization problem:

$$\begin{aligned} &\min \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \log(a_{ij}) - \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij} \Theta_{ij} + \alpha \|\mathbf{w}\|_1 \\ &\text{s.t. } 0 \le w_m \le 1 \ \ (\forall m = 1, ..., d), \end{aligned} \tag{7}$$

where α controls the sparsity of feature indicator vector $\mathbf{w}$.

To summarize, the adaptive neighborhood structure around each node v_i is in the form of the tie strength vector $\mathbf{a}_i$. It can be learned from Eq. (2) and is related to two variables (or functions) - the mapping function $f(.)$, and the noisy pseudo class label $\mathbf{y}_i$. Also, we assume that the probability of observed links is determined by the node similarity in the feature space $\mathcal{S}$ and we enforce the probability of *target nodes* in the context of the *source node* to preserve the adaptive neighborhood structure measured by the tie strength vector. Concretely, we minimize the KL divergence of these two probabilistic distributions, i.e., the probability of observed links and the adaptive neighborhood structure. In this way, the adaptive neighborhood structure acts as a bridge to capture the correlation between node attributes and network structure.

4 THE OPTIMIZATION FRAMEWORK

In this section, we discuss how to obtain the adaptive neighborhood structure $\mathbf{a}_i$ for each node v_i and the feature indicator vector $\mathbf{w}$.

4.1 Learning the Adaptive Neighborhood

First, we learn the optimal adaptive neighborhood structure $\mathbf{a}_i$ for each node v_i through Eq. (2). However, the update of the adaptive neighborhood structure $\mathbf{a}_i$ is difficult as $\mathbf{a}_i$ is related to the mapping function $f(.)$ and the noisy pseudo class label $\mathbf{y}_i$. In unsupervised feature selection, both $f(.)$ and $\mathbf{y}_i$ are unknown. Meanwhile, the constraint Ω_i on $\mathbf{a}_i$ further complicates the optimization problem on $\mathbf{a}_i$. Following [2], we reformulate the optimization problem in Eq. (2) into the following problem which yields an upper bound guarantee of high confidence:

$$\min_{\mathbf{a}_i} \|\mathbf{a}_i\|_2 + M \sum_{j=1}^{n} a_{ij} d(\mathbf{x}_j, \mathbf{x}_i), \ \ \text{s.t. } \mathbf{a}_i \in \Omega_i, \tag{8}$$

where M is a positive constant. The Lagrangian of the above problem is as follows:

$$\begin{aligned} L = \|\mathbf{a}_i\|_2 + \mathbf{a}_i^T \mathbf{u}_i + \lambda (1 - \sum_{j=1}^{n} a_{ij}) \\ - \sum_{v_j \in \mathcal{N}(v_i)} \theta_j a_{ij} + \sum_{v_j \notin \mathcal{N}(v_i)} \eta_j a_{ij}, \end{aligned} \tag{9}$$

where $\mathbf{u}_i = [M.d(\mathbf{x}_1, \mathbf{x}_i), ..., M.d(\mathbf{x}_n, \mathbf{x}_i)]^T$. The parameters $\lambda \in \mathbb{R}$, $\eta_j \in \mathbb{R}$ $(\forall v_j \notin \mathcal{N}(v_i))$ and $\theta_j \ge 0$ $(\forall v_j \in \mathcal{N}(v_i))$ are the Lagrange multipliers. As Eq. (8) is convex, thus any solution that satisfies the KKT condition [5] guarantees a global optimum. By setting the derivative of the Lagrangian w.r.t. $\mathbf{a}_i$ to zero, we obtain:

$$\frac{a_{ij}}{\|\mathbf{a}_i\|_2} = \begin{cases} \lambda + \theta_j - u_{ij}, & \forall v_j \in \mathcal{N}(v_i) \\ \lambda - \eta_j - u_{ij}, & \forall v_j \notin \mathcal{N}(v_i). \end{cases} \tag{10}$$

Let $\mathbf{a}_i^*$ be the optimal solution, according to the complementary slackness condition, for any $a_{ij}^* > 0$, we obtain the following:

$$a_{ij}^*/\|\mathbf{a}_i^*\|_2 = \lambda - u_{ij}. \tag{11}$$

And the optimal solution of $\mathbf{a}_i^*$ is given as follows:

$$a_{ij}^* = \frac{\{\lambda - u_{ij}\} \cdot \mathbf{1}\{\lambda > u_{ij}\}}{\sum_{j=1}^{n}\{\lambda - u_{ij}\} \cdot \mathbf{1}\{\lambda > u_{ij}\}}, \tag{12}$$

for $\forall v_j \in \mathcal{N}(v_i)$, where $\mathbf{1}(.)$ is an indicator function. According to the definition of $\mathbf{a}_i$, it has a cutoff effect as $a_{ij} = 0$ if $v_j \notin \mathcal{N}(v_i)$. In addition to that, it can be observed that $\mathbf{a}_i$ also has a cutoff effect for $v_j \in \mathcal{N}(v_i)$ when the condition $\lambda > u_{ij}$ is satisfied. In other words, for each node v_i, there exists $0 \le k_i^* \le |\mathcal{N}(v_i)|$ such that only the tie strengths to these k_i^* neighbors are nonzero. Without the loss of generality, we assume that for each node v_i, the index of the other nodes are ordered in an ascending order w.r.t. u_{ij}, then by squaring and summing Eq. (12) over all nonzero entries in $\mathbf{a}_i^*$ and solving the equation, we have:

$$\lambda = \frac{1}{k_i^*}\Big(\sum_{i=1}^{k_i^*} u_{ij} + \sqrt{(\sum_{j=1}^{k_i^*} u_{ij})^2 - k_i^* \sum_{j=1}^{k_i^*} u_{ij}^2 + k_i^*}\Big). \tag{13}$$

Following [2], a greedy algorithm is leveraged to add neighbors v_j of node v_i according to the value of u_{ij} until the optimal number of neighbors k_i^* is fulfilled (until the condition $\lambda > u_{ij}$ does not hold any more). From Eq. (12), we can find the optimal solution a_{ij}^* is closely related to u_{ij}, while u_{ij} is proportional to the distance between node v_i and v_j in the observed features $\mathcal{F}$. In other words, the optimal adaptive neighborhood structure can be obtained from the original feature space $\mathcal{F}$ in a greedy manner.

4.2 Learning the Feature Indicator Vector

In the previous subsection, we have discussed how to obtain the formulation of the adaptive neighborhood structure $\mathbf{a}_i$ $(i = 1, ..., n)$ for each node on the attributed network. To update the feature indicator vector $\mathbf{w}$, we plug the optimal solution of $\mathbf{a}_i^*$ into Eq. (7), resulting in the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij}^* \log(a_{ij}^*) - \sum_{(i,j)\in\mathcal{E}} \gamma_i a_{ij}^* \Theta_{ij} + \alpha\|\mathbf{w}\|_1 \\ &\text{s.t. } 0 \le w_m \le 1 \ \ (\forall m = 1, ..., d). \end{aligned} \tag{14}$$

We apply projected stochastic gradient descent method to optimize the above objective function. Specifically, at each step, we sample a mini-batch of edges and then update the feature indicator vector $\mathbf{w}$. Suppose the edge (v_i, v_j) is sampled, then the objective function in Eq. (14) is reduced to:

$$\begin{aligned} \mathcal{L}(i,j) = &\underbrace{\gamma_i a_{ij}^* \log(a_{ij}^*)}_{\mathcal{L}_1(i,j)} + \underbrace{(-\gamma_i a_{ij}^* \log\sigma(s_{ij}))}_{\mathcal{L}_2(i,j)} \\ &+ \underbrace{(-\gamma_i a_{ij}^* \sum_{m=1}^{K} \mathbb{E}_{v_m\sim P_v}[\log\sigma(-s_{im})])}_{\mathcal{L}_3(i,j)} + \alpha\|\mathbf{w}\|_1. \end{aligned} \tag{15}$$

In the above formulation, the first term $\mathcal{L}_1(i,j)$ is independent of the feature indicator vector $\mathbf{w}$, while the partial derivative of $\mathcal{L}_2(i,j)$, and $\mathcal{L}_3(i,j)$ w.r.t. w_k can be calculated as follows:

$$\begin{aligned} \frac{\partial\mathcal{L}_2(i,j)}{\partial w_k} &= -\gamma_i a_{ij}^*(1 - \sigma(s_{ij}))x_{ik}x_{jk} \\ \frac{\partial\mathcal{L}_3(i,j)}{\partial w_k} &= \gamma_i a_{ij}^* \sum_{m=1}^{K} \mathbb{E}_{v_m\sim P_v}\sigma(s_{im})x_{ik}x_{mk}. \end{aligned} \tag{16}$$

Normally, the ℓ_1-norm regularization term on $\mathbf{w}$ is not smooth and its derivative is not achievable everywhere over its feasible region. However, as we relax the constraint on $\mathbf{w}$ to make it in the range of [0,1], the partial derivative of $\alpha\|\mathbf{w}\|_1$ w.r.t. w_k is α. To summarize, we have:

$$\frac{\partial\mathcal{L}(ij)}{\partial w_k} = \frac{\partial\mathcal{L}_2(i,j)}{\partial w_k} + \frac{\partial\mathcal{L}_3(i,j)}{\partial w_k} + \alpha. \tag{17}$$

With the partial derivative, the update step of w_k is given as follows:

$$w_k \leftarrow P\Big[w_k - \rho\frac{\partial\mathcal{L}(ij)}{\partial w_k}\Big], \tag{18}$$

where $P[x]$ is a box projection operation which maps $x \in \mathbb{R}$ in the bounded region of [0,1]. Specifically, if $0 \le x \le 1$, then $P[x] = x$; and if $x > 1$, then $P[x] = 1$; otherwise if $x < 0$, then $P[x] = 0$. Meanwhile, ρ is the learning rate, and in practice, it can be adaptively adjusted to facilitate the convergence [4].

Algorithm 1 The proposed ADAPT framework.

Input: Attributed network $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, the number desired features k, parameters α, and M.
Output: The top-k ranked features.

```
Initialize the feature indicator vector w;
Update a_i via the greedy algorithm in Section 4.1 (∀i = 1, ..., n);
for epoch = 1 : maxepoch do
    for batch = 1 : numbatch do
        Sample a mini-batch of edges and negative samples;
        Update w_k (∀k = 1, ..., d) according to Eq. (18) ;
    end for
end for
Rank features according to the entries in w.
```

Summary: Now we summarize the proposed optimization algorithm for the developed adaptive unsupervised feature selection framework. At the very beginning, we initialize the feature indicator vector $\mathbf{w}$ and update the adaptive neighborhood structure $\mathbf{a}_i$ $(i = 1, ..., n)$ for all nodes via the greedy algorithm in Section 4.1; then we sample a mini-batch of edges and a number of negative samples to update the feature indicator vector $\mathbf{w}$ through projected stochastic gradient descent, where the update rule is given in Eq. (18). The detailed pseudocode is illustrated in Algorithm 1.

Time Complexity Analysis: As previously discussed, we first make use of the observed features to update the tie strength vector $\mathbf{a}_i$ for each node. The complexity of updating $\mathbf{a}_i$ is $O(k_i + d|\mathcal{N}(v_i)| + |\mathcal{N}(v_i)|\log|\mathcal{N}(v_i)|)$, where k_i denotes the optimal number of neighbors for node v_i on the attributed network. Then in each epoch, we update the feature indicator vector $\mathbf{w}$ by sampling a mini-batch of edges and the negative samples. For each epoch, the complexity of updating $\mathbf{w}$ is $O(edK) + \sum_{i=1}^{n} O(dk_i|\mathcal{N}(v_i)|)$. The negative sampling in each epoch takes $O(e)$ with the alias table [20].

5 EXPERIMENTAL EVALUATIONS

In this section, we perform experiments on real-world attributed networks of various types to validate the effectiveness of the proposed ADAPT framework. We first introduce the datasets and experimental settings before presenting details of the experiments.

5.1 Datasets

In the experiments, we collect four widely used real-world attributed networks Wiki, BlogCatalog, Flickr, and ACM [16, 23, 40].

Wiki: Wiki is a collection of Wikipedia documents that are inherently connected with each other via hyperlinks. Each document is categorized into a number of predefined classes. The dataset contains 2,405 Wikipedia documents from 19 different classes. Each document is described by 4,973-dimensional TFIDF features. In total, there are 12,178 hyperlinks among these Wikipedia documents.

BlogCatalog: BlogCatalog is a social blogging website where users follow each other and post blogs with certain categories. The tags of the blogs posted by users are taken as the features while the main category of blogs by the users is regarded as the ground truth label. In the dataset, we have 5,196 users, 171,743 social relations, 8,189 features and 6 categories of blogs.

Flickr: Flickr is an image sharing website where users interact with each other via photo sharing under predefined categories. Specifically, each user can specify a set of tags as their interests which are considered as their attributes. Meanwhile, the main category of images the user post is regarded as the ground truth label of users. In the collected dataset, there are 7,575 users, 12,047 features, 239,738 social relations, and 9 classes.

ACM: This dataset is a subgraph of citation network of papers published before 2016 in ACM organized venues. Each publication is described by the bag-of-words features based on the abstract and is categorized into one of the 9 predefined classes such as machine learning and data mining. In the dataset, we have 16,484 publications, 8,337 features, and 71,980 links.

5.2 Experimental Settings

To verify the effectiveness of the proposed method, we compare ADAPT with the following baseline methods:

- **LS** [14]: selects features that can best preserve the local manifold structure of data.
- **MCFS** [6]: performs feature selection based on spectral analysis and sparse regression.
- **GreedyFS** [9]: selects features in a greedy manner by measuring the reconstruction error of the data.
- **LUFS** [34]: selects discriminative features on the attributed networks with the aid of extracted social dimensions.
- **NetFS** [24]: enables feature selection on attributed networks with embedded network latent representation learning.
- **MMOP** [38]: selects features on attributed networks that can maximally preserve the partial order preserving principle.
- **GFS** [37]: finds relevant features on attributed networks by modeling network and attributes with a generative process.

Among them, LS, MCFS, GreedyFS are conventional unsupervised feature selection methods with node attribute information only. They respectively belong to the similarity based, sparse learning based, and reconstruction based methods, which are the three most widely used categories of unsupervised feature selection methods. LUFS and NetFS exploit the link information at a coarse granularity level via community analysis, while MMOP and GFS directly make use of the link information for unsupervised feature selection but treat all links equally. As these methods are from different categories, their comparisons against ADAPT could further reveal the superiority of the developed framework.

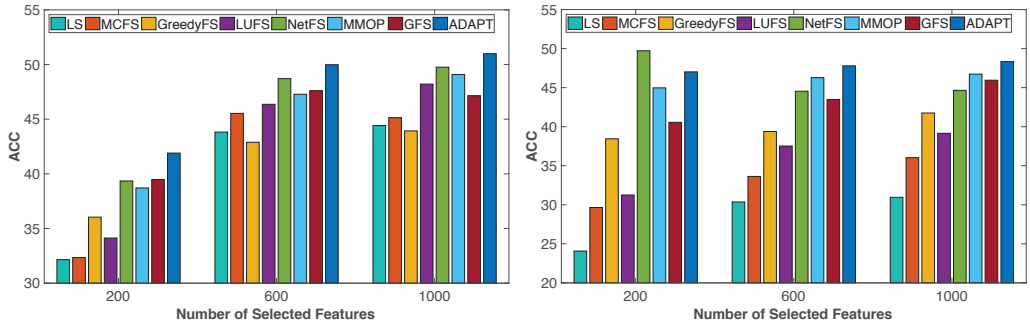
Following the commonly adopted setting [8, 9, 14, 24, 37, 38] in evaluating unsupervised feature selection, we assess the performance of the developed framework in terms of clustering. Specifically, each feature selection method is first applied to select relevant features on the attributed networks, then we perform K-means clustering based on the selected features. As K-means may converge to a local optimum, we repeat the K-means process repeatedly over 20 times and report the average clustering results. Two widely used evaluation metrics, clustering accuracy (ACC) and normalized mutual information (NMI) are used. Normally, the higher the ACC and NMI are, the better the selected features are.

For LS and MCFS, we set the number of neighborhood size to be 5 to construct the affinity graph with the heat kernel. In GreedyFS, the number of feature partitions is set to be 5. For MCFS, LUFS, and NetFS, the number of clusters or pseudo labels is specified to be the true number of classes. In ADAPT, the number of negative samples K is specified as 5. In addition, different methods have different sets of regularization parameters, and these parameters are often difficult to determine in an unsupervised scenario. To have a fair comparison between these methods, we tune these parameters via grid search and the best average clustering results are reported.

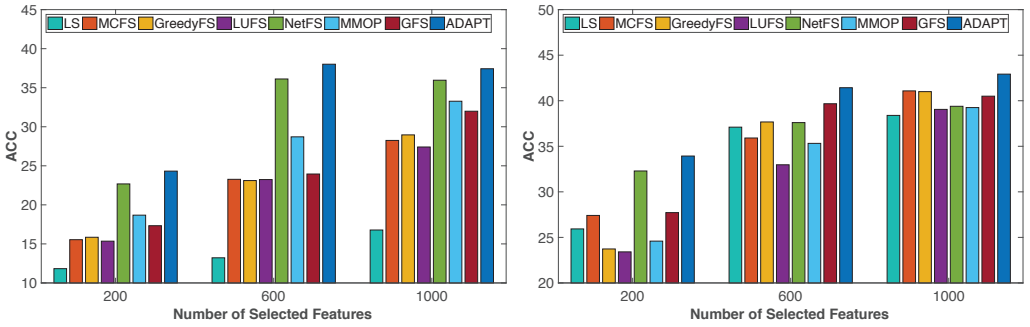
5.3 Effectiveness of ADAPT

First, we investigate the effectiveness of the proposed ADAPT framework by comparing the clustering performance against other baseline methods after feature selection. The number of selected features is varied in the range of $\{200, 600, 1000\}$. The comparison results are shown in Figure 2 and Figure 3. We make the following observations from these figures:

- The proposed ADAPT framework obtains the best clustering performance in almost all cases w.r.t. different numbers of selected features. To further validate it, we perform a pairwise Wilcoxon signed rank test between ADAPT and baseline methods and the results indicate that the improvement of ADAPT is significant at the level of 0.05.
- The proposed ADAPT is superior to LUFS and NetFS which model the network information at the community level. Meanwhile, it also achieves better performance than MMOP and GFS which treat all observed links equally. The improvement of ADAPT over these methods corroborate the importance of characterizing adaptive neighborhood structure around each node for feature selection. As LUFS is very sensitive to the noisy and incomplete network structure, its performances are the worst among these five methods.
- LS, MCFS, and GreedyFS are conventional unsupervised feature selection methods which only make use of the node attribute information. Their performance is inferior to NetFS, MMOP, and GFS in many cases. The observation supports the assumption that network structure complement node attribute information for feature selection.

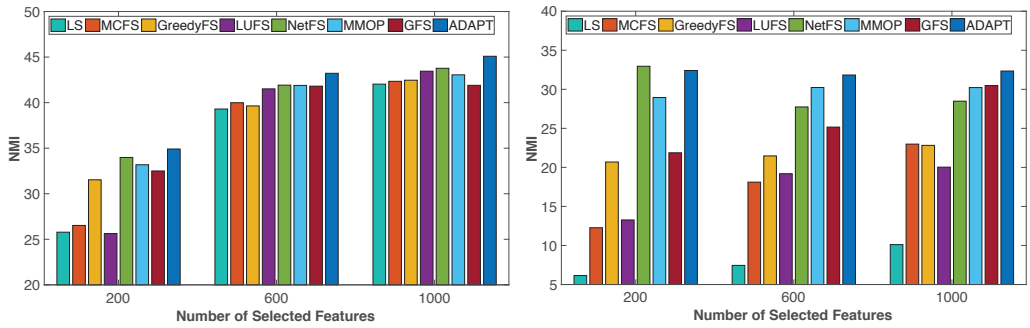


(a) Wiki (b) BlogCatalog

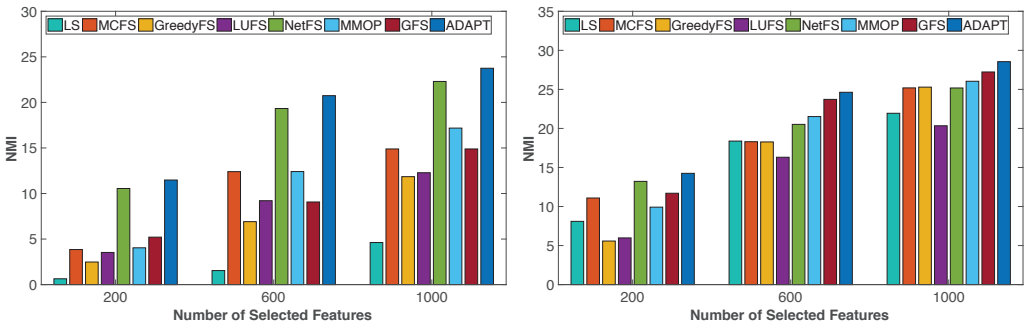


(c) Flickr (d) ACM

Figure 2: Comparisons of clustering results (ACC) of different unsupervised feature selection algorithms.



(a) Wiki (b) BlogCatalog



(c) Flickr (d) ACM

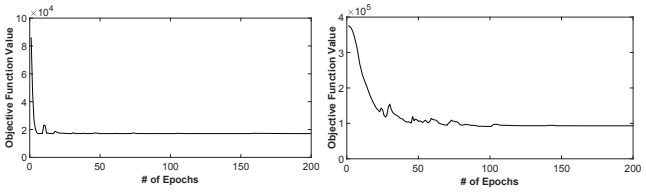
Figure 3: Comparisons of clustering results (NMI) of different unsupervised feature selection algorithms.

- The improvement of feature selection algorithms with network information (i.e., ADAPT, NetFS) over conventional feature selection methods (i.e., LS, MCFS, GreedyFS) is relatively higher on the BlogCatalog and Flickr datasets. The

reason is that the density of networks (the ratio between the number of edges and the number of nodes) is much higher on these two datasets, thus network information could provide more constraints in finding relevant features.

5.4 Convergence Analysis

Now we empirically show the convergence of Eq. (7) on the BlogCatalog and Wiki datasets in Figure 4. The observations are similar on the other datasets and we do not list the results to save space. Here, the mini-batch size of the stochastic gradient descent is set as 256, the learning rate is specified with the starting value of $\rho_0 = 0.01$ and it decays half every 10 epochs. From the figures, we can observe that the objective function value decreases very quickly at the very beginning and converges within 20 epochs in the BlogCatalog dataset; on the Wiki dataset, the convergence is relatively slower but it also converges within 100 epochs. The above observations demonstrate the effectiveness of the optimization algorithm.

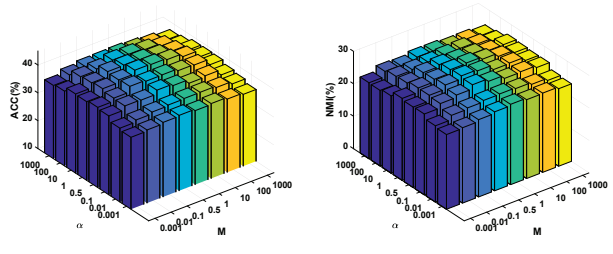


(a) BlogCatalog (b) Wiki

Figure 4: Convergence analysis of ADAPT.

5.5 Parameter Sensitivity Study

The proposed ADAPT framework has two important model parameters: (1) M controls the number of optimal neighbors around each node for tie strength characterization; and (2) α controls the sparsity of the feature indicator vector $\mathbf{w}$. To investigate how the variation of these two model parameters affects the feature selection performance, we vary them among $\{0.001, 0.01, 0.1, 0.5, 1, 10, 100, 1000\}$. We only show the parameter study results on the ACM dataset (with 1000 selected features) to save space as we have similar observations on the other datasets. Firstly, as can be observed from Figure 5, when we vary the value of M, the clustering performance first increases, then reaches its peak, and then gradually decreases. It implies that finding a suitable number of optimal neighbors around each node could advance feature selection. Secondly, we can observe that when the parameter α is between 0.1 and 10, the clustering performance is relatively stable. Hence, it is safe to tune α in a wide range without jeopardizing the clustering performance too much.



(a) ACC (b) NMI

Figure 5: Parameter study of ADAPT.

6 RELATED WORK

In this work, we briefly review related work on: (1) conventional feature selection; and (2) feature selection with networked data.

Conventional Feature Selection Feature selection is important in alleviating the curse of dimensionality [7, 13] by finding a subset of features of high quality from the original feature space and is essentially useful when the original features are indispensable for model interpretation and knowledge distillation. The merits of feature selection include the improvement of learning performance, the increase of model comprehensibility, and the reduction of storage and computational costs. Depending on whether label information is involved, existing methods can be broadly classified as supervised [32] and unsupervised algorithms [1]. Supervised feature selection directly makes use of the discriminative information embedded in the class labels to differentiate instances from different classes. However, in most cases, the label information is time-consuming and expensive to obtain which motivates a surge of research on unsupervised feature selection. Without label information in guiding the selection phase, existing efforts seek for alternative criteria to assess the relevance of features. For example, Laplacian Score [14] and SPEC [44] evaluate the importance of features by their ability in preserving the data similarity structure. Another prevalent choice is the so-called pseudo label based methods, examples that fall in this family include MCFS [6], UDFS [41], NDFS [27], RUFS [30], and FSASL [8]. Typically, they generate pseudo labels from data via some clustering algorithms and select features based on their utility in predicting the pseudo labels with sparse learning based framework. Furthermore, data reconstruction error recently emerged as a new criterion to define feature relevance by measuring its capacity in approximating the original data through a reconstruction function. The few existing attempts are GreedyFS [9] and REFS [25].

Feature Selection on Networked Data A vast majority of existing feature selection methods are fundamentally based on the data *i.i.d.* assumption. However, this assumption is untenable on networked data as instances are inherently correlated. In this regard, a lot of efforts have been made toward feature selection on networked data. Gu and Han [12] first studied supervised feature selection on networked data, in particular, a graph regularized sparse learning framework is developed to capture the correlation between network structure and node attributes. Tang and Liu [33] further investigated how to find relevant features from social media data by incorporating various types of social relations, and it was later extended to jointly find relevant instances and features simultaneously [35] as both instances and features could be noisy. As instances in networks could be highly idiosyncratic, Li et al. [26] customized the feature selection process for each node on the network by finding a subset of shared features and instance-specific features. The above-mentioned attempts, however, are limited with the use of label information, which is often tedious to obtain in practice. LUFS [34] was among one of the first unsupervised feature selection framework on networks. In particular, it leverages the community structure of nodes to facilitate the selection of relevant features, which is performed in two separate steps. Li et al. [24] proposed a robust framework NetFS to embed the community detection into feature selection, and the proposed framework is robust

to the noise among the observed links. However, it is argued that both LUFS and NetFS fail to take advantage of the fine-grained link information for feature selection. Therefore, Wei et al. [37, 38] proposed a framework based on the partial order relations among node pairs to exploit link information directly. Our studied problem also differs from attributed network embedding [15, 22, 40] as we focus on finding a subset of original node features for learning, which often gives models better readability and interpretability.

7 CONCLUSIONS AND FUTURE WORK

Two distinct but highly correlated data representations are naturally observed on real-world attributed networks. In many cases, high-dimensional nodal attributes increase the possibility of including noisy, redundant and network topologically irrelevant features, which hinder us to gain insights from such networks. Without the label supervision, efforts have been made to find a subset of features that are can be fused with network topology seamlessly for synergistic knowledge discovery. These methods, however, overwhelmingly exploit network structure at a rather coarse granularity, either at a macro-level through community analysis or treating all links equally with the partial order relations. In this regard, they fail to capture the finer-grained tie strength information embedded on the network, and the importance of which has been reinforced by the sociology findings. In this paper, we make the initial investigation to develop a principled adaptive unsupervised feature selection framework on attributed networks. Specifically, we study how to characterize the adaptive neighborhood structure around each node, and meanwhile, target at finding a high-quality feature subset to generate the observed links with a probabilistic framework. Additionally, to capture the inherent correlation among these two data representations, we constrain the probability of generated links to preserve the adaptive neighborhood structure measured by the tie strength. To validate the effectiveness of the proposed framework, we perform empirical evaluations on various real-world attributed networks, the results imply that ADAPT outperforms the state-of-the-art unsupervised feature selection methods. Further work includes generalizing ADAPT on more complex networks such as signed and heterogeneous attributed networks.

ACKNOWLEDGEMENTS

This material is, in part, supported by the National Science Foundation (NSF) under grant number 1614576. The authors would like to thank the anonymous reviewers for their constructive feedback.

REFERENCES

[1] Salem Alelyani, Jiliang Tang, and Huan Liu. 2013. Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications* 29 (2013), 110–121.
[2] Oren Anava and Kfir Levy. 2016. k*-Nearest Neighbors: From Global to Local. In *NIPS*. 4916–4924.
[3] Lars Backstrom and Jure Leskovec. 2011. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *WSDM*. 635–644.
[4] Yoshua Bengio. 2012. Practical Recommendations for Gradient-based Training of Deep Architectures. In *Neural Networks: Tricks of the Trade*. Springer, 437–478.
[5] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.
[6] Deng Cai, Chiyuan Zhang, and Xiaofei He. 2010. Unsupervised Feature Selection for Multi-Cluster Data. In *SIGKDD*. 333–342.
[7] Thomas M Cover and Joy A Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons.
[8] Liang Du and Yi-Dong Shen. 2015. Unsupervised Feature Selection with Adaptive Structure Learning. In *SIGKDD*. 209–218.
[9] Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. 2013. Efficient Greedy Feature Selection for Unsupervised Learning. *KAIS* 35, 2 (2013), 285–310.
[10] Eric Gilbert and Karrie Karahalios. 2009. Predicting Tie Strength With Social Media. In *SIGCHI*. 211–220.
[11] Mark S Granovetter. 1973. The Strength of Weak Ties. *Amer. J. Sociology* 78, 6 (1973), 1360–1380.
[12] Quanquan Gu and Jiawei Han. 2011. Towards Feature Selection in Network. In *CIKM*. 1175–1184.
[13] Isabelle Guyon and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *JMLR* 3, Mar (2003), 1157–1182.
[14] Xiaofei He, Deng Cai, and Partha Niyogi. 2006. Laplacian Score for Feature Selection. In *NIPS*. 507–514.
[15] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label Informed Attributed Network Embedding. In *WSDM*. 731–739.
[16] Xiao Huang, Qingquan Song, Jundong Li, and Xia Hu. 2018. Exploring Expert Cognition for Attributed Network Embedding. In *WSDM*. 270–278.
[17] Solomon Kullback. 1997. *Information Theory and Statistics*. Courier Corporation.
[18] Timothy La Fond and Jennifer Neville. 2010. Randomization Tests for Distinguishing Social Influence and Homophily Effects. In *WWW*. 601–610.
[19] Lubor Ladicky and Philip Torr. 2011. Locally Linear Support Vector Machines. In *ICML*. 985–992.
[20] Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. 2014. Reducing the Sampling Complexity of Topic Models. In *SIGKDD*. 891–900.
[21] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. *ACM CSUR* 50, 6 (2017), 94.
[22] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed Network Embedding for Learning in A Dynamic Environment. In *CIKM*. 387–396.
[23] Jundong Li, Xia Hu, Jiliang Tang, and Huan Liu. 2015. Unsupervised Streaming Feature Selection in Social Media. In *CIKM*. 1041–1050.
[24] Jundong Li, Xia Hu, Liang Wu, and Huan Liu. 2016. Robust Unsupervised Feature Selection on Networked Data. In *SDM*. 387–395.
[25] Jundong Li, Jiliang Tang, and Huan Liu. 2017. Reconstruction-based Unsupervised Feature Selection: An Embedded Approach. In *IJCAI*. 2159–2165.
[26] Jundong Li, Liang Wu, Osmar R Zaïane, and Huan Liu. 2017. Toward Personalized Relational Learning. In *SDM*. 444–452.
[27] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, Hanqing Lu, et al. 2012. Unsupervised Feature Selection using Nonnegative Spectral Analysis. In *AAAI*. 1026–1032.
[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*. 3111–3119.
[29] Joseph J Pfeiffer III, Sebastian Moreno, Timothy La Fond, Jennifer Neville, and Brian Gallagher. 2014. Attributed Graph Models: Modeling Network Structure with Correlated Attributes. In *WWW*. 831–842.
[30] Mingjie Qian and Chengxiang Zhai. 2013. Robust Unsupervised Feature Selection. In *IJCAI*. 1621–1627.
[31] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine* 29, 3 (2008), 93.
[32] Jiliang Tang, Salem Alelyani, and Huan Liu. 2014. Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications* (2014).
[33] Jiliang Tang and Huan Liu. 2012. Feature Selection with Linked Data in Social Media. In *SDM*. 118–128.
[34] Jiliang Tang and Huan Liu. 2012. Unsupervised Feature Selection for Linked Social Media Data. In *SIGKDD*. 904–912.
[35] Jiliang Tang and Huan Liu. 2013. Coselect: Feature Selection with Instance Selection for Social Media Data. In *SDM*. 695–703.
[36] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-Scale Information Network Embedding. In *WWW*. 1067–1077.
[37] Xiaokai Wei, Bokai Cao, and S Yu Philip. 2016. Unsupervised Feature Selection on Networks: A Generative View. In *AAAI*. 2215–2221.
[38] Xiaokai Wei, Sihong Xie, and Philip S Yu. 2015. Efficient Partial Order Preserving Unsupervised Feature Selection on Networks. In *SDM*. 82–90.
[39] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling Relationship Strength in Online Social Networks. In *WWW*. 981–990.
[40] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. 2015. Network Representation Learning with Rich Text Information. In *IJCAI*. 2111–2117.
[41] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. 2011. $\ell_{2,1}$-norm Regularized Discriminative Feature Selection for Unsupervised Learning. In *IJCAI*. 1589–1594.
[42] Kai Yu, Tong Zhang, and Yihong Gong. 2009. Nonlinear Learning using Local Coordinate Coding. In *NIPS*. 2223–2231.
[43] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.
[44] Zheng Zhao and Huan Liu. 2007. Spectral Feature Selection for Supervised and Unsupervised Learning. In *ICML*. 1151–1157.