Label Informed Attributed Network Embedding

WSDM 2017.

Xiao Huang
Texas A&M University
College Station

Jundong Li
Arizona State University

Xia Hu
Texas A&M University
College Station

Attributed networks

Different from plain networks in which only node-to-node interactions and dependencies are observed, each node in an attributed network is often associated with a rich set of features. such as academic networks and health care systems.

ANE targets at leveraging both network proximity and node attribute affinity.

Related Work

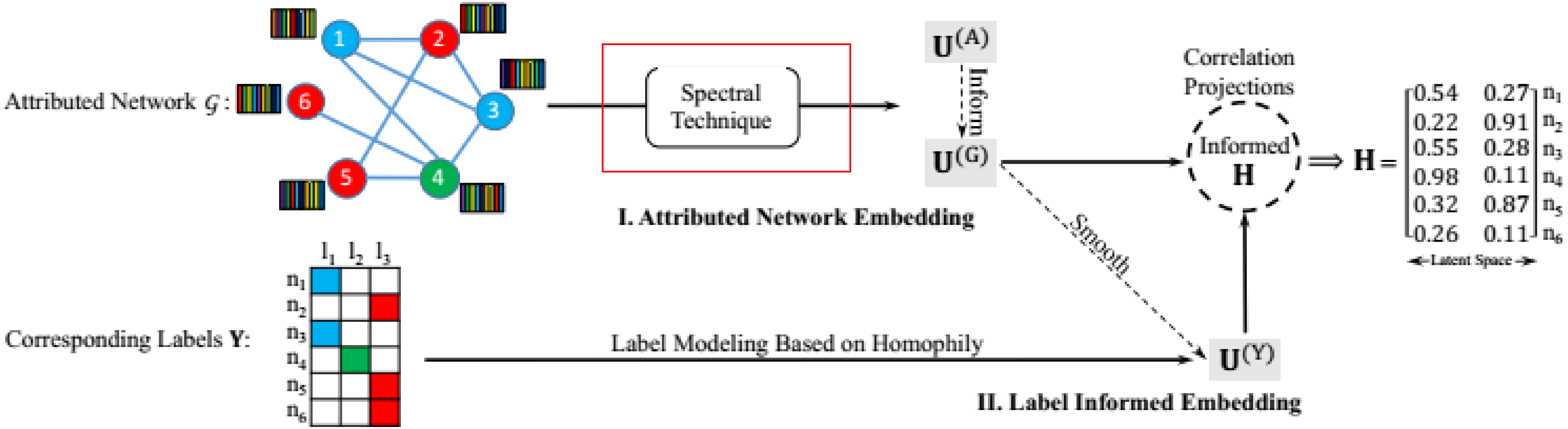
- Propose a novel framework LANE, which can affiliate labels with the attributed network and smoothly embed them into a low-dimensional representation by modeling their structural proximities and correlations;
- Present an effective alternating algorithm to solve the optimization problem of LANE;
- Empirically evaluate and validate the effectiveness of LANE on real-world attributed networks.

Notations

小写字母表示标量，加粗表示向量，大写字母加粗表示矩阵，表示矩阵的第i行,$\|\cdot\|_2$表示欧几里得范数，**I**表示单位矩阵

Notations	Definitions
$\mathbf{G}$	weighted adjacency matrix
$\mathbf{A}$	attribute information matrix
$\mathbf{Y}$	label information matrix
$\mathbf{H}$	final embedding representation
$\mathbf{S}^{(G)}$	network affinity matrix
$\mathbf{S}^{(A)}$	node attribute affinity matrix
n	number of nodes in the network
d	dimension of the embedding representation

LANE



Attributed Network $\mathcal{G}$:
Spectral Technique
$\mathbf{U}^{(A)}$
Inform
$\mathbf{U}^{(G)}$
I. Attributed Network Embedding
Correlation Projections
Informed H
$\Longrightarrow \mathbf{H} = \begin{bmatrix} 0.54 & 0.27 \\ 0.22 & 0.91 \\ 0.55 & 0.28 \\ 0.98 & 0.11 \\ 0.32 & 0.87 \\ 0.26 & 0.11 \end{bmatrix} \begin{matrix} n_1 \\ n_2 \\ n_3 \\ n_4 \\ n_5 \\ n_6 \end{matrix}$
Latent Space
Smooth
Corresponding Labels $\mathbf{Y}$:
l_1 l_2 l_3
n_1 n_2 n_3 n_4 n_5 n_6
Label Modeling Based on Homophily
$\mathbf{U}^{(Y)}$
II. Label Informed Embedding

Attributed networks Embedding

Specifically, we aim at allocating similar vector representations for nodes with similar geometrical or attribute proximities respectively

- the network structure U(G)

The key idea is to focus on each pair of nodes i and j. If they have similar locality properties, then their vector representations $\mathbf{u}_i$ and $\mathbf{u}_j$ should also be similar in the learned space. We use distance $\|\mathbf{u}_i - \mathbf{u}_j\|_2^2$ to measure this. For ex-

例如：节点1和节点3

Attributed networks Embedding

cosine measure s_{ij} to calculate the similarity of two nodes, and it is straightforward to extend to other measures. Since s_{ij} would be large if nodes i and j have similar network structures, and approach small value otherwise, we can use the following product to measure the *degree of disagreement* between s_{ij} and $\{\mathbf{u}_i, \mathbf{u}_j\}$:

$$s_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2. \qquad (1)$$

Attributed networks Embedding

$$\underset{\mathbf{U}^{(G)}}{\text{minimize}} \quad \frac{1}{2}\sum_{i,j=1}^{n} s_{ij}\left\|\frac{\mathbf{u}_i}{\sqrt{d_i}} - \frac{\mathbf{u}_j}{\sqrt{d_j}}\right\|_2^2. \tag{2}$$

$\mathbf{u}_i$ and $\mathbf{u}_j$ are the i^{th} and j^{th} rows of network latent representation $\mathbf{U}^{(G)}$. We represent the pairwise similarities as a graph affinity matrix $\mathbf{S}^{(G)}$, where s_{ij} is its $(i,j)^{\text{th}}$ element. d_i and d_j are the sum of i^{th} and j^{th} rows of $\mathbf{S}^{(G)}$. We utilize

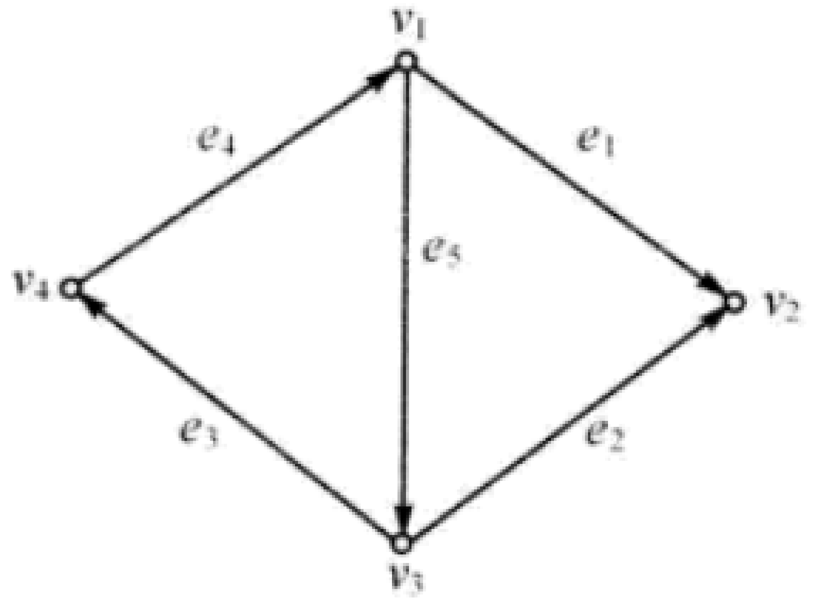
设任意图$G=(V,E)$，其中顶点集$V=v_1,v_2,\dots,v_n$，边集$E=e_1,e_2,\dots,e_\varepsilon$。用$m_{ij}$表示顶点$v_i$与边$e_j$关联的次数，可能取值为0,1,2,…，称所得矩阵$\mathbf{M}(G)=(m_{ij})_{n\times\varepsilon}$为图G的**关联矩阵**

类似地，有向图D的关联矩阵$\mathbf{M}(D)=(m_{ij})_{n\times\varepsilon}$的元素$m_{i\times j}$定义为：

$$m_{ij}=\begin{cases}1, & v_i\text{是有向边}a_j\text{的始点}\\ \\ -1, & v_i\text{是有向边}a_j\text{的终点}\\ \\ 0, & v_i\text{是有向边}a_j\text{的不关联点}\end{cases}$$

关联矩阵:

$$\begin{bmatrix}1 & 0 & 0 & -1 & 1\\ -1 & -1 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & -1\\ 0 & 0 & -1 & 1 & 0\end{bmatrix}$$



注：关联矩阵是描述图的另一种矩阵表示。

Attributed networks Embedding

Based on the definition of normalized graph Laplacian [8], we can reformulate Eq. (2) into a maximization problem, and model the geometrical proximity via the objective function as follows

$$\begin{aligned} \underset{\mathbf{U}^{(G)}}{\text{maximize}} \quad & \mathcal{J}_G = \mathrm{Tr}(\mathbf{U}^{(G)^T} \mathcal{L}^{(G)} \mathbf{U}^{(G)}) \\ \text{subject to} \quad & \mathbf{U}^{(G)^T} \mathbf{U}^{(G)} = \mathbf{I}. \end{aligned} \tag{3}$$

Laplacian $\mathcal{L}^{(G)} = \mathbf{D}^{(G)^{-\frac{1}{2}}} \mathbf{S}^{(G)} \mathbf{D}^{(G)^{-\frac{1}{2}}}$, and degree matrix $\mathbf{D}^{(G)}$ is a diagonal matrix with sum of each row of $\mathbf{S}^{(G)}$ on the diagonal. A constraint is added to avoid being arbitrary.

$$L(u,v) = \begin{cases} d_v & \text{if } u = v, \\ -1 & \text{if } u \text{ and } v \text{ are adjacent}, \\ 0 & \text{otherwise.} \end{cases}$$

Let T denote the diagonal matrix with the (v,v)-th entry having value d_v. The *Laplacian* of G is defined to be the matrix

$$\mathcal{L}(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } d_v \neq 0, \\ -\frac{1}{\sqrt{d_u d_v}} & \text{if } u \text{ and } v \text{ are adjacent}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} Tr(I-\mathcal{L})^2 &= Tr(T^{-1/2}AT^{-1}AT^{-1/2}) \\ &= \sum_{x,y} \frac{1}{\sqrt{d_x}} A(x,y) \frac{1}{d_y} A(y,x) \frac{1}{\sqrt{d_x}} \\ &= \sum_x \frac{1}{d_x} - \sum_{x \sim y} \left(\frac{1}{d_x} - \frac{1}{d_y}\right)^2, \end{aligned}$$

$$\mathcal{L} = T^{-1/2} L T^{-1/2}$$

Attributed networks Embedding

- the attribute latent representation U(A)

sine similarity to construct the attribute affinity matrix $\mathbf{S}^{(A)}$, and aim at minimizing the degree of disagreement between $\mathbf{U}^{(A)}$ and $\mathbf{S}^{(A)}$. We denote the corresponding Laplacian as $\mathcal{L}^{(A)} = \mathbf{D}^{(A)^{-\frac{1}{2}}}\mathbf{S}^{(A)}\mathbf{D}^{(A)^{-\frac{1}{2}}}$, where $\mathbf{D}^{(A)}$ is the degree matrix of $\mathbf{S}^{(A)}$. Then the objective function of node attributes embedding is defined as

$$\begin{aligned} \underset{\mathbf{U}^{(A)}}{\text{maximize}} \quad & \mathcal{J}_A = \mathrm{Tr}(\mathbf{U}^{(A)^T}\mathcal{L}^{(A)}\mathbf{U}^{(A)}) \\ \text{subject to} \quad & \mathbf{U}^{(A)^T}\mathbf{U}^{(A)} = \mathbf{I}. \end{aligned} \tag{4}$$

Attributed networks Embedding

- incorporate $\mathbf{U}(A)$ into $\mathbf{U}(G)$

  we project $\mathbf{U}(A)$ into the space of $\mathbf{U}(G)$, and employ variance of the projected matrix as a measurement of the correlations

$$\rho_1 = \mathrm{Tr}(\mathbf{U}^{(A)^T}\mathbf{U}^{(G)}\mathbf{U}^{(G)^T}\mathbf{U}^{(A)}). \qquad (5)$$

投影矩阵， $P = A(A^TA)^{-1}A^T$

Label Informed Embedding

In this step, we map the node proximities in labels into a latent representation U(Y) , The basic idea is to employ the learned attribute network proximity to smooth label information modeling.

the same label into the same clique $\mathbf{Y}\mathbf{Y}^T$

the cosine similarity of $\mathbf{Y}\mathbf{Y}^T$ $\mathbf{S}^{(YY)}$

$$\mathcal{L}^{(YY)} = \mathbf{D}^{(Y)^{-\frac{1}{2}}}\mathbf{S}^{(YY)}\mathbf{D}^{(Y)^{-\frac{1}{2}}}$$

$\mathbf{D}^{(Y)}$ is the degree matrix of $\mathbf{S}^{(YY)}$

Label Informed Embedding

However, due to the special structure, the rank of matrix $\mathbf{S}^{(YY)}$ is limited by the number of label categories k, which might be smaller than the embedding dimension d. It leads to unsatisfactory performance of the eigen-decomposition of $\mathcal{L}^{(YY)}$. To address this issue, we utilize the learned proximity $\mathbf{U}^{(G)}\mathbf{U}^{(G)^T}$ to smooth the modeling, and leverage the

$$\begin{aligned} \underset{\mathbf{U}^{(Y)}}{\text{maximize}} \quad & \mathcal{J}_Y = \mathrm{Tr}\left(\mathbf{U}^{(Y)^T}(\mathcal{L}^{(YY)} + \mathbf{U}^{(G)}\mathbf{U}^{(G)^T})\mathbf{U}^{(Y)}\right) \\ \text{subject to} \quad & \mathbf{U}^{(Y)^T}\mathbf{U}^{(Y)} = \mathbf{I}. \end{aligned} \tag{6}$$

Correlation Projections

since all of the latent representations are constrained by corresponding Laplacian, we project all of them into a new space H

$$\rho_2 = \mathrm{Tr}(\mathbf{U}^{(G)^T}\mathbf{H}\mathbf{H}^T\mathbf{U}^{(G)}). \tag{7}$$

Similarly, we project $\mathbf{U}^{(A)}$ and $\mathbf{U}^{(Y)}$ into the space of $\mathbf{H}$ and measure their correlation as

$$\rho_3 = \mathrm{Tr}(\mathbf{U}^{(A)^T}\mathbf{H}\mathbf{H}^T\mathbf{U}^{(A)}), \text{ and} \tag{8}$$

$$\rho_4 = \mathrm{Tr}(\mathbf{U}^{(Y)^T}\mathbf{H}\mathbf{H}^T\mathbf{U}^{(Y)}). \tag{9}$$

The loss function for entire three projections is defined as

$$\underset{\mathbf{U}^{(\cdot)},\mathbf{H}}{\text{maximize}} \quad \mathcal{J}_{corr} = \rho_2 + \rho_3 + \rho_4, \tag{10}$$

Joint Representation Learning via LANE

$$
\begin{aligned}
\underset{\mathbf{U}^{(\cdot)},\mathbf{H}}{\text{maximize}} \quad & \mathcal{J} = (\mathcal{J}_G + \alpha_1 \mathcal{J}_A + \alpha_1 \rho_1) + \alpha_2 \mathcal{J}_Y + \mathcal{J}_{corr} \\
\text{subject to} \quad & \mathbf{U}^{(G)^T}\mathbf{U}^{(G)} = \mathbf{I}, \quad \mathbf{U}^{(A)^T}\mathbf{U}^{(A)} = \mathbf{I}, \\
& \mathbf{U}^{(Y)^T}\mathbf{U}^{(Y)} = \mathbf{I}, \quad \mathbf{H}^T\mathbf{H} = \mathbf{I},
\end{aligned} \tag{11}
$$

Optimization Algorithm for LANE

The second order derivative of $\mathcal{J}$ w.r.t. $\mathbf{U}^{(G)}$ is formed as

$$\nabla^2_{\mathbf{U}^{(G)}}\mathcal{J} = \mathcal{L}^{(G)} + \alpha_1 {\mathbf{U}^{(A)}}^T \mathbf{U}^{(A)} + \alpha_2 {\mathbf{U}^{(Y)}}^T \mathbf{U}^{(Y)} + \mathbf{H}^T\mathbf{H}, \tag{12}$$

When $\mathbf{U}^{(A)}$, $\mathbf{U}^{(Y)}$ and $\mathbf{H}$ are fixed, Eq. (11) becomes convex w.r.t. $\mathbf{U}^{(G)}$, and we are able to obtain the optimal solution via Lagrange multipliers method. Let $\boldsymbol{\lambda}_i (i = 1, \ldots, 4)$

$$(\mathcal{L}^{(G)} + \alpha_1 \mathbf{U}^{(A)} \mathbf{U}^{(A)T} + \alpha_2 \mathbf{U}^{(Y)} \mathbf{U}^{(Y)T} + \mathbf{H}\mathbf{H}^T)\mathbf{U}^{(G)} = \boldsymbol{\lambda}_1 \mathbf{U}^{(G)}, \tag{13}$$

$$(\alpha_1 \mathcal{L}^{(A)} + \alpha_1 \mathbf{U}^{(G)} \mathbf{U}^{(G)T} + \mathbf{H}\mathbf{H}^T)\mathbf{U}^{(A)} = \boldsymbol{\lambda}_2 \mathbf{U}^{(A)}, \tag{14}$$

$$(\alpha_2 \mathcal{L}^{(YY)} + \alpha_2 \mathbf{U}^{(G)} \mathbf{U}^{(G)T} + \mathbf{H}\mathbf{H}^T)\mathbf{U}^{(Y)} = \boldsymbol{\lambda}_3 \mathbf{U}^{(Y)}, \tag{15}$$

$$(\mathbf{U}^{(G)} \mathbf{U}^{(G)T} + \mathbf{U}^{(A)} \mathbf{U}^{(A)T} + \mathbf{U}^{(Y)} \mathbf{U}^{(Y)T})\mathbf{H} = \boldsymbol{\lambda}_4 \mathbf{H}. \tag{16}$$

Complexity Analysis

$\mathcal{O}(n^2)$.

Extensions

LANE w/o Label

$$
\begin{aligned}
\underset{\mathbf{U}^{(G)},\mathbf{U}^{(A)},\mathbf{H}}{\text{maximize}} \quad & \mathcal{J}_G + \beta_1 \mathcal{J}_A + \beta_2 \rho_1 + \rho_2 + \rho_3 \\
\text{subject to} \quad & {\mathbf{U}^{(G)}}^T \mathbf{U}^{(G)} = \mathbf{I}, \quad {\mathbf{U}^{(A)}}^T \mathbf{U}^{(A)} = \mathbf{I}, \\
& \mathbf{H}^T \mathbf{H} = \mathbf{I}.
\end{aligned} \tag{17}
$$

Experiments

Datasets

	# Nodes	# Edges	# Attributes	# Labels
BlogCatalog	5,196	171,743	8,189	6
Flickr	7,575	239,738	12,047	9

Table 2: Detailed information of the datasets.

Baselines

- *LCMF* [47]: It conducts a joint matrix factorization on the linkage and attribute information, and maps them into a shared subspace. It uses this subspace as the learned representation.

- *SpecComb*: It concatenates the attributed network $\mathcal{G}$ and labels $\mathbf{Y}$ into one matrix, and performs normalized spectral embedding [41] on this combined matrix. The corresponding top d eigenvectors are collected as the embedding representation.

- *MultiView* [17]: It considers the network, attributes, and labels as three views, and applies co-regularized spectral clustering on them collectively.

- *LANE_on_Net* and *LANE_w/o_Label*: They are two variations of LANE, which have been described in Section 3.6. The former one is for a plain network. The latter one only leverages the attributed network, without the help of label informed embedding.

Performance Evaluation

1、嵌入表示学习的影响（嵌入向量维度d）5-100

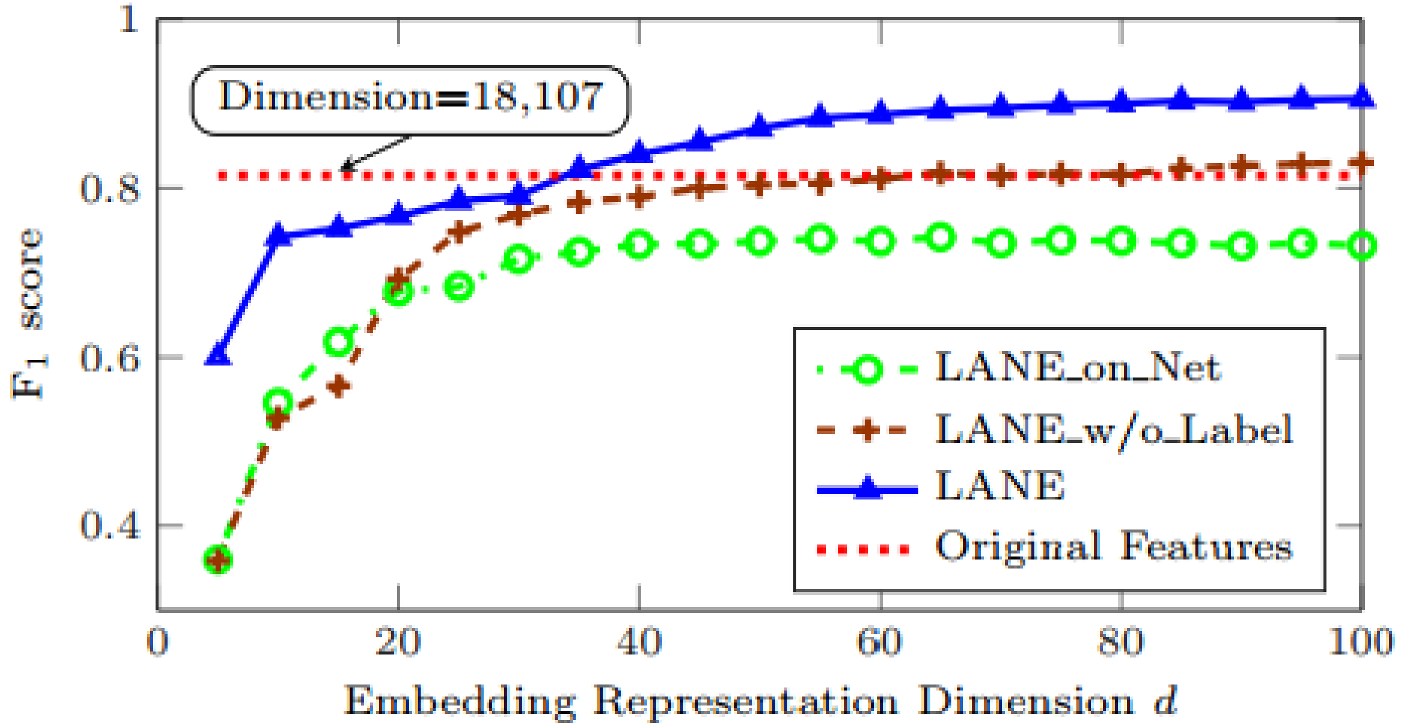


Figure 2: Classification performance of Original Features, LANE and its variations on Flickr dataset.

2、LANE的影响

	BlogCatalog					Flickr				
	1/16	1/8	1/4	1/2	1	1/16	1/8	1/4	1/2	1
DeepWalk	0.5488	0.7000	0.7824	0.7937	0.8100	0.3438	0.4597	0.5818	0.6819	0.7382
LINE	0.6663	0.7255	0.7332	0.6959	0.6931	0.3587	0.4920	0.5733	0.6500	0.6413
LANE_on_Net	0.6553	0.6985	0.7590	0.8046	0.8126	0.4298	0.5063	0.5698	0.6300	0.7319
LCMF	0.7119	0.7920	0.8366	0.8646	0.8401	0.3531	0.5065	0.5884	0.7026	0.7381
LANE_w/o_Label	0.7638	0.7977	0.8361	0.8513	0.8685	0.5426	0.6046	0.6578	0.6809	0.8300
SpecComb	0.6138	0.6027	0.6360	0.6965	0.5895	0.4618	0.4943	0.5800	0.7054	0.7816
MultiView	0.6478	0.7046	0.8207	0.8078	0.7903	0.4942	0.4856	0.5843	0.5870	0.8061
LANE	**0.8065**	**0.8523**	**0.8856**	**0.8964**	**0.9008**	**0.6658**	**0.7645**	**0.8267**	**0.8276**	**0.9054**

Table 3: Classification performance (F_1 score) of different methods on different datasets with $d = 100$.

3、参数敏感度分析

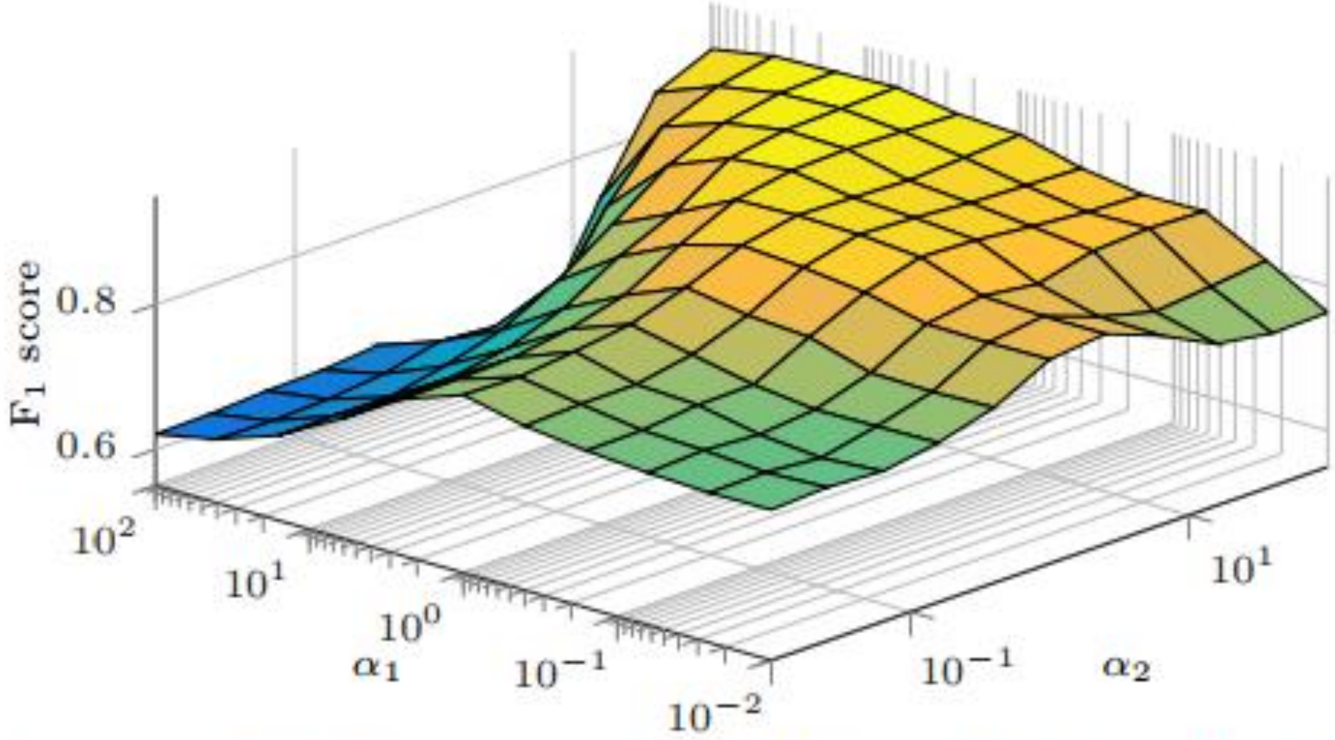


Figure 3: Performance of LANE on Flickr with different parameters α_1 and α_2.

Related References

Laplacian eigenmaps andspectral techniques for embedding and clustering

Spectral graph theory. *American Mathematical Soc.*, 1997

Learning spectral clustering. *NIPS*, 2004

.Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS*, 2001