Graph Spectral Feature Learning for Mixed Data of Categorical and Numerical Type

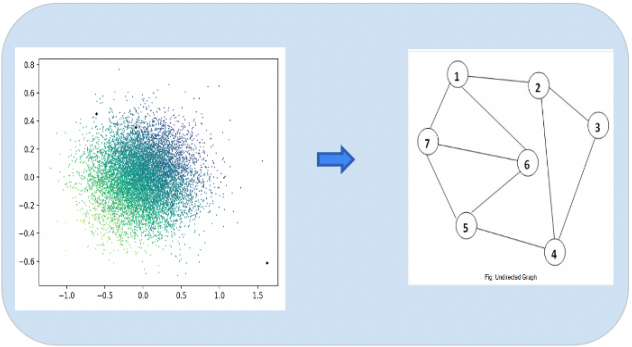
Saswata Sahoo* (Gartner), Souradip Chakraborty* (Walmart Labs)

Overview :
Representation Learning in a heterogeneous space with mixed variables of numerical and categorical types has interesting challenges due to its complex feature manifold. We propose a novel strategy to explicitly model the probabilistic dependence structure among the mixed type of variables by an undirected graph.

Methodology

We assume the components of X as vertices of an undirected graph and the edges of the graph indicate the conditional dependence structure among the random variables. Let's say, the graph *G = (V,E)* with vertices V = {1,2,...,p} and edge set E.

Undirected Graphical Model & Ising Model



$$f(x) \propto exp[\sum_{(s,t)\in\mathcal{E}} \psi(x_s, x_t)]$$

$$\psi(x_s, x_t) = \theta_{st} h(x_s, x_t)$$

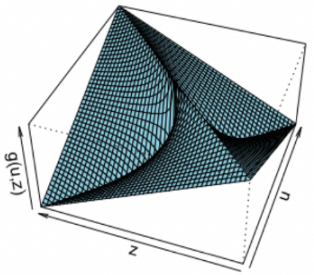
Collective Factorization of Numerical and Categorical Space

$$\min_{W,H_1,H_2} : ||\mathcal{D}^{(num)} - WH_1|| + ||\mathcal{D}^{(cat)} - WH_2||$$

$$\hat{x}^{(num)} = \hat{w}\hat{H}_1$$
$$\hat{x}^{(cat)} = \hat{w}\hat{H}_2$$

Similarity Function for the Edge potential

$$g(u,z) = \frac{\min(|u|,|z|)}{\max(|u|,|z|)} \operatorname{sgn}(uz), \quad if \min(|u|,|z|) > \epsilon,$$
$$= \frac{\epsilon \operatorname{sgn}(uz)}{\max(|u|,|z|)}, \quad if \min(|u|,|z|) \le \epsilon \text{ and } \max(|u|,|z|) > \epsilon,$$
$$= 1, \quad if \max(|u|,|z|) \le \epsilon.$$



Model Estimation and Pseudo Likelihood

$$\log L(\theta, \mathcal{D}) = \sum_{i=1}^{n} \sum_{\forall(s,t)\in\mathcal{E}} \theta_{st} h(x_{(i)s}, x_{(i)t}) - n \log(Z(\theta))$$

$$Z(\theta) = \sum_{i=1}^{n} \exp(\textstyle\sum_{\forall(s,t)\in\mathcal{E}} \theta_{st} h(x_{(i)s}, x_{(i)t}))$$

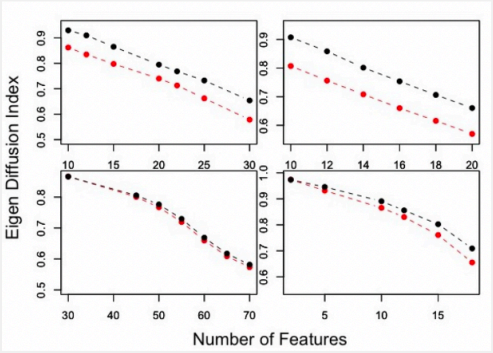
$$\hat{\theta}_{st} = \theta_{st} - \eta \frac{\partial \log L(\theta, \mathcal{D})}{\partial \theta_{st}}$$

Graph Spectral Feature Learning for Mixed Data of Categorical and Numerical Type

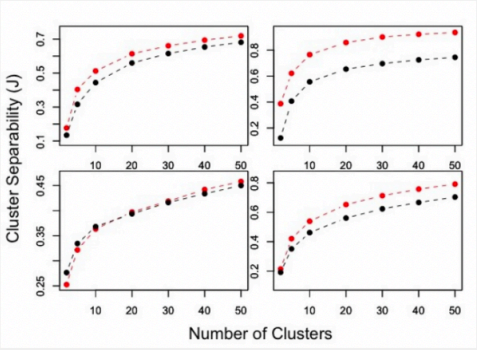
Saswata Sahoo* (Gartner), Souradip Chakraborty* (Walmart Labs)

Results & Numerical Investigation

Eigen Diffusion index (α) for varying feature dimensions of the proposed spectral embedding feature map (in red) and naive principal component features (in black)



Cluster Separability index (J) for varying number of clusters with naive K-means on the proposed spectral embedding features (in red) and actual mixed data points (in black):



Cluster quality obtained using different clustering techniques. Clusters based on the proposed feature map outperforming all the competitors are shown in bold font

	Datasets	Credit Approval Dataset		Adult Salary Dataset		German Credit Dataset		Heart Disease Dataset	
	Methods	RI	NMI	RI	NMI	RI	NMI	RI	NMI
Clustering on Proposed Features	SE-KMeans	0.625	0.257	**0.639**	**0.227**	**0.599**	**0.028**	**0.669**	**0.271**
	SE-KMediods	**0.699**	**0.328**	0.582	**0.208**	0.562	**0.036**	**0.674**	**0.323**
	SE-HAgglo	**0.672**	**0.306**	**0.639**	**0.226**	**0.678**	**0.017**	0.629	0.245
Competitor Clustering Methods for Mixed Data	K-Mediods	0.670	0.281	0.550	0.126	0.537	0.008	0.669	0.260
	K-Prototype	0.571	0.172	0.540	0.162	0.587	0.001	0.669	0.262
	WKM	0.662	0.281	0.603	0.092	0.504	0.003	0.611	0.188
	EWKM	0.654	0.243	0.626	0.012	0.511	0.013	0.643	0.221
	OCIL	0.601	0.182	0.624	0.004	0.559	0.003	0.641	0.231
	AE-KMeans	0.580	0.142	0.581	0.006	0.552	0.019	0.584	0.143