Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Non-linear Dimensionality Reduction of Social Network Content using Cross-Covariance Based Local Tangent Space Alignment

[1]Sunil Kumar Mishra and [2] Ajay Agarwal

[1]Department of CSE, Accurate Institute of Management and Technology, Greater Noida, India-201308

[2]Department of IT, KIET Group of Institutions, Ghaziabad, India-201206

E-mail: [1]sunilmishra.accurate@gmail.com, [2]ajay.agarwal@kiet.edu

ABSTRACT

The rise in the involvement of people on social media has exponentially increased the volume of the data. Along with the publicly available information of the individuals, such as name, date of birth, address, phone number, etc., the data intrinsically hides a vast information about individual usage pattern such as likes, dislikes, shopping history, browsed products, and friends' network. This information has been proved to be useful in targeting customers with specific products, services, or offers to increase revenue. Generally, the domain experts extract this information through manual feature engineering to build a classification or regression model, but on the given high dimensional data with high feature correlation and noise, the task becomes complicated. The existing state-of-the-art non-linear dimensionality reduction methods solve the problem by preserving the maximum non-linear relationship among the data while projecting it to its intrinsic dimensional space. Local tangent space alignment (LTSA) also aims to give true intrinsic representation by aligning the tangent spaces of spatially close data points using covariance metric. In this paper, we propose a cross-covariance based local tangent space alignment (CcLTSA) method to preserve the maximum non-linear relationship among data. The proposed method preserves both local and global information by exploiting the statistical independence between the locally connected instances using the Hilbert-Schmidt independence criterion. Extensive experiments on synthetic data set shows that the proposed CcLTSA gives better intrinsic space representation. The classification results on real-world data set proves that CcLTSA outperforms the existing state-of-the-art methods by≈3%.

Keywords: *Social Network Mining, Dimensionality reduction, Manifold Learning, Local Tangent Space Alignment, Hilbert-Schmidt Independence Criterion*

1. INTRODUCTION

The recent increase in social media engagement has changed the whole scenario of data mining [1][2][3]. Just like in the past, miners extract precious metals out from mines; the data miners also try to extract meaningful information and insights about an individual or group through data mining. This information generally hides in plain sight of the data and can be extracted only by employing sophisticated machine learning (ML) tools [4][5]. ML on social media content can give a good overview of individual usage pattern and his/her interactions with others [6][7][8]. Further, this can be used for multiple purposes such as showing relevant contents similar to their interest, suggest already available social groups of similar choice [9], target advertisement [10], etc. The same set of information is also used by companies to understand the requirements of the users, poll, and opinion about their products, market survey for products to be launched, and so on. On the darker side of social media, the recent increase in online trolling and bullying [11][12] has made a negative impact on its users. ML algorithms can also help in curbing these activities by an accurate sentiment analysis or content filtering.

Though the big data captured through social media is useful, it becomes a great deal to extract the required information from it. Previously, this was handled by a domain expert who through manual feature engineering performed feature extraction, selection, clustering, etc. However, with the given high dimensional data, manual efforts fall short of achieving the target with an increase in cost, time, and result in erroneous inferences. In such situations, dimensionality reduction using ML comes handy [13][14]. The ML models inherently analyse the statistical and geometrical properties of data to find their most optimal representation using a minimum number of dimensions containing sufficient discriminative power [15].

It is well known that not all the dimensions constitute in information extraction. Some dimensions are strongly co-related to others, and few of them are present due to unknown transformations and noise in the data. If, the data contains a linear relationship among all the instances, the traditional algorithms like principal component analysis (PCA) [16][17], multidimensional scaling (MDS) [18], etc. can be applied. As not all data follow linear properties, it is also essential to preserve the non-linear relationship present in the data even after dimensionality reduction has been applied. Manifold learning [19][20] helps in achieving this goal through a set of non-linear dimensionality reduction methods. In the core, manifold learning assumes that the given high dimensional data lie on a very low dimensional space where all instances follow Euclidean properties. On this low dimensional space, data visualization, clustering, classification, and regression can be performed much easily and accurately than on the given high dimensional space [21]. A few manifold learning methods include: local tangent space alignment (LTSA) [22], isometric mapping (ISOMAP) [23], local linear embedding (LLE) [24], Laplacian eigenmaps (LE) [25] and Hessian eigenmaps (HE) [26]. The idea is to exploit the local linear geometrical properties present in the data and perform a dimensionality reduction by preserving those properties. The better the properties preserved, the more accurate is the low dimensional representation.

Assume from a smooth Riemannian manifold $\mathcal{M} \in \mathbb{R}^D$, data samples $x_i \in X$ have been drawn. In LTSA, the aim is to align the local tangent space around each x_i to determine the true connectivity between all manifold samples. To accomplish the task, LTSA assumes the following:

1. The samples drawn follow the uniform distribution and cover the whole sample space.
2. The spatial relationship among the instances should be preserved i.e., instances close on high dimension should lie close on the low dimensional space.

However, the randomly drawn samples fail to assure the first assumption, and due to unknown transformations and noise, the spatial relationship between instances cannot be used alone to measure the similarity between them. Due to highly correlated and the noise, the data generally lie on the periphery of the space, which leads to erroneous inferences. In manifold learning, it is well known that the actual data resides on a much lower dimension having sufficient discriminative power. In the core, the manifold learning method heavily relies on the measure of similarity between the instances. The distance or coordinate-based similarity such as Euclidean, Cosine, etc. fails to deliver accurate metric due to noise present in the data. Similarly, covariance used in LTSA also gets affected by similar factors leading to erroneous results. The problem lies in the covariance assumption that all data observations lie on the same space; however, due to varying curvature of the underlying manifold structure, this assumption does not hold. The concept of classification using dimensionality reduction as a pre-processing step is depicted in fig. 1.

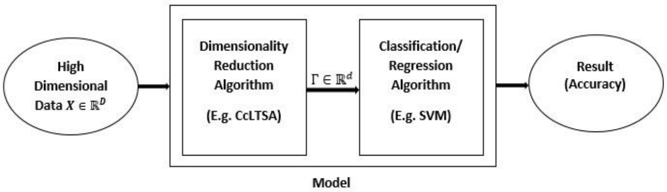


Figure 1: Concept of classification with Dimensionality Reduction

In this paper, we propose cross-covariance based LTSA (CcLTSA) using Hilbert-Schmidt independence criterion (HSIC) to measure statistical dependence/independence between the neighboring instances. CcLTSA aims to preserve this statistical information along with the given spatial relation on the lower-dimensional space to obtain the true intrinsic geometrical co-ordinates of the given data.

The rest of the paper has been organized by describing the problem definition in section 2, followed by the proposed method, and its preliminaries in section 3. Extensive experiments using CcLTSA and existing state-of-the-art dimensionality reduction methods have been listed and compared in section 4. The paper concludes with the findings of the proposed work in section 5.

2. PROBLEM DEFINITION

Given D dimensional input data $(x_i)_{i=1}^n \in \mathbb{R}^D$ captured from social network modelled using a graph $G = (V, E)$ where, each instance becomes a vertex $x_i \in V$ and the link between two instance x_i and x_j is represented using a weighted edge $e_{ij} \in E$. As D is very large and due to the strong co-relation between them, most of the data lie on the periphery of the input space. Thus, the underlying ML model fails to extract accurate information from it. However, the true discriminative complement information lies on a much lower dimension $d \ll D$. The aim is to exploit the intrinsic information present in x_i to obtain its optimal representation on $\mathbb{R}^d$ with maximum property preservation.

In its core, manifold learning algorithm assumes that if two instances are close on given input space then they should also lie close on lower-dimensional space. This is ensured by measuring the similarity between such instances generally through some distance function. However, when the original data is affected by noise, the distance automatically becomes noisy, and hence, the resultant lower-dimensional representation becomes erroneous. The aim is to discard the effects of noise and extract the true local geometrical data representation which contains the highest discriminative information through cross-covariance based LTSA.

3. PROPOSED METHODOLOGY

3.1 Hilbert-Schmidt independence criterion

The HSIC is defined by Hilbert-Schmidt operators [27].

Definition 1: Let H_x and H_y be two separate Hilbert spaces where, the orthonormal basis of H_x is given by $u_i \in I$. $\mathcal{F}: H_x \to H_y$ is a compact operator with $\sum_{i \in I} \parallel \mathcal{F} e_i \parallel_y^2 < +\infty$ then, $\mathcal{F}$ is a Hilbert-Schmidt (HS) operator [28].

Let, $HS(H_x \to H_y)$be the space of all HS operators from H_x to H_y . Then the Hilbert space is obtained from

$(HS(H_x \to H_y), \langle \cdot,\cdot \rangle_{HS})$

where, $\langle \cdot,\cdot \rangle_{HS}$ defines the inner product. HSIC involves two reproducing kernel Hilbert space (RKHS) to measure the independence between them.

Let, $H_x = (L^2(\Omega_x), \langle \cdot,\cdot \rangle_x)$ be a RKHS where, $\kappa_x: \Omega_x \times \Omega_x \to \mathbb{R}$ is the reproducing kernel of H_x. Define $\varphi_x: \Omega_x \to H_x$ such that for all $x_i \in \Omega_x$, $\varphi_x(x_i) = \kappa_x(\cdot, x_i) \in H_x$ and $\langle \varphi_x(x_i'), \varphi_x(x_i'') \rangle_x = \kappa_x(x_i', x_i'')$. Similarly, for Ω_y on H_y.

Theorem 1: Let, $\Phi: HS(H_x \to H_y) \to \mathbb{R}$ such that for all $\mathcal{F} \in HS(H_x \to H_y)$ then,

$$\Phi(\mathcal{F}) = \mathbf{E}_{xy}[\langle \varphi_x(x) \otimes \varphi_y(y), \mathcal{F} \rangle_{HS}]$$

If, $\mathbf{E}_{xy}\left[\| \varphi_x(x) \otimes \varphi_y(y) \|_{HS}\right] < +\infty$, then Φ is continuous linear function on $HS(H_x \to H_y)$. Based on representation theorem (Riesz theorem) [29] of continuous linear functions, there must be a unique operator $\mathcal{F}_\Phi \in HS\left(H_x \to H_y\right)$ such that for all HS operators $\mathcal{M} \in HS(H_x \to H_y)$,

$$\Phi(\mathcal{F}) = \mathbf{E}_{xy}[\langle \varphi_x(x) \otimes \varphi_y(y), \mathcal{F} \rangle_{HS}] = \langle \mathcal{F}, \mathcal{F}_\Phi \rangle_{HS}$$

where, $\mathcal{F}_\Phi$ is the cross-covariance operator represented using C_{xy}.

Definition 2: Given two RKHS spaces H_x and H_y with joint measure p_{xy}, the HSIC is defined as the squared HS-norm of C_{xy}

$$HSIC(p_{xy}, H_x, H_y) := \| C_{xy} \|_{HS}^2$$

Lemma 1:

$$HSIC(p_{xy}, H_x, H_y) = \mathbf{E}_{x,x',y,y'}[\kappa_x(x, x')\kappa_y(y, y')] + \mathbf{E}_{x,x'}[\kappa_x(x, x')]\mathbf{E}_{y,y'}[\kappa_y(y, y')] - 2\mathbf{E}_{x,y}[\mathbf{E}_{x'}[\kappa_x(x, x')]\mathbf{E}_{y'}[\kappa_y(y, y')]]$$

Here, $\mathbf{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x, y) and (x', y') drawn from p_{xy}.

Definition 3: Let, $Z := \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\} \subseteq X \times Y$ be set of n independent observations drawn from p_{xy}. The empirical HSIC can be estimated using

$$HSIC(Z, H_x, H_y) := \frac{1}{(n-1)^2} tr(KHLH) \tag{1}$$

where, $K, H, L \in \mathbb{R}^{n \times n}$, $K_{ij} = \kappa_x\left(x_i, x_j\right)$, $L_{ij} = \kappa\left(y_i, y_j\right)$and $H_{ij} = I - \frac{\delta_{ij}}{n}$ is the centering matrix.

3.2 CROSS-COVARIANCE BASED LOCAL TANGENT SPACE ALIGNMENT (CcLTSA)

The given n data samples from a smooth Riemannian manifold $\mathcal{M}$ can be represented as $(x_i)_{i=1}^n \in \mathbb{R}^D$. It is well known that the data actually lies on a much lower dimension space $d \ll D$ and can be represented by

$$f : C \subset \mathbb{R}^d \to \mathbb{R}^D$$

where, C is a compact subset of $\mathbb{R}^d$ and f is the data generation function i.e.

$$x_i = f(\tau_i) + \eta_i$$

where, τ_i are original feature vectors or the lower dimensional complement information and η_i is the noise. A linear manifold learning can be solved by minimizing the reconstruction error

$$min \| E \| = argmin_{c,U,T} \| X - (ce^T + UT) \|_F$$

where, $\|\cdot\|_F$ defines the Frobenius norm on the matrix, $X = [x_1, x_2 \dots x_n]$, $T = [\tau_1, \tau_2 \dots \tau_n]$, $E = [\eta_1, \eta_2 \dots \eta_n]$, $e \in \mathbb{1}^n$ column vector, c is the centering matrix and Ucontains orthonormal basis of intrinsic subspace. The co-relation matrix is obtained using

$$\hat{X} = (X - \overline{X}e^T)(X - \overline{X}e^T)^T \tag{2}$$

where, $\overline{x}_i \in \overline{X} = \frac{1}{n}\sum_{j=1}^n x_j$ is the centered matrix. Then, the problem can solved by performing singular value decomposition (SVD) on $\hat{X}$

$$SVD(\hat{X}) = \mathbf{Q}\Sigma\mathbf{V}^T \tag{3}$$

where, both $\mathbf{Q} \in \mathbb{R}^{D \times D}$ and $\mathbf{V} \in \mathbb{R}^{n \times d}$ are orthonormal matrices, and $\Sigma \in \mathbb{R}^{D \times n}$ is the singular value matrix. The measure of similarity in the local region completely relies on result from eqn. 2. However, as it depends on the given coordinate system which is affected by noise and co-related dimensions, the similarity cannot be assured to be accurate. By replacing the co-relation matrix with cross-covariance matrix using HSIC, the statistical similarity between instances are obtained in Hilbert space which minimizes the effect of noise and gives accurate result. Briefly, a cross-covariance operator maps from one space to another, whereas a covariance operator maps from a space to itself. In the linear algebraic case, the covariance is $C_{xx} := \mathbf{E}_x[xx^T] - \mathbf{E}_x[x]\mathbf{E}_x[x^T]$, while the cross-covariance is $C_{xy} := \mathbf{E}_{x,y}[xy^T] - \mathbf{E}_x[x]\mathbf{E}_y[y^T]$. On replacing covariance in eqn. 2 to cross-covariance from eqn. 1, the similarity within a neighborhood is measured using

$$\tilde{X} = X - \overline{X}e^T \; ; K = \kappa_x(\tilde{X}, \tilde{X}^T) \; ; H = I - \frac{\mathbb{1}}{n}$$

$$\hat{X} = HKH \tag{4}$$

where, κ must be a positive definite kernel.

$$\Rightarrow \Phi = SVD\left(\hat{X}\right) = \mathbf{Q}_d\Sigma_d\mathbf{V}_d^T$$

The optimal Φ^* is given by eigenvectors $\mathbf{Q}_d$ corresponding to d largest singular values. Then, the linear manifold can be represented as

$$f(\tau) = \overline{X} + \Phi^*\tau$$

and the coordinate matrix Γ is given by

$$\Gamma = (\Phi^*)^T\hat{X} = \mathrm{diag}(\sigma_1 \dots \sigma_d)\mathbf{V}_d^T$$

Γ contains the low dimensional representation of the given data. In case of non-linear dimensionality reduction, it is required to explore and exploit local linear region around each observation. The local linear structure can be extracted by representing each sample x_i with weighted linear sum of its neighbors N_i.

$$N_i = \left[x_{i,1}, x_{i,2} \dots x_{i,k}\right]$$

In its core LTSA works on the assumption that if the manifold is correctly unfolded then all tangent spaces will be aligned. The d dimensional sub-space for each N_i can be approximated by

$$argmin_{x,\theta,\mathbf{Q}} \sum_{j=1}^k \| x_{ij} - (x + \mathbf{Q}\theta_j) \|_2^2 = argmin_{x,\Theta,\mathbf{Q}} \| N_i - (xe^T + \mathbf{Q}\Theta) \|_2^2 \tag{5}$$

where, $\mathbf{Q}$ consists of d orthonormal columns and $\Theta = [\theta_1 \dots \theta_k]$. It holds the local tangent coordinates of neighborhood data points. Further these local tangent coordinates will be aligned in lower dimensional space using affine transformations to obtain global coordinate system.

Algorithm 1: CcLTSA

```
Input: $X \in \mathbb{R}^D$: Data
       $nn$: Nearest neighbor count
Output: $\Gamma \in \mathbb{R}^d$: Low dimensional representation
for $i \leftarrow 1$ to $n$ do
    $N_i \leftarrow find\_knn(x_i, nn)$;
    $\overline{x}_i \leftarrow \frac{1}{n} n \sum_{j=1}^{nn} x_{N_{i,j}}$;
    $\tilde{x}_i = N_i - \overline{x}_i e^T$ ;              \\Centered data
    $K = \kappa_x(\tilde{x}_i, \tilde{x}_i^T)$;          \\Positive definite kernel
    $H = I - \frac{\mathbb{1}}{n}$;                   \\Centering matrix
    $t \leftarrow HKH$ ;              \\HSIC cross-covariance
    $[g_1 \dots g_d] \leftarrow pca(t,d)$;        \\Local coordinates
    $G_i \leftarrow [e/\sqrt{k}, g_1 \dots g_d]$;
    $B(I_i, I_i) \leftarrow B(I_i, I_i) + I - G_i G_i^T$;
end
\* Get global coordinates                    *\
$\Gamma \leftarrow svd(B, d, 'small')$;
$return\ \Gamma$;
```

Then,

$$\Theta_i = \mathbf{Q}_i^T N_i (I - \frac{1}{k} ee^T) = \left[\theta_{i,1} \dots \theta_{i,k}\right]$$

$$\theta_{i,j} = \mathbf{Q}_i^T \left(x_{i,j} - \overline{x}_i\right)$$

$$\therefore x_{i,j} = \overline{x}_i + \mathbf{Q}_i \theta_{i,j} + \xi_{i,j}$$

$$\xi_{i,j} = \left(I - \mathbf{Q}_i \mathbf{Q}_i^T\right)\left(x_{i,j} - \overline{x}_i\right)$$

where, $\xi_{i,j}$ is the tangent reconstruction error. The global coordinate $\{\tau_i\}_{i=1}^n$ is constructed using the local coordinates $\theta_{i,j}$ where each $\tau_{i,j}$ should fulfill

$$\tau_{i,j} = \overline{\tau}_i + L_i \theta_{i,j} + \epsilon_{i,j}$$

for $i = 1 \dots n$ and $j = 1 \dots k$ defines each x_i's local neighborhood.

$$\Rightarrow \Gamma_i = \tfrac{1}{k} \Gamma_i ee^T + L_i \Theta_i + E_i \qquad (6)$$

where, $\Gamma_i = \left[\tau_{i,1} \dots \tau_{i,k}\right]$ and $E_i = \left[\epsilon_{i,1} \dots \epsilon_{i,k}\right]$ is the local reconstruction error.

$$E_i = \Gamma_i \left(I - \tfrac{1}{k} ee^T - L_i \Theta_i\right) \qquad (7)$$

The optimal L_i for a fixed Γ_i is given by

$$L_i = \Gamma_i (I - \frac{1}{k} ee^T) \Theta_i^+ = \Gamma_i \Theta_i^+$$

$$\therefore E_i = \Gamma_i w_i \text{ where, } w_i = \left(I - \frac{1}{k} ee^T\right)\left(I - \Theta_i^+ \Theta_i\right)$$

where, Θ_i^+ represents pseudo inverse of Θ_i.

$$\Rightarrow \sum_{i=1}^{n} \| E_i \|_F^2 = \| TSW \|_F^2$$

where, $S = [s_1 \dots s_n]$ is selection matrix such that $\Gamma s_i = \tau_i$ and $W = \text{diag}(w_1 \dots w_n)$. Constraint $\Gamma\Gamma^T = I_d$ helps finding unique Γ. Then, the vector e becomes the eigenvector of $B = SWW^T S^T$ with respect to eigenvalue zero hence, optimal Γ is given by the eigenvectors corresponding to 2nd to $\mathrm{d} + 1$ smallest eigenvalues of B.

4. EXPERIMENTS AND RESULTS

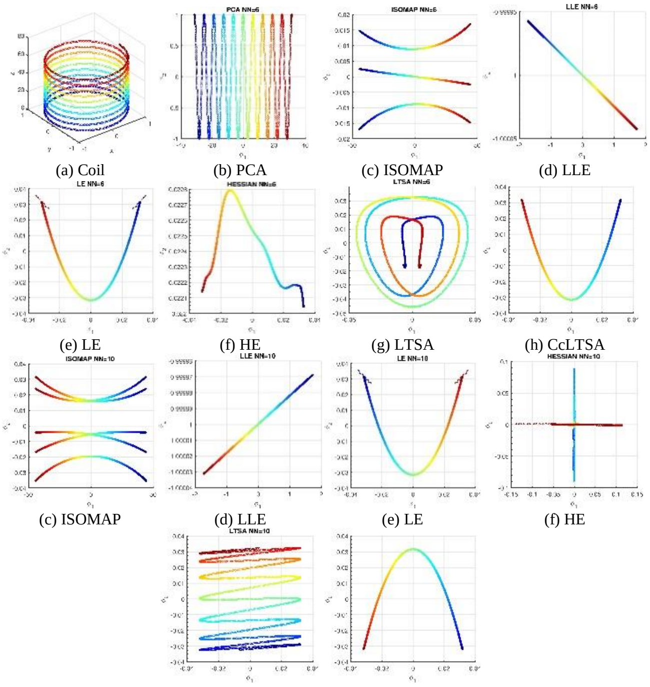
The effectiveness of the proposed CcLTSA based non-linear dimensionality reduction technique has been tested by performing extensive experiments on seven synthetic and four real-world data set. The performance of CcLTSA has been compared with existing state-of-the-art dimensionality reduction methods: PCA, ISOMAP, LLE, LE, HE, and LTSA. The synthetic data set contains high dimensional data for which their respective intrinsic dimension and geometry is well known. In the real-world data set, two linear classifiers i.e. kNN and SVM [30] have been trained assuming that the intrinsic dimensional representation obtained using the above-mentioned methods contains linear decision boundary.

4.1 Synthetic Datasets

Coil: The original data is shown in fig. 2(a). In its intrinsic dimension, it is actually a straight line embedded as a coilin 3D space. In first set of results shown from fig. 2(c) to 2(h), the nearest neighborhood graph for the methods were created by taking 6 nearest neighbors (NN) from the point of interest (PCA does not require a graph). As evident, on $NN = 6$, LLE and CcLTSA were able to give optimal intrinsic geometry representation while ISOMAP struggled and created 3 segments of the same line. PCA, HE and LTSA were able to extract the line but failed to straighten the curvature which leads to sub-optimal representation. LE gave representation similar to CcLTSA but failed to hold the tail instances to make them appear like outliers. The same experiment was performed again for graph based algorithms with $NN = 10$ as shown in fig. 2(i) to 2(n). By increasing the NN, the connectivity in the graph was artificially increased. ISOMAP, LTSA and HE got adversely affected, and the representation became poor than $NN = 6$. LLE and CcLTSA remained unaffected which proves the robustness of the proposed method. The representation of LE got improved, and the tail instances appeared more connected than $NN = 6$.

Sine cylinder: The data generated in sine cylinder is a sine wave with head and tail connected around a cylinder. Intrinsically it is a circle embedded in a 3D space, as shown in fig. 3(a). Fig. 3(b) to 3(h) show the low dimensional representation obtained using various dimensionality reduction method with $NN = 5$.

The simple geometrical shape of the data allowed PCA, ISOMAP, and CcLTSA to extract the accurate intrinsic representation of sine cylinder. An irregular ellipse was recovered through LLE, which shows the incompetence of the method on simple geometric shapes. Other methods LE, HE and LTSA either failed to preserve the connectivity of the data points or remain unable to extract the smooth circle from the data. This proves that proposed CcLTSA is effective even for simple data set.



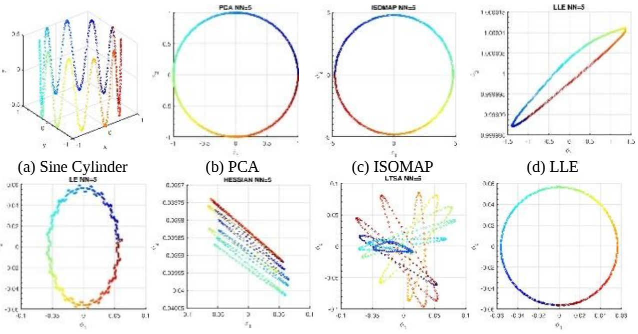
(a) Coil (b) PCA (c) ISOMAP (d) LLE

(e) LE (f) HE (g) LTSA (h) CcLTSA

(c) ISOMAP (d) LLE (e) LE (f) HE

(g) LTSA (h) CcLTSA

Figure 2: Dimensionality reduction on coil

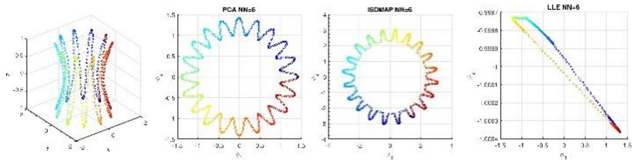


(a) Sine Cylinder (b) PCA (c) ISOMAP (d) LLE

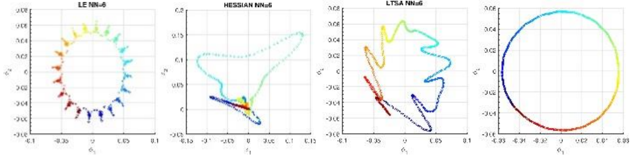
(e) LE (f) HE (g) LTSA (h) CcLTSA

Figure 3: Dimensionality reduction on Sine cylinder

Sine hyperboloid: The sine hyperboloid data is similar to sine cylinder except for extra curvature introduced in the middle of the shape to make it wrap around a hyperboloid as shown in fig. 4(a). Among all the methods with $NN = 6$, CcLTSA was the only method which was able to extract the true intrinsic geometry of a smooth circle from given data, as shown in fig. 4(h). LE (fig. 4(e)) failed to preserve the intermediate data connectivity, which shows its inefficiency. All other methods were able to preserve the data connectivity; however, they failed to extract the smooth circle representation from the data. LLE (fig. 4(d)), HE (fig. 4(f)) and LTSA (fig. 4(g)) extracted an irregular shape which destroyed the spatial information available in data. PCA (fig. 4(b)) and ISOMAP (fig. 4(c)) gave similar results which proved to be sub-optimal.



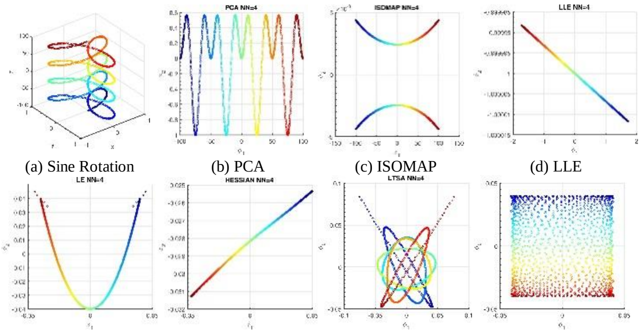
(a) Sine Hyperboloid (b) PCA (c) ISOMAP (d) LLE



(e) LE (f) HE (g) LTSA (h) CcLTSA

Figure 4: Dimensionality reduction on Sine hyperboloid

Sine rotation: The sine wave rotation data consist of a simple 2D sine wave twisted in the third dimension. Originally it is a straight line embedded in 3D space as shown in fig. 5(a). Fig. 5(b) to 5(h) show the intrinsic representation extracted using above mentioned dimensionality reduction techniques using $NN = 6$.



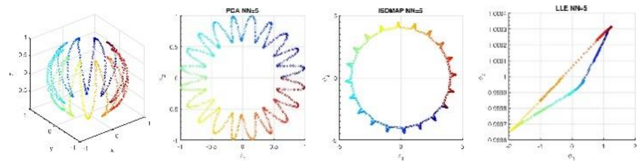
(a) Sine Rotation (b) PCA (c) ISOMAP (d) LLE

(e) LE (f) HE (g) LTSA (h) CcLTSA

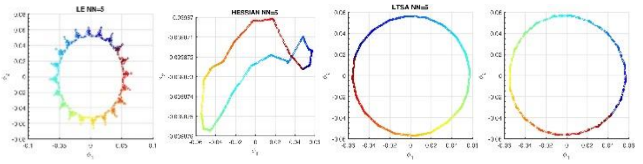
Figure 5: Dimensionality reduction on Sine Rotation

In this experiment, LLE (fig. 5(d)) and HE (fig. 5(f)) gave the most accurate low dimensional representation. Both ISOMAP (fig. 5(c)) and LE (fig. 5(e)) lost the spatial information present in the data by projecting disjoint instances section. PCA (fig. 5(b)) and LTSA (fig 5(g)) preserved the data connectivity information but they were not able to straighten the twists in the original data. The proposed method CcLTSA exposed the additional connectivity hidden in the data along with preserving the original spatial information, as shown in fig. 5(h). The additional connectivity could prove useful in further action of clustering, classification, or regression.

Sine sphere: The sine sphere data is a smooth 2D circle modelled using a sine wave wrapped around a sphere in a 3D space, as shown in fig. 6(a). Among all the methods, only LTSA and CcLTSA were able to extract the circle from original data as depicted in fig. 6(g) and fig. 6(h) respectively. Rest all the methods were able to preserve the data connectivity information at the cost of sub-optimal low dimensional representation. This proves that the CcLTSA leverages basic LTSA properties whenever required to exploit the accurate intrinsic geometrical information.



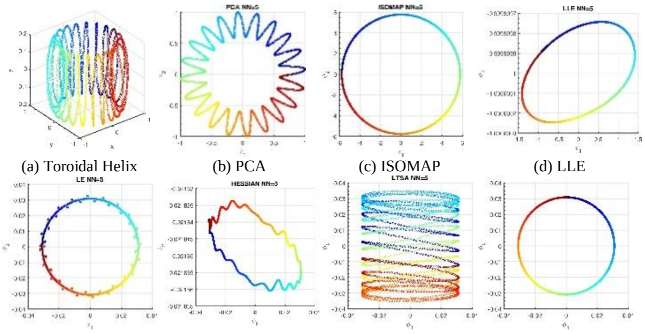
(a) Sine Sphere (b) PCA (c) ISOMAP (d) LLE



(e) LE (f) HE (g) LTSA (h) CcLTSA

Figure 6: Dimensionality reduction on Sine sphere

Toroidal helix: Fig. 7(a) shows the toroidal helix data, which is originally a smooth 2D circle embedded in a coil structure in 3D space. In this experiment using $NN = 5$, all the methods preserved the local data connectivity information, as shown in fig. 7(b) to 7(h) but only ISOMAP (fig. 7(c)) and CcLTSA (fig. 7(h)) were able to produce the accurate low dimensional representation of the toroidal helix. LLE gave a sub-optimal representation using an ellipse as shown in fig. 7(d). Rest of all the methods proved incompetent, as they were not able to extract the intrinsic 2D circle hidden in the data. This proves that the cross-covariance enables CcLTSA to exploit the global connectivity of the graph, along with preserving the local information.

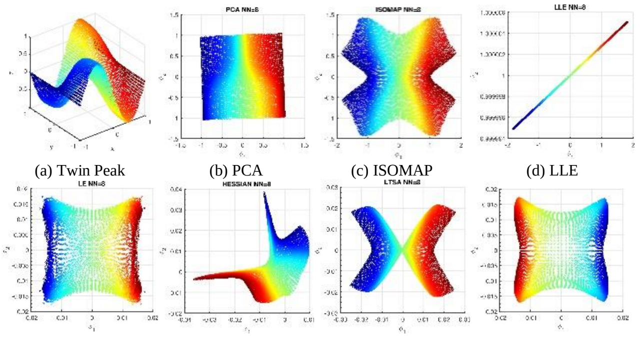


(a) Toroidal Helix (b) PCA (c) ISOMAP (d) LLE

(e) LE (f) HE (g) LTSA (h) CcLTSA

Figure 7: Dimensionality reduction on Toroidal Helix

Twin peaks: The twin peaks data is originally a 2D flat strip with peaks at the two corners, as shown in fig. 8(a). Among all the methods, PCA (fig. 8(b)) gave the most accurate low dimensional representation.



(a) Twin Peak (b) PCA (c) ISOMAP (d) LLE

(e) LE (f) HE (g) LTSA (h) CcLTSA

Figure 8: Dimensionality reduction on Twin Peak

After PCA, LE (fig. 8(e)) and CcLTSA (fig. 8(h)) gave the best representation as compared to other methods using $NN = 8$. LLE (fig. 8(d)), HE (fig. 8(f)) and LTSA (fig. 8(g)) gave grossly inaccurate low dimensional representation as they either discarded the local information, which changed the strip to a line or proved incapable of exploiting global connectivity information leaving the peaks and valleys as it is. This shows that CcLTSA holds properties similar to LE which can be used to exploit the data whenever required.

4.2 Real World Datasets

Four real-world data include Facebook metrics, fashion MNIST, Lego bricks, and mobile pricing have been evaluated using kNN and SVM linear classifiers. The data exhibits the dynamics of the content generated and consumed over various social networks. The feature co-relation in the input data adversely affects the underlying ML model. Hence, in the first step, the input data is reduced to its intrinsic dimension, which contains enough discriminative power useful for classification. In the next step, kNN and SVM classifiers are trained on the reduced input space for the given labels. This process is repeated 10 times by randomly selecting training and testing instances in each round. The accuracy of the dimensionality reduction techniques employed is measured by the mean classification error of both kNN and SVM across the 10 rounds. Other than PCA, all methods require a graph to proceed. In these experiments, the graph is created through the nearest neighbor method. To further to observe the change in low dimension representation from graph-based methods, the value of NN varies from 7 to 17.

Facebook metrics: The Facebook metric data [1] contains textual posts, images, and promotional videos, etc., of a leading cosmetic brand in 2014 on Facebook. The aim of the study is to predict the impact of these online posts. The data consists of 500 observations spanned over 18 attributes. Out of them, first 7 features constitute the input space containing data/time of post, unique post identification, the content of the post, post type/category, and paid post. The rest of the 11 features contain the respective posts' impact in terms of lifetime post total reach, impression, user engagement, post-consumption, likes, comments, share, etc.

The data is reduced from 7 dimensions to one dimension for an optimal input space representation using CcLTSA and other 6 methods. The resultant intrinsic dimensional representation is further modelled using kNN and SVM classifier trained using 300 instances, the rest 200 instances were used for the testing purpose. Table I lists the mean error of all 10 rounds for NN=7-17. As PCA remains independent of NN, the result remained constant for all NN values.

As evident, the proposed CcLTSA remained more accurate than other methods for NN= 13 and 17. The vanilla LTSA gave high accuracy result for NN=7 using kNN classifier. As ISOMAP creates a fully connected graph using Dijkstra or other shortest path method, for NN=7-11, it failed to create a single connected graph and hence, the model was not trained for those values. For NN=11, HE gave most accurate results as compared to others on both kNN and SVM classifiers. Overall for kNN classifier, CcLTSA gave most accurate results followed by PCA, LTSA, LE, HE, LLE and ISOMAP in the same order based on the mean of all NN error values.

Table I
Facebook Metrics (d=1) Mean Error

Method	Classifier	NN=7	NN=11	NN=13	NN=17
CcLTSA	SVM	25.19	24.71	25.46	**25.07**
	kNN	28.85	28.86	**27.75**	29.04
PCA	SVM	**25.16**	25.16	**25.16**	25.16
	kNN	28.42	28.42	28.42	28.42
LLE	SVM	25.71	25.75	26.35	26.03
	kNN	28.23	28.05	29.38	28.65
LTSA	SVM	25.75	26.35	25.34	25.66
	kNN	**27.41**	28.28	28.09	29.61
LE	SVM	25.75	25.75	25.75	25.71
	kNN	27.64	28.88	30.71	**27.73**
HE	SVM	25.89	**24.56**	25.71	25.71
	kNN	29.47	**27.82**	29.24	30.16
ISOMAP	SVM	∞		25.74	25.67
	kNN	∞		29.07	29.02

Fashion MNIST: Similar to original MNIST handwritten data, the fashion MNIST [31] consists of 70000 images from 10 different fashion categories (top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot). Each image consists of 28×28 grayscale pixels.

Fig. 9 shows a snapshot of images. It is essential to learn the link between the images available over the social media and their respective label or category. The data was reduced from 784 dimensions to 8 using CcLTSA and other methods. Out of 70000 instances, in each round out of 10, 42469 instances were randomly selected for training leaving rest for testing.

Figure 9: Fashion MNIST (10 categories)

Table II depicts the mean classification error for both kNN and SVM classifiers using all dimensionality reduction methods. As evident, the proposed CcLTSA gave accurate predictions against all NN values, which shows the robustness of the proposed method. CcLTSA was able to increase the classification accuracy to ≈85% by giving the optimal low dimensional representation of the data. ISOMAP in here also failed to find a single connected component for NN=7-13 and thus no model was trained using ISOMAP for the said values. On comparing the methods by taking mean across all NN values for both kNN and SVM classifier, CcLTSA gave most accurate results followed by PCA, LLE, LE, LTSA, HE and ISOMAP.

Table II
Fashion MNIST (d=8) Mean Error

Method	Classifier	NN=7	NN=11	NN=13	NN=17
CcLTSA	SVM	**20.97**	23.16	**21.17**	**13.82**
	kNN	**15.32**	**19.79**	**20.49**	**15.18**
PCA	SVM	21.23	**21.23**	21.23	21.23
	kNN	22.07	22.07	22.07	22.07
LLE	SVM	23.77	23.84	22.83	22.29
	kNN	23.87	25.12	24.61	24.30
LTSA	SVM	24.86	24.25	23.65	23.27
	kNN	27.23	24.98	24.76	24.57
LE	SVM	23.03	23.03	23.26	23.26
	kNN	26.24	24.89	24.86	25.35
HE	SVM	31.40	28.90	27.30	26.31
	kNN	36.74	27.84	27.51	26.57
ISOMAP	SVM	∞			27.14
	kNN	∞			35.18

Lego Bricks Set: Fig. 10 shows a snapshot of the given data [32]. The data is spanned across 16 categories of different building bricks manufactured by Lego. Each category contains ≈400 images wherein each image consist of 200×200 grayscale pixels. However, by exploiting the intrinsic geometrical information, the data can be represented using just 5 dimensions. The above listed state-of-the-art dimensionality reduction methods along with CcLTSA have been employed to change the given data representation to its optimal form.

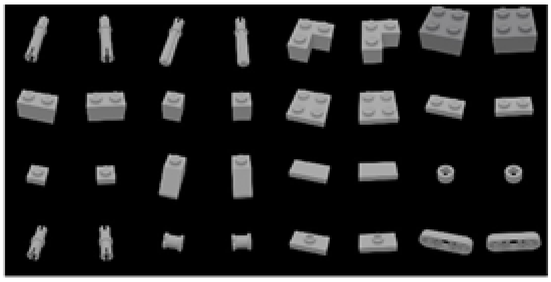


Figure 100: Lego Bricks (16 Categories)

Table III
Lego Bricks (d=5) Mean Error

Method	Classifier	NN=7	NN=11	NN=13	NN=17
CcLTSA	SVM	**8.84**	**9.54**	**10.06**	**9.23**
	kNN	**17.63**	**20.01**	**16.61**	**17.35**
PCA	SVM	24.37	24.37	24.37	24.37
	kNN	27.32	27.32	27.32	27.32
LLE	SVM	22.67	32.93	24.11	26.51
	kNN	19.62	28.53	21.40	30.87
LTSA	SVM	58.83	83.75	23.64	22.59
	kNN	44.82	76.39	23.28	29.82
LE	SVM	60.12	62.22	69.52	69.28
	kNN	60.36	63.25	64.20	62.24
HE	SVM	34.08	19.44	21.68	22.94
	kNN	28.49	23.14	21.79	22.43
ISOMAP	SVM	64.18	86.48	25.62	14.86
	kNN	40.52	76.63	24.85	29.54

Table III contains the mean classification error obtained using kNN and SVM classifiers. Except for NN independent PCA method, the change in representation through all other methods were observed by varying NN values as mentioned in the table. Across all parameters, the proposed CcLTSA method gave optimal representation, thus increasing the underlying models' accuracy to ≈92%. HE gave secondmost accurate inferences followed by PCA, LLE, LTSA, ISOMAP, and LE.

Mobile Pricing: Today the Internet is full of description of electronic devices with the majority of mobile devices containing their features, reviews, and price. In this experiment, a similar data set has been used containing 3000 mobile handset description spanned across 20 features [33]: battery capacity, Bluetooth, WiFi, processor, 4G, 3G, camera, etc. The aim is to predict the price segment (0-3; where 0-low cost and 3-very high cost) of the respective handset based on input features. The data was reduced from 20 input dimensions to 4 dimensions optimally representing the data on its intrinsic dimensional space.

Table IV
Mobile Pricing (d=4) Mean Error

Method	Classifier	NN=7	NN=11	NN=13	NN=17
CcLTSA	SVM	12.49	6.90	**3.60**	**3.95**
	kNN	14.66	**6.22**	**10.49**	**5.96**
PCA	SVM	**6.22**	6.22	6.22	6.22
	kNN	**12.19**	12.19	12.19	12.19
LLE	SVM	21.85	5.59	6.22	4.95
	kNN	22.87	7.24	11.69	7.49
LTSA	SVM	59.08	**5.33**	5.71	6.22
	kNN	29.22	11.05	11.05	12.07
LE	SVM	11.94	9.14	9.53	10.54
	kNN	13.72	11.94	11.56	12.83
HE	SVM	12.83	10.67	14.99	13.59
	kNN	14.10	12.32	16.13	16.90
ISOMAP	SVM	∞	14.30	12.61	16.54
	kNN	∞	16.13	11.69	22.87

Table IV lists the mean classification error of kNN and SVM classifiers for listed dimensionality reduction methods on different NN value for graph creation. As evident, the proposed CcLTSA achieved ≈94% accurate prediction for kNN classifier on NN=11. For small values of NN, ISOMAP failed to create a fully connected graph, and hence, no classifier model was trained for the same. On the mean error of both classifiers on all NN values, CcLTSA remained the most accurate method followed by PCA, LLE, LE, HE, LTSA, and ISOMAP.

REMARKS: The big data generated and consumed on the social network remains difficult for ML algorithms to give accurate inference due to the inherent large number of dimensions. It is well known that not all dimensions span the data, but few of them are present due to unknown transformations and noise in the data. Thus, an optimal dimensionality reduction method is required to project this high-dimensional data to its true intrinsic dimension containing sufficient discriminative power.

Though PCA performs well on linear data-set, it loses the non-linear relationship between the data. The state-of-the-art methods LTSA, LLE, ISOMAP, LE and HE preserve the global non-linear properties by exploiting and preserving the local linear property hidden in the data. However, the measure of similarity between data instances in these algorithms relies on some distance function like Euclidean or cosine, etc. The absence of statistical similarity measurement does not allow these algorithms to exploit the intrinsic information present in the data. CcLTSA utilizes HSIC based cross-covariance operator to measure the independence between two random variables, i.e., point of interest and its local neighbors. If the HSIC value is closer to 1, then, the two random variables are more dependent. Similarly, on the lower dimension, this local property should be preserved. The more similar are two instances, they should lie spatially closer to each other on the target dimension, and this is assured by cross-covariance statistical independence measurement.

Extensive experiments performed on seven synthetic and four real-world datasets show that the proposed CcLTSA gave a true low-dimensional representation of the given high dimensional data which further increased the underlying classifiers' model accuracy. The existing state-of-the-art method fails to perform due to their sole dependence on the spatial relation between the data instances. Due to unknown transformations and inherent noise in the data, the spatial similarity proves to be grossly inaccurate. Here, statistical independence measurement using the cross-covariance operator in CcLTSA allows the method to increase the weightage of dependent instances. The proposed method optimizes the low dimensional representation by maximizing the statistical dependence between instances. This allows CcLTSA to preserve local linear property along with global non-linear property. This makes visualization, modelling, clustering, classification, regression, etc. on the data easy and accurate.

5. CONCLUSION

The massive content generated on social media needs effective and efficient algorithms for data analysis, visualization, or clustering, etc. Due to high feature correlation and noise present in data, these tasks become difficult. Extensive experiments on both synthetic and real-world data set prove that the proposed cross-covariance based LTSA efficiently removes the effect of noise and co- related attributes allowing the underlying algorithms to extract the true intrinsic properties of data. On the known synthetic manifolds, the results show that the proposed method preserved the smoothest and true intrinsic geometry. Moreover, the increased accuracy of linear classifiers on the real-world data set proves that the representation obtained from proposed method holds maximum discriminative features and minimum noised. In performance bench-marking, on the data set containing product details scraped from social media, the proposed method achieved ≈ 92% of accuracy and outperformed the existing state-of-the-art dimensionality reduction methods by $\geq 3\%$.

REFERENCES

[1] S. Moro, P. Rita, B. Vala, Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach, Journal of Business Research 69 (9) (2016) 3341–3351.

[2] J. Surma, A. Furmanek, Data mining in on-line social network for marketing response analysis, in: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 537–540. doi:10.1109/PASSAT/SocialCom.2011.72.

[3] L. del Carmen Contreras Chinchilla, K. A. R. Ferreira, Analysis of the behavior of customers in the social networks using data mining techniques, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 623–625. doi:10.1109/ASONAM.2016.7752301.

[4] D. Tiwari, M. Kumar, Social media data mining techniques: A survey, in: Information and Communication Technology for Sustainable Development, Springer, 2020, pp. 183–194.

[5] S. Wang, J. Cao, P. S. Yu, Deep learning for spatio-temporal data mining: A survey, arXiv preprint arXiv:1906.04928.

[6] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, M. M. Bronstein, Geometric deep learning on graphs and manifolds using mixture model cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5115–5124.

[7] L. Liu,W. K. Cheung, X. Li, L. Liao, Aligning users across social networks using network embedding., in: IJCAI, 2016, pp. 1774–1780.

[8] T. You, H.-M. Cheng, Y.-Z. Ning, B.-C. Shia, Z.-Y. Zhang, Community detection in complex networks using density-based clustering algorithm and manifold learning, Physica A: Statistical Mechanics and its Applications 464 (2016) 221–230.

[9] C. Campbell, C. Ferraro, S. Sands, Segmenting consumer reactions to social network marketing, European Journal of Marketing 48 (3/4) (2014) 432–452.

[10] C. Van den Bulte, S. Wuyts, Social networks and marketing, Marketing Science Institute Cambridge, MA, 2007.

[11] G. Huitsing, R. Veenstra, Bullying in classrooms: Participant roles from a social network perspective, Aggressive behavior 38 (6) (2012) 494–509.

[12] B. S. Nandhini, J. Sheeba, Online social network bullying detection using intelligence techniques, Procedia Computer Science 45 (2015) 485–492.

[13] S. P. Crain, K. Zhou, S.-H. Yang, H. Zha, Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond, in: Mining text data, Springer, 2012, pp. 129–161.

[14] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, Line: Large-scale information network embedding, in: Proceedings of the 24th international conference on world wide web, International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.

[15] D.Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2016, pp. 1225–1234.

[16] I. Jollie, Principal component analysis, Springer, 2011.

[17] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.

[18] J. B. Kruskal, M. Wish, Multidimensional scaling, Vol. 11, Sage, 1978.

[19] T. Jia, M. Jian, L. Wu, Y. He, Modular manifold ranking for image recommendation, in: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), 2018, pp. 1–5. doi:10.1109/BigMM.2018.8499455.

[20] W. Zhao, S. Tan, Z. Guan, B. Zhang, M. Gong, Z. Cao, Q. Wang, Learning to map social network users by unified manifold alignment on hypergraph, IEEE Transactions on Neural Networks and Learning Systems 29 (12) (2018) 5834–5846. doi:10.1109/TNNLS.2018.2812888.

[21] K. Kim, J. Lee, Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction, Pattern Recognition 47 (2) (2014) 758–768.

[22] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, SIAM JOURNAL ON SCIENTIFIC COMPUTING (2004) 313–338.

[23] J. B. Tenenbaum, V. De Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, science 290 (5500) (2000) 2319–2323.

[24] L. K. Saul, S. T. Roweis, An introduction to locally linear embedding.

[25] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation 15 (6) (2003) 1373–1396.

[26] D. L. Donoho, C. Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data,

Proceedings of the National Academy of Sciences 100 (10) (2003) 5591–5596.

[27] A. Gretton, O. Bousquet, A. Smola, B. Sch¨olkopf, Measuring statistical dependence with hilbert-schmidt norms, in: International conference on algorithmic learning theory, Springer, 2005, pp. 63–77.

[28] I. Gohberg, S. Goldberg, M. A. Kaashoek, Hilbert-schmidt operators, in: Classes of Linear Operators Vol. I, Springer, 1990, pp. 138–147.

[29] E. Kreyszig, Introductory functional analysis with applications, Vol. 1, wiley New York, 1978.

[30] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[31] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017). arXiv:cs.LG/1708.07747.

[32] J. Hazelzet, Images of lego bricks. URL https://www.kaggle.com/joosthazelzet/lego-brick-images

[33] A. Sharma, Mobile price classification. URL https://www.kaggle.com/iabhishekofficial/mobile-price-classification

AUTHOR PROFILES

Sunil Kumar Mishra pursed his Bachelor of Technology in Computer Science from Dr. A.P.J. Abdul Kalam Technical University Uttar Pradesh, Lucknow in year 2005 and Master of Technology in year 2012. He is currently pursuing Ph.D. from IFTM University, Moradabad under the supervision of Professor Ajay Agarwal and working as an Assistant Professor in the Department of Computer Sciences, Accurate Institute of Management and Technology, Greater Noida since 2012. Previously he worked as a lecturer in KNIT, Sultanpur from 2005 to 2007 and as an Assistant professor in UCER, Greater Noida from 2007 to 2012, making it approximately 14 years of academic experience. He has published many research papers in reputed journals and conferences. His main research work focuses on social network mining, web mining and literature mining etc.

Dr. Ajay Agarwal received his B. Tech. degree from IET Lucknow (a constituent college of AKTU Lucknow), in year 1979. He completed his Master of Engineering (Computer Science and Engineering) with honours from M.N.N.I.T. Allahabad in year 1995 and Ph.D. from I.I.T. Delhi in year 2006. He is a versatile professional with a strong academic background. His area of research includes mobile adhoc networking, soft computing, data base management system. He has received many awards. He has published many papers in reputed journals and conferences. He is an author of a book. Presently he is working as Professor in Information Technology Department at KIET Group of Institutions, Ghaziabad