# A semi-supervised clustering approach using nonlinear canonical correlation analysis with t-SNE



Hong, X. ORCID: https://orcid.org/0000-0002-6832-2298, Xiao, J. and Wei, H. ORCID: https://orcid.org/0000-0002-9664-5748
(2024) A semi-supervised clustering approach using nonlinear canonical correlation analysis with t-SNE. In: 2024 International Joint Conference on Neural Networks (IJCNN), 30 Jun- 5 Jul 2024, Yokohoma, Japan. doi: https://doi.org/10.1109/ijcnn60899.2024.10650841 Available at https://centaur.reading.ac.uk/116055/

# A Semi-Supervised Clustering Approach Using Nonlinear Canonical Correlation Analysis with t-SNE

Xia Hong
Department of Computer Science
University of Reading, UK

James Xiao
Independent Researcher
Vancouver, BC, Canada

Hong Wei
Department of Computer Science
University of Reading, UK

***Abstract*—Clustering of high-dimensional data is a challenging task, since the usual distance measures in high-dimensional space cannot reflect how clusters are partitioned. In this work, by assuming there are some data examples with known labels, a new semi-supervised clustering approach is proposed using a modified canonical correlation analysis and t-SNE. Initially, t-SNE projects high dimensional data onto 3D embedding. While the clusters in the t-SNE embedding space may be visually separable, it is still challenging to achieve very good clustering performance with a conventional unsupervised clustering algorithm. In this work, by using radial basis functions (RBFs) in t-SNE embedding space, centred as some labelled points, a modified canonical correlation analysis algorithm is introduced. The proposed algorithm is referred to as RBF-CCA, which learns the associated projection matrix using supervised learning on the small labelled data set, followed by projection of the associated canonical variables to a large amount of unlabelled data. Then, k-means clustering is applied as the final clustering step. To demonstrate its effectiveness, the proposed algorithm is experimented on several benchmark image data sets.**

***Index Terms*—semi-supervised learning, embedding, canonical correlation analysis, radial basis function, spectral clustering.**

## I. Introduction

When dealing with empirical data, it is common to begin by identifying groups of similar behaviour in the data sets. Over the past decades, various clustering methods have been proposed, including centroid-based clustering [1], density-based clustering [2], spectral clustering [3], distribution-based clustering [4], ensemble clustering [5], etc. Clustering is one of the most prominent research topics in machine learning, computer vision, data mining and various scientific applications [3], [6], [7]. Data visualization via dimensionality reduction is critically important for understanding and interpreting the clustering structure of large data sets. The t-distributed stochastic neighbor embedding (t-SNE) algorithm, proposed by van der Maaten and Hinton [8], is a state-of-art technique of dimensionality reduction and visualization for a wide range of applications [9]. Cai and Ma investigated the theoretical foundations of t-SNE, to provide theoretical guidance for applying t-SNE and help select its tuning parameters in various applications [10]. Note that since the t-SNE technique cannot be used for predicting embeddings of new data samples, the data set needs to be defined in advance.

Data clustering, in most cases, is an unsupervised classification paradigm which divides observed data into different subsets (clusters), such that similar objects are allocated to the same subset while dissimilar objects are assigned to different subsets. However, sometimes there is prior knowledge in the clustering output space which can be used to enhance the clustering results, such as some cluster labels being available for a subset of observations, or we may know that some observations belong to the same cluster, or must belong to different clusters. Clustering algorithms which combine complementary information to supervise the cluster learning process are called semi-supervised clustering methods [11]. One of which, the concept of few shots learning (FSL) is proposed [12], [13] which refers to the learning from a limited number of examples with supervised information. This learning paradigm is desired since large-scale labelled datasets can be expensive in many applications, e.g. in semantic segmentation to associate a label or category with every pixel in an image [14].

The canonical correlation analysis (CCA) was originally proposed by H. Hotelling in the seminal works [15], [16]. The subject of this work is finding the best predictors among the linear functions from each set by maximizing the correlation coefficient between two sets. CCA can simplify the statistical analysis for two set variables and properly solve the aforementioned problem, by defining a sequence of pairs of variates as canonical variates and the correlations between them as canonical correlations. The complicated high-dimensional variable sets are projected in the common latent subspace with a desired low dimension. In an attempt to increase the flexibility of CCA for nonlinear relationships between two random variables, kernel CCA (KCCA) has been introduced [17]. CCA/KCCA was applied for effective feature selection tool as a preprocessing step for classifiers of support vector machine or random forests [18]. As a powerful tool for multimodal feature fusion, CCA/KCCA has received widespread attentions in both theoretical advances and applications (see [19] and references within).

In this work, a novel semi-supervised clustering algorithm is proposed using a modified canonical correlation analysis and t-SNE. The contributions of this work are summarized as follows:

1) The proposed algorithm operates over a 3D embedding

space via t-SNE, aimed at clustering of high dimensional data sets.
2) In order to provide an approximation of the nonlinear shapes of t-SNE clusters, radial basis functions using labelled-data input features, are employed as the first set of variables for RBF-CCA, the class labels become indicative variables as a second set of variables in CCA.
3) The proposed algorithm transfers knowledge from labelled data sets, via projection of the associated canonical variables, to a large amount of unlabelled data, based on which k-means clustering is applied as used by a spectral clustering algorithm [20].
4) The clustering algorithm can be viewed as a generalized spectral clustering algorithm, due to that CCA is based on eigen-decomposition of data matrices.

The remainder of paper is organized as follows: Section II introduces the proposed semi-supervised algorithm based on RBF-CCA and t-SNE. Section II-A presents the problem statement and the objective. Section II-B provides preliminaries of CCA, followed by Section II-C which presents the problem of the semi-supervised clustering and the proposed algorithmic solution. In Section III, numerical experiments are performed based on four well-known image data sets to compare with the baseline approaches of a k-means clustering algorithm based on the original data and the t-SNE embedding, respectively. Section IV is devoted to conclusions.

## II. Proposed semi-supervised algorithm based on RBF-CCA and t-SNE

### A. Problem statement

To consider a task of semi-supervised few-shots clustering, we have some labelled data points $N_l \ll N_u$ , as $D_{train} = \{\boldsymbol{x}_i, t_i\}_{i=1}^{N_l}$, where $\boldsymbol{x}_i = [x_{i,1}, ..., x_{i,n}]^{\mathrm{T}} \in \Re^n$ is a high dimensional input feature vector, there is also a large test data set $D_{test} = \{\boldsymbol{x}_i, t_i\}_{i=N_l+1}^{N}$, $N = N_l + N_u$. The problem is to assign each $\boldsymbol{x}_i$ a cluster $\hat{t}_i \in \{1, ...K\}$, that minimizes the average discrepancies between $t_i$ and $\hat{t}_i$ for all $i$. We initially adopt an unsupervised learning stage via the t-SNE embedding construction, followed by building the so-called canonical variables using radial basis functions from $D_{train}$'s t-SNE embedding. Finally, k-means clustering is applied to these canonical variables, learnt from RBF-CCA projection, over both $D_{train}$ and $D_{test}$, in order to obtain cluster membership of each index $i$.

Note that the problem and solution differ fundamentally from a classification problem, in which one trains a classifier $H : \boldsymbol{x}_i \rightarrow \hat{t}_i$ using an optimization algorithm, which involves minimizing training errors over $D_{train}$, then the classifier is applied to $D_{test}$ or any new data samples. The main objective of the proposed semi-supervised clustering algorithm is to offer improvement over clustering performance of an unsupervised algorithm. The clustering algorithms, with limited labels in large amounts of data, is desired in many applications, since a classification algorithm can be expensive due to its dependency on data collection and labelling, or being computationally costly to train.

### B. Canonical correlation analysis

Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two sets of multidimensional variables. Consider two sets of random variables $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_i\}$, $i = 1, ..., d_1$ and $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_i\}, i = 1, ..., d_2$ with zero mean. It is assumed that $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ are full rank, and $d = \min(rank(\boldsymbol{\Phi}), rank(\boldsymbol{\Psi}))$.

Denote the total covariance matrix as

$$C = \begin{bmatrix} C_{\boldsymbol{\Phi}\boldsymbol{\Phi}} & C_{\boldsymbol{\Phi}\boldsymbol{\Psi}} \\ C_{\boldsymbol{\Psi}\boldsymbol{\Phi}} & C_{\boldsymbol{\Psi}\boldsymbol{\Psi}} \end{bmatrix} = E\left[ \begin{pmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\Psi} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Phi} \\ \boldsymbol{\Psi} \end{pmatrix}^{\mathrm{T}} \right] \quad (1)$$

Define a set of two projection matrices $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_d] \in \Re^{d_1 \times d}$ and $\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_d] \in \Re^{d_2 \times d}$, which generate a set of linear combinations named $\boldsymbol{U} = \boldsymbol{\Phi}\boldsymbol{A} = [\boldsymbol{u}_1, ...\boldsymbol{u}_d]$ and $\boldsymbol{V} = \boldsymbol{\Psi}\boldsymbol{B} = [\boldsymbol{v}_1, ...\boldsymbol{v}_d]$. Each member of $\boldsymbol{U}$ is paired with a member of $\boldsymbol{V}$, as a set of canonical variables pairs $(\boldsymbol{u}_i, \boldsymbol{v}_i)$.

The task in CCA is to find $\boldsymbol{A}, \boldsymbol{B}$ such that the correlations $\rho_i(\boldsymbol{u}_i, \boldsymbol{v}_i)$ are maximized. Represent

$$\rho_i = \frac{\boldsymbol{a}_i^T C_{\boldsymbol{\Phi}\boldsymbol{\Psi}} \boldsymbol{b}_i}{\sqrt{\boldsymbol{a}_i^T C_{\boldsymbol{\Phi}\boldsymbol{\Phi}} \boldsymbol{a}_i} \sqrt{\boldsymbol{b}_i^T C_{\boldsymbol{\Psi}\boldsymbol{\Psi}} \boldsymbol{b}_i}}, \quad i = 1, ..., d. \quad (2)$$

Equivalently,

$$\max_{\boldsymbol{a}_i, \boldsymbol{b}_i} \rho_i = \boldsymbol{a}_i^T C_{\boldsymbol{\Phi}\boldsymbol{\Psi}} \boldsymbol{b}_i, \forall i \quad (3)$$

subject to $\boldsymbol{a}_i^T C_{\boldsymbol{\Phi}\boldsymbol{\Phi}} \boldsymbol{a}_i = 1$, $\boldsymbol{b}_i^T C_{\boldsymbol{\Psi}\boldsymbol{\Psi}} \boldsymbol{b}_i = 1$.

To obtain the CCA solution, initially define [21]

$$\mathcal{K} = C_{\boldsymbol{\Phi}\boldsymbol{\Phi}}^{-1/2} C_{\boldsymbol{\Phi}\boldsymbol{\Psi}} C_{\boldsymbol{\Psi}\boldsymbol{\Psi}}^{-1/2} \quad (4)$$

and perform singular value decomposition of $\mathcal{K}$ as

$$\mathcal{K} = \Gamma \Lambda \Delta^T \quad (5)$$

with $\Gamma = [\gamma_1, ...\gamma_d]$, $\Delta = [\delta_1, ..., \delta_d]$. $\Lambda = \mathrm{diag}\{\lambda_1^{1/2}, ...\lambda_d^{1/2}\}$. and $\lambda_1 \geq \lambda_2 \geq ...\lambda_d$ are the nonzero eigenvalues of $\mathcal{K}^T\mathcal{K}$. $\gamma_i$ and $\delta_i$ are the left and right eigenvectors of $\mathcal{K}$.

Now define

$$\begin{aligned} \boldsymbol{a}_i &= C_{\boldsymbol{\Phi}\boldsymbol{\Phi}}^{-1/2} \gamma_i \\ \boldsymbol{b}_i &= C_{\boldsymbol{\Psi}\boldsymbol{\Psi}}^{-1/2} \delta_i \end{aligned} \quad (6)$$

so that

$$cov(\boldsymbol{u}_i, \boldsymbol{u}_j) = \boldsymbol{a}_i^T C_{\boldsymbol{\Phi}\boldsymbol{\Phi}} \boldsymbol{a}_j = \gamma_i^T \gamma_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (7)$$

$$cov(\boldsymbol{v}_i, \boldsymbol{v}_j) = \boldsymbol{b}_i^T C_{\boldsymbol{\Psi}\boldsymbol{\Psi}} \boldsymbol{b}_j = \delta_i^T \delta_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (8)$$

and the correlation between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ has maximal of

$$\rho(\boldsymbol{u}_i, \boldsymbol{v}_i) = \gamma_i^T \Gamma \Lambda \Delta^T \delta_i = \lambda_i^{1/2}. \quad (9)$$

### C. Proposed semi-supervised spectral clustering algorithm using RBF-CCA

Suppose that we have some $N_l \ll N_u$ labelled data points, as $D_{train} = \{\boldsymbol{x}_i, t_i\}_{i=1}^{N_l}$, where $\boldsymbol{x}_i = [x_{i,1}, ..., x_{i,n}]^{\mathrm{T}} \in \Re^n$ is a high dimensional input feature vector. $t_i \in \{1, ...K\}$ for a given $K \ll N_l$. Let the input data points with labels be denoted as $\boldsymbol{X}^{(L)} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{N_l}]^{\mathrm{T}}$. Simultaneously, there are $N_u \gg N_l$ unlabelled data points given as $\boldsymbol{X}^{(U)} = [\boldsymbol{x}_{N_l+1}, ..., \boldsymbol{x}_N]^{\mathrm{T}}$. Denote the completed input data points

$$\boldsymbol{X} = \left[ \begin{array}{c} \boldsymbol{X}^{(L)} \\ \boldsymbol{X}^{(U)} \end{array} \right] \in \Re^{N \times n},$$

The goal of semi-supervised clustering is to partition these points into $K$ disjoint sets, with partial assistance of supervised learning from the data set $D_{train}$.

Initially, we apply t-SNE algorithm (Appendix A) for dimensionality reduction $\boldsymbol{x}_i \in \Re^n \longrightarrow \boldsymbol{y}_i \in \Re^m$. Denote $\boldsymbol{Y} = \text{t-SNE}(\boldsymbol{X})$ as the full set of input features.

It is known that the t-SNE algorithm can generate visually irregular shaped clusters. However, if a conventional clustering algorithm, e.g. k-means, is applied to assign cluster labels to $\boldsymbol{y}_i$, it may not be very effective, due to that the clustering criterion in k-means is oversimplified, minimizing average distances to each cluster centre, which is not suitable for irregularly shaped clusters. An improvement would be to have a better approximation to the actual cluster shapes.

Radial basis functions are a cornerstone in approximation theory [22]. The output of a radial basis function reduces as the distance between the input and its fixed centre increases, giving its ability to locally identify new data to these centres. In order to model the irregularly shaped t-SNE clusters, we propose to use a modified CCA based on radial basis functions using $D_{train}$, referred to as RBF-CCA. Since within the visible t-SNE separate clusters, the training data embeddings are assigned known labels, these are set to be centres using $D_{train}$. For convenience, we denote the embedding of labelled data points as $\boldsymbol{c}_j = \boldsymbol{y}_j$, $i = 1, ..., N_l$. Over $\boldsymbol{Y}$, let $\phi_{i,j} = \exp\left(-\frac{\|\boldsymbol{y}_i - \boldsymbol{c}_j\|^2}{2\sigma^2}\right)$, in which $\boldsymbol{c}_j$ are the centres of radial basis functions, $\sigma$ is a preset proper hyperparameter. In this work, we simply use

$$\sigma = \sqrt{\frac{1}{NN_l} \sum_{i=1}^{N} \sum_{j=1}^{N_l} \|\boldsymbol{y}_i - \boldsymbol{c}_j\|^2} \tag{10}$$

to construct a matrix

$$\boldsymbol{\Phi}^{full} = \left[ \begin{array}{ccc} \phi_{1,1} & \cdots & \phi_{1,N_l} \\ \vdots & \ddots & \vdots \\ \phi_{N,1} & \cdots & \phi_{N,N_l} \end{array} \right] \tag{11}$$

followed by mean removal as

$$\boldsymbol{\Phi}^{full} \leftarrow \boldsymbol{\Phi}^{full} - \text{mean}(\boldsymbol{\Phi}^{full}).$$

Let

$$\boldsymbol{\Phi}^{full} = \left[ \begin{array}{c} \boldsymbol{\Phi} \\ \boldsymbol{\Phi}^U \end{array} \right], \tag{12}$$

where $\boldsymbol{\Phi} \in \Re^{N \times n_l}$ is based on labelled data $D_{train}$, which is used as the first set of variables in our proposed RBF-CCA. By examining (11), we can see that this fills up all pair-wise RBFs between a query (input features in t-SNE) to that of training data (with known labels). At each row, only data points close to any given labelled training data will be excited, otherwise their values are close to zero. Since t-SNE have visible clusters, this means that each data query, will generate some significant values according to the labelled points associated to some cluster, linked to the second set of variable in CCA designed as follows.

In order to create the second set of variables with the objective of semi-supervised clustering, $\boldsymbol{\Psi} \in \Re^{N_l \times K}$ is generated as

$$\boldsymbol{\Psi} = \left[ \begin{array}{ccc} \psi_{1,1} & \cdots & \psi_{1,K} \\ \vdots & \ddots & \vdots \\ \psi_{N_l,1} & \cdots & \psi_{N_l,K} \end{array} \right] \tag{13}$$

in which

$$\psi_{i,j} = \left\{ \begin{array}{ll} 1 & j = t_i \\ 0 & j \neq t_i \end{array} \right. \tag{14}$$

followed by mean removal as

$$\boldsymbol{\Psi} \leftarrow \boldsymbol{\Psi} - \text{mean}(\boldsymbol{\Psi}).$$

Clearly, using canonical correlation analysis (CCA) for the two sets of multidimensional variables $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, as defined above, aims to capture the relationship between the nonlinear RBF transform based on t-SNE clusters and the cluster labels for the labelled data set $D_{train}$. The proposed algorithm, as shown in Algorithm 1, modifies CCA in Sec II-B, based on RBF functions constructed in labelled data via t-SNE embedding mapping, and is referred to as RBF-CCA. Note that Line 5 returns the canonical variable corresponding to the full input features, which are used for spectral clustering. The proposed semi-supervised spectral clustering based on RBF-CCA is given in Algorithm 2.

**Algorithm 1** Modified CCA using radial basis functions (RBF-CCA).

**Require:** Number $K$ of clusters to construct; Labelled data $D_{train} = \{\boldsymbol{x}_i, t_i\}_{i=1}^{N_l}$; Unlabelled data points $\boldsymbol{X}^{(U)} = [\boldsymbol{x}_{N_l+1}, ..., \boldsymbol{x}_N]^{\mathrm{T}}$. Complete data in space of t-SNE $\boldsymbol{Y}$.

1: Construct $\boldsymbol{\Phi}^{full}$ and $\boldsymbol{\Psi}$ using (11) and (13) respectively, then remove their mean.
2: Recover $\boldsymbol{\Phi}$ using (12).
3: Perform CCA to obtain $\boldsymbol{A} \in \Re^{N_l \times K}$ and $\boldsymbol{B} \in \Re^{K \times K}$ (Section II-B)
4: Calculate $\boldsymbol{U}^{full} = \boldsymbol{\Phi}^{full}\boldsymbol{A}$ using complete input data set.
5: Return $\boldsymbol{U}^{full} = \{u_{i,j}\} \in \Re^{N \times K}$.

**Remarks:**

1) The computational complexities of each part of the proposed algorithm is in the order of $O(nN^2)$ (t-SNE), which is further scaled by iterations of gradient descent algorithms, $O(N_l^3)$ (RBF-CCA), and $O(N)$ (k-means

**Algorithm 2** Proposed semi-supervised algorithm based on RBF-CCA and t-SNE

**Require:** Number $K$ of clusters to construct; Labelled data $D_{train} = \{\boldsymbol{x}_i, t_i\}_{i=1}^{N_l}$; Unlabelled data points $\boldsymbol{X}^{(U)} = [\boldsymbol{x}_{N_l+1}, ..., \boldsymbol{x}_N]^{\mathrm{T}}$.

1: Form complete input data matrix $\boldsymbol{X}$. Apply $\boldsymbol{Y}$ = t-SNE($\boldsymbol{X}$) to each point (Appendix A).
2: Call Algorithm 1.
3: Form the matrix $\boldsymbol{Z} = \{z_{i,j}\} \in \Re^{N \times K}$ by normalising the rows to norm one, i.e. to set

$$z_{i,j} = u_{i,j} \Big/ \sqrt{\sum_{j=1}^{K} u_{i,j}^2}. \tag{15}$$

4: **for** $i = 1, ..., N$ **do**
5: Let $\hat{\boldsymbol{z}}_i \in \Re^K$ be the vector corresponding to the $i$th row of $\boldsymbol{Z}$.
6: **end for**
7: Cluster the points $\hat{\boldsymbol{z}}_i$, $i = 1, ..., N$ with the $k$-means algorithm [6] into clusters $C_1, ..., C_K$.
8: Return: Find clusters $k \in \{1, ..., K\}$ with $\{k, \hat{\boldsymbol{z}}_i \in C_k\}$ and assign original data points $\boldsymbol{x}_i$ according to cluster's index set of $k = 1, ..., K$.

clustering). Since $N \gg N_l$, the main cost is due to that of obtaining t-SNE embedding, which is a drawback for problems with large $N$ and $n$.

2) The proposed algorithm can transfer knowledge that are learnt from RBF-CCA using labelled data sets via projection of the associated canonical variables. Because canonical variables, which are related to eigenvectors, are applied for clustering, this approach is similar to spectral clustering algorithms [20].

## III. Experiments

Four image data sets *pendigits*, *usps*, *mnist* [7] and *fashion-mnist* [23] are used for validating the proposed algorithm. A brief summary of the four data sets is provided in Table I. We used this type of data set since it is generally large with high dimensionality, thus likely to benefit from dimensionality reduction for improved performance. The data set *pendigits* is a handwritten digit data set of 250 samples from 44 writers, collected as sampled coordinate information of each digit from a tablet. The data set *usps* is a standard handwritten database, and *mnist* is a handwritten digits data set, presented as a fixed-sized image. The *fashion-mnist* data set is proposed as a more challenging replacement data set for *mnist*. It is a data-set comprised of 70,000 28×28 pixel gray scale images of items of 10 types of clothing, such as shoes, T-shirts, dresses, and more. A visualization comparison between different types is shown in Figure 1. Each data sample in $\boldsymbol{X}$ has undergone a 3D t-SNE algorithm, to obtain the embedding $\boldsymbol{Y}$, as shown in Figure 2. Note that these data sets are originally divided into training and test data sets for supervised classification tasks. In this work, we merge the two parts for our semi-supervised setting. Then the whole data set are divided by a randomly drawn labelled subset with size ($N_l$) for semi-supervised, i.e, the labels are used in training of the modified CCA model. The average results over repeated random experiments are demonstrated.

The comparative methods are explained as follows:

1) The k-mean algorithm was applied to original data feature.
2) The k-means algorithm was applied in 3D t-SNE as described in this work.
3) Linear CCA algorithm: This semi-supervised method is devised to validate effectiveness of RBF-CCA, in which the same amount of labelled data are used for training, and the t-SNE features and CCA are also applied. There exists a key difference, which is that $\boldsymbol{\Phi}^{full}$ is formed differently. In linear CCA algorithm, the t-SNE features are used directly without any nonlinear transformation.

The results of the proposed RBF-CCA algorithm are presented in Tables II-V, based on an average of 20 random experiments carried out by varying the number of labelled data samples for training. The ratio of $N_l/N$ is set as low as possible, yet as large as necessary to cover the data space, so there are meaningful estimation results. Therefore, for *pendigits* and *usps* data sets, a minimum of $5\%$ data sample is experimented, while for *mnist* and *fashion-mnist*, a minimum of $1\%$ data sample is experimented, according to their data size. This means only a few hundred labelled data samples over $K = 10$ are used at most, which can lead to much improved clustering performance over both unsupervised k-means algorithm for original data and t-SNE embedding data. It can be seen that as $N_l$ increases, the clustering results improves for all data sets. All algorithms outperform significantly over k-means algorithm in original features, which don't use any labelled data to guide the clustering algorithm. When k-means algorithm was applied in 3D t-SNE space, the performance is much better than in original high dimensional data space for all data sets. For the linear CCA algorithm, the same range of $N_l/N$ is experiments, it can also be seen that as $N_l$ increases, the clustering results also improve. Yet linear CCA algorithm performance is much worse than the proposed RBF-CCA algorithm, and is in general even slightly worse than k-means algorithm in 3D t-SNE. This is because linear CCA is unable to explain nonlinearity in 3D t-SNE clusters, which is crucially important in achieving high clustering performance. The proposed RBF-CCA algorithm is shown to have the best performance for all data sets, due to the use of radial basis functions with a small amount of labelled data sets as centres, thus to fit irregular 3D t-SNE clusters.

## IV. Conclusions

This paper has introduced a new semi-supervised clustering algorithm by modifying canonical correlation analysis in the manner that two sets of random variables are designed using an RBF function. Given a set of large amount of data features, within which a subset has known clustering labels, we have proposed to perform t-SNE embedding initially, based on which RBF-CCA algorithm is then developed. The first set


Fig. 1. 10 sample images per class are shown as visual comparison between (a) *mnist* and (b) *fashion-mnist*; Each data-set has 70000 images.

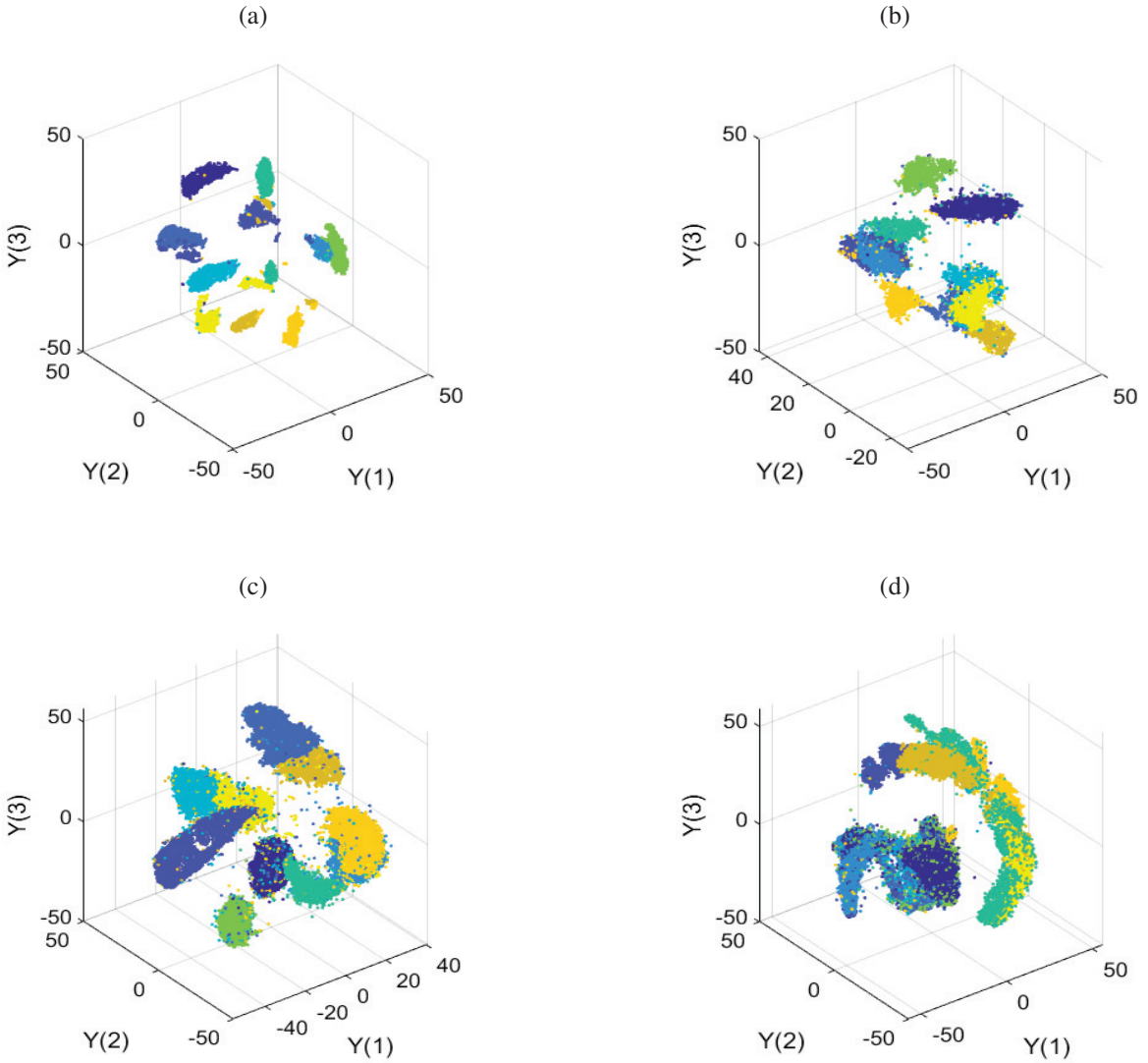


Fig. 2. 3D t-SNE embedding with the true labels indicated by color; (a) *pendigits* and (b) *usps*; (c) *mnist* and (d) *fashion-mnist*;

TABLE I
A SUMMARY OF FOUR IMAGE DATA SETS.

| Data sets | Number of data size ($N$) | Number of features($d$) | Number of classes ($K$) |
|---|---|---|---|
| *pendigits* | 10992 | 16 | 10 |
| *usps* | 9298 | 256 | 10 |
| *mnist* | 70000 | 784 | 10 |
| *fashion-mnist* | 70000 | 784 | 10 |

TABLE II
CLUSTERING ACCURACY OF *pendigits* DATA SET.

| $N_l/N$ | k-means | k-means (t-SNE) | Linear CCA | Proposed CCA-RBF |
|---|---|---|---|---|
| 0 | 77.06 | 86.45 | - | - |
| 0.05 | - | - | $82.90 \pm 4.62$ | $96.54 \pm 3.69$ |
| 0.1 | - | - | $82.57 \pm 3.62$ | $98.49 \pm 0.15$ |
| 0.15 | - | - | $85.59 \pm 1.80$ | $98.57 \pm 0.18$ |

TABLE III
CLUSTERING ACCURACY OF *usps* DATA SET.

| $N_l/N$ | k-means | k-means (t-SNE) | Linear CCA | Proposed CCA-RBF |
|---|---|---|---|---|
| 0 | 66.66 | 66.51 | - | - |
| 0.05 | - | - | $76.49 \pm 3.78$ | $92.34 \pm 2.74$ |
| 0.1 | - | - | $84.82 \pm 6.46$ | $93.44 \pm 3.34$ |
| 0.15 | - | - | $87.96 \pm 6.59$ | $94.22 \pm 3.20$ |

TABLE IV
CLUSTERING ACCURACY OF *minst* DATA SET.

| $N_l/N$ | k-means | k-means (t-SNE) | Linear CCA | Proposed CCA-RBF |
|---|---|---|---|---|
| 0 | 55.33 | 96.22 | - | - |
| 0.01 | - | - | $91.99 \pm 4.58$ | $96.06 \pm 1.00$ |
| 0.02 | - | - | $92.64 \pm 3.62$ | $96.72 \pm 0.10$ |
| 0.05 | - | - | $92.58 \pm 3.78$ | $96.89 \pm 0.04$ |

TABLE V
CLUSTERING ACCURACY OF *fashionMinst* DATA SET.

| $N_l/N$ | k-means | k-means (t-SNE) | Linear CCA | Proposed CCA-RBF |
|---|---|---|---|---|
| 0 | 54.21 | 59.21 | - | - |
| 0.01 | - | - | $56.07 \pm 0.55$ | $76.58 \pm 0.75$ |
| 0.02 | - | - | $56.09 \pm 0.46$ | $80.02 \pm 0.40$ |
| 0.05 | - | - | $56.23 \pm 0.54$ | $81.32 \pm 0.25$ |

of random variables is designed based on a set of RBFs using t-SNE embedding of input features, and the second set of random variables are from their labels. The proposed algorithm learns the associated projection matrix from RBF-CCA, followed by learning the associated canonical variates for the full input features. Finally, the canonical variates are clustered using k-means clustering. The algorithm has been experimented using several benchmark data sets to demonstrate its effectiveness, in comparison with baseline approaches. Since clustering approaches do not make many assumptions on the availability of labelled data, they are desirable in many applications. Our future research will be focused on semi-supervised or unsupervised clustering tasks in computer vision applications, such as object detection and segmentation, by combining other techniques in image analysis. It is also worth investigating alternative design of CCA variables, with the aim to achieve more effective performance for unsupervised clustering approaches.

## APPENDIX A: T-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING (T-SNE)

Given a set of $N$ high-dimensional $\boldsymbol{X} = \{\boldsymbol{x}_i\}, i = 1, ...N$. As a tool for visualizing high-dimensional data by giving each data point a location in a two or three-dimensional map, the t-SNE [8] aims to learn a very low $m$-dimensional map $\boldsymbol{Y} = \{\boldsymbol{y}_i \in \Re^m\}, i = 1, ...N$, with $m$ typically chosen as 2 or 3.

The locations of the points $\boldsymbol{y}_i$ in the map are determined by minimizing the (non-symmetric) Kullback–Leibler divergence of the distribution of the pairwise similarity probability $P$ (of $\boldsymbol{X}$) from the distribution $Q$ (of $\boldsymbol{Y}$):

$$KL(P\|Q) = \sum_{i \neq j} \log \frac{p_{ij}}{q_{ij}} \tag{16}$$

where $p_{ij}$ is joint probabilities of similarity $\{\boldsymbol{x}_j, \boldsymbol{x}_i\}$, $q_{ij}$ is joint probabilities of similarities of $\{\boldsymbol{y}_j, \boldsymbol{y}_i\}$.

To this end, the similarity of two distinctive $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$ is the conditional probability, $p_{j|i}$, defined as:

$$p_{j|i} = \frac{\exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(2\sigma_i^2)\right)}{\sum_{k\neq i}\exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2/(2\sigma_i^2)\right)} \tag{17}$$

for $i \neq j$. Set $p_{i|i} = 0$, and

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{18}$$

By doing this, the distances between high dimensional data points are transferred into probabilities using a Gaussian distribution.

In t-SNE, a Student t-distribution with one degree of freedom (which is the same as a Cauchy distribution) as the heavy-tailed distribution in the low-dimensional map. Set $q_{ii} = 0$. The similarity of two distinctive $\boldsymbol{y}_j$ to $\boldsymbol{y}_i$ is defined as:

$$q_{ij} = \frac{\left(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\right)^{-1}}{\sum_k \sum_{l\neq k}\left(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2\right)^{-1}} \tag{19}$$

for $i \neq j$. $\boldsymbol{y}_i$ in the map are determined using gradient descent algorithm, with a user set hyperparameter using the concept of perplexity, which is then used to determine $\sigma_i$ in (17).

## REFERENCES


[1] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.

[4] R. Wang, S. Han, J. Zhou, Y. Chen, L. Wang, T. Du, K. Ji, Y.-o. Zhao, and K. Zhang, "Transfer-learning-based gaussian mixture model for distributed clustering," *IEEE Transactions on Cybernetics*, vol. 53, no. 11, pp. 7058–7070, 2023.

[5] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar, "From clustering to clustering ensemble selection: A review," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104388, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621002360

[11] Z. Ghasemi, H. A. Khorshidi, and U. Aickelin, "A survey on optimisation-based semi-supervised clustering methods," in *2021 IEEE International Conference on Big Knowledge (ICBK)*, 2021, pp. 477–482.

[6] S. S. Haykin, *Neural networks and learning machines*, 3rd ed., Upper Saddle River, NJ, 2009.

[7] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1669–1680, 2015.

[8] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. nov, pp. 2579–2605, 2008, pagination: 27.

[9] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1061.html

[10] T. T. Cai and R. Ma, "Theoretical foundations of t-sne for visualizing high-dimensional clustered data," *Journal of Machine Learning Research*, vol. 23, no. 301, pp. 1–54, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-0524.html

[12] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[13] Y. Wang and Q. Yao, "Few-shot learning: A survey," *CoRR*, vol. abs/1904.05046, 2019. [Online]. Available: http://arxiv.org/abs/1904.05046

[14] N. Catalano and M. Matteucci, "Few shot semantic segmentation: a review of methodologies and open challenges," 2023.

[15] H. Hotelling, "The most predictable criterion." *Journal of Educational Psychology*, vol. 26, pp. 139–142, 1935.

[16] ——, "Relations between two sets of variables," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936. [Online]. Available: https://doi.org/10.1093/biomet/28.3-4.321

[17] S. AKAHO, "A kernel method for canonical correlation analysis," *International Meeting of Psychometric Society, 2001*, vol. 1, 2001. [Online]. Available: https://cir.nii.ac.jp/crid/1574231874767776512

[18] Y. Wang, S. Cang, and H. Yu, "Mutual information inspired feature selection using kernel canonical correlation analysis," *Expert Systems with Applications: X*, vol. 4, p. 100014, 2019.

[19] X. Yang, W. Liu, W. Liu, and D. Tao, "A survey on canonical correlation analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2349–2368, 2021.

[20] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, p. 395–416, dec 2007.

[21] L. Hardle, Wolfgang; Simar, in *Applied Multivariate Statistical Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Canonical Correlation Analysis, pp. 321–330.

[22] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[23] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017. [Online]. Available: http://arxiv.org/abs/1708.07747