A New Supervised t-SNE with Dissimilarity Measure for Effective Data Visualization and Classification

Laureta Hajderanj
School of Engineering
London South Bank University
London, United Kingdom
hajderal@lsbu.ac.uk

Isakh Weheliye
School of Engineering
London South Bank University
London, United Kingdom
weheliyi@lsbu.ac.uk

Daqing Chen
School of Engineering
London South Bank University
London, United Kingdom
chend@lsbu.ac.uk

ABSTRACT

In this paper, a new version of the Supervised t- Stochastic Neighbor Embedding (S-tSNE) algorithm is proposed which introduces the use of a dissimilarity measure related to class information. The proposed S-tSNE can be applied in any high dimensional dataset for visualization or as a feature extraction for classification problems. In this study, the S-tSNE is applied to three datasets MNIST, Chest x-ray, and SEER Breast Cancer. The two-dimensional data generated by the S-tSNE showed better visualization and an improvement in terms of classification accuracy in comparison to the original t- Stochastic Neighbor Embedding(t-SNE) method. The results from *k*-nearest neighbors (*k*-NN) classification model which used the lower dimension space generated by the new S-tSNE method showed more than 20% improvement on average in accuracy in all the three datasets compared with the t-SNE method. In addition, the classification accuracy using the S-tSNE for feature extraction was even higher than classification accuracy obtained from the original high dimensional data.

CCS Concepts

- **Networks → Network algorithms**

Keywords

High dimensional data; dissimilarity measure; supervised dimensionality reduction; supervised t-SNE; *k*-NN classification.

1. INTRODUCTION

Real-world data such as digital images, government records, gene expression micro array data, customer's transactions contain large number of features. Curse of dimensionality phenomenon is encountered when dealing with such high dimensional space [1]. Aggarwal et al. [2] stated that in high dimensional spaces, the Euclidean distance becomes meaningless as they may contain many irrelevant dimensions that only create noisy data. Furthermore, it is not guaranteed that a data mining model performs well in high dimensional space [3]. As a consequence, a dimensionality reduction technique is considered to be an important step in data analytic. The common dimensionality reduction techniques include Principal Component Analyses (PCA), Isomap [4], LLE (Local Linear Embedding) [5], Laplacian Eigenmaps [6] and t-SNE [7]. PCA works well with linear data structures whereas the other three can handle complex non-linear real-world data by preserving the intrinsic geometry of the high dimensional dataset. However, Isomap and LLE require low dimensional representation to be embedded into a single location in the original data space [8]. So, when the low dimensional representation of the original data is embedded into different locations, these methods do not work well. On the contrary, t-SNE is more suitable to deal with such situation. Although these non-linear methods can handle reducing data dimensionality, the general limitation of all unsupervised dimensionality reduction techniques is that they cannot fit well in classification/regression problems as they lack the class label information. Supervised approaches such as WeightedIso, S-Isomap, Supervised LLE, Supervised Enhanced LLE [8,9,10,11] improved the classification accuracy by introducing different kinds of dissimilarity measures instead of using the Euclidean distance. Similarly, Yu et al. [12] and Cheng et al. [13] proposed supervised t-SNE methods using the class information combined with specific dataset related information (angle information of the Synthetic Aperture Radar (SAR) dataset) [12] and (silhouette frame information of IXMAS and Weizmann datasets) [13]. Consequently, these methods are closely tied to datasets and cannot be generalized to other datasets which do not include the angular or silhouette frame information.

Motivated by the lack of a general supervised t-SNE approach, this paper aims to improve the unsupervised t-SNE algorithm by adding class related information to dissimilarity measure in order to make it suitable for classification tasks.

The remaining sections are structured as follows. In Section 2, other supervised dimensionality reduction approaches are presented. Section 3 describes t-SNE algorithm and presents a new supervised t-SNE approach. Section 4 presents experimental results and discussion. Then a conclusion and future research work is given in Section 5.

2. RELATED WORK ON SUPERVISED DIMENSIONALITY REDUCTION TECHNIQUES

Dimensionality reduction techniques map high dimensional data points to a lower space. Suppose that a dataset is represented by a matrix A with N vectors and each vector a_i $(i \in \{1,2,\ldots,N\})$ has D dimensions, they map the dataset $A_{N\,x\,D}$ into a new dataset $B_{N\,x\,d}$ where $(d \ll D)$and preserve as much as possible from the fundamental information of the dataset A. In this paper, features in the high dimensional data are denoted as input variable A, whereas the features of low dimensional space are donated as output variable B. The variable which contains the label information is denoted as response variable T. For any given i it is denoted with a_i a sample from the high dimensional space A, with

ICSIE 2019, April 9-12, 2019, Cairo, Egypt

ACM ISBN 978-1-4503-6105-7/19/04…$15.00
https://doi.org/10.1145/3328833.3328853

b_i a sample from the low dimensional space B and with t_i the i^{th} sample from T. $dist(a_i, a_j)$ represents the Euclidian distance of two high dimensional data points a_i and a_j.

Supervised dimensionality reduction techniques have found a renewed interest recently with WeightedIso [8] and S-Isomap [9] being proposed as supervised versions of Isomap, for pattern recognition problems. In WeightedIso approach, the same class data points are forced to be nearby one another by scaling their distances with a parameter σ (σ > 1), whereas the distance between points of different classes does not change.

$$WeightedIso = \begin{cases} \frac{1}{\sigma} \text{ dist}(a_i, a_j) & t_i = t_j \\ \text{dist}(a_i, a_j) & t_i \neq t_j \end{cases}, (\sigma > 1) \quad (1)$$

On the contrary, $S - Isomap$ introduced the concept of dissimilarity measure.

$$\textit{S-Isomap}=\begin{cases} \sqrt{1 - e^{\frac{-dist^2(a_i,a_j)}{\beta}}} & t_i = t_j \\ \sqrt{e^{\frac{dist^2(a_i,a_j)}{\beta}}} - \alpha & t_i \neq t_j \end{cases} \quad (2)$$

where β is usually calculated as the average distance over all data points distances [9] and α being the parameter that decreases with α the distance of different class data points [9]. Likewise, supervised variants of LLE were proposed and defined as a Supervised LLE (SLLE) [10] and a Supervised Enhanced LLE (ESLLE) [11]. The SLLE increases different class points distances and keeps the same class points distances unchanged.

$$SLLE = \begin{cases} \text{dist}(a_i, a_j) & t_i = t_j \\ \text{dist}(a_i, a_j) + \alpha \, max(\text{dist}(a_i, a_j)) \, \lambda_{ij} & t_i \neq t_j \end{cases} \quad (3)$$

where $\alpha\epsilon[0,1]$ $and\ \lambda_{ij}$ is 0 if samples i and j are in the same class and 1 otherwise.

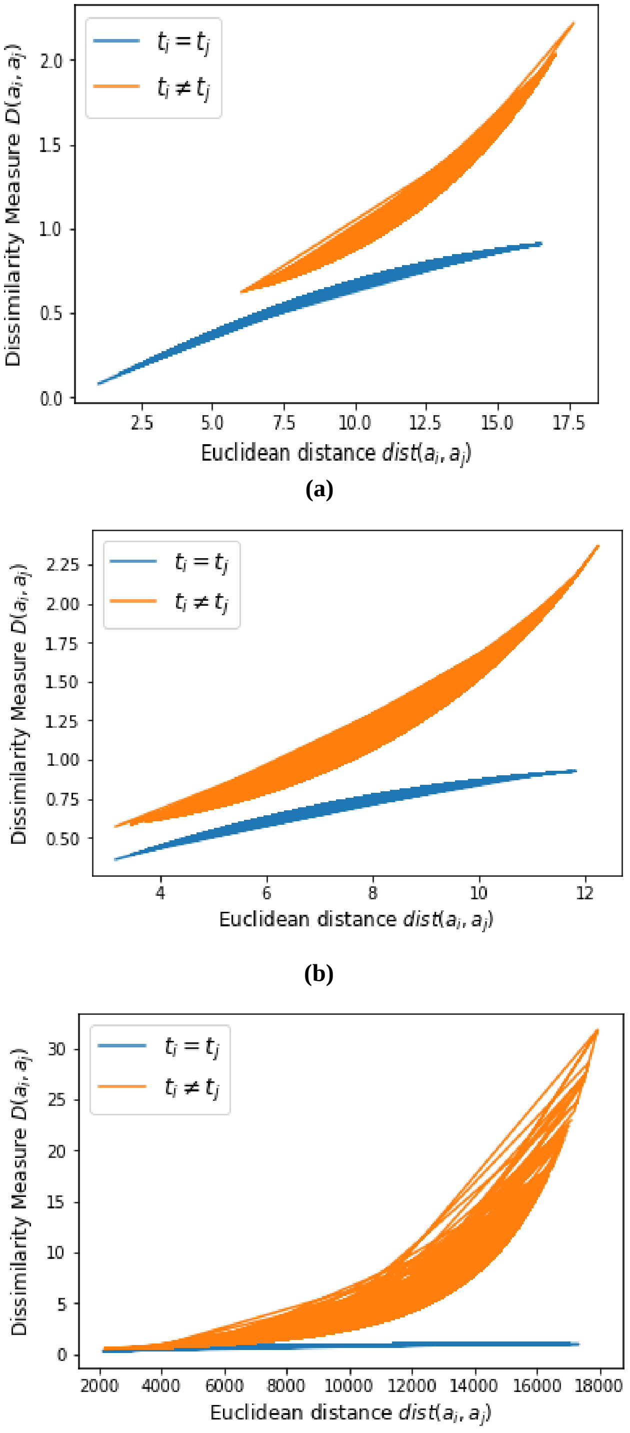
The ESLLE uses the dissimilarity measure as in Eq. (2) to increase the distance of different class points and decrease the distance of the same class points.

In addition, Yu et al. [12] and Cheng et al. [13] proposed supervised versions of t-SNE. In this method, the distances between the different classes are calculated as follows.

$$\textit{St-SNE}=\begin{cases} \text{dist}(a_i, a_j) * exp\left(-\theta(a_i) - \theta(a_j)\right) & t_i = t_j \\ \text{dist}(a_i, a_j) & t_i \neq t_j \end{cases} \quad (4)$$

where $\theta(a_i)$ refers to the angle information [12] and the silhouette frame information [13] of the sample a_i.

The approach proposed in this paper uses the dissimilarity measure of Eq. (2) to calculate the dissimilarity between any two given points. According to Geng et al. [9], there are some good properties using the dissimilarity measure $S - Isomap$. One main advantage is that for the same Euclidean distance of any two given points, the dissimilarity measure of points belonging to the same class is smaller than the points of different classes as it is shown in Figure 1.. This fact makes the use of dissimilarity measure more appropriate in classification problems than the Euclidean distance.



(a)

(b)

(c)

Figure 1. The relationship between Euclidean distance $dist(a_i, a_j)$ and Dissimilarity measure $D(a_i, a_j)$ using MNIST (a), Chest X-ray (b), SEER Breast Cancer (c) datasets.

The advantage of the proposed S-tSNE approach over the Yu et al. [12] and Cheng et al. [13] approaches is that it can be applied to any high dimensional dataset. Consequently, the S-tSNE approach can be considered as a generalized case of the other supervised t-SNE approaches.

3. T-SNE ALGORITHM AND SUPERVISED T-SNE APPROACH

3.1 t-SNE algorithm

t-SNE is a feature extraction technique and an extension of Stochastic Neighbour Embedding (SNE) [15]. Although SNE is good for visualization, its cost function suffers from crowding problem, which is due to the fact that not enough space is accommodated for points far apart in comparison with nearby points in the low dimensional space. For the completeness of the relevant concepts and the following discussions, the details of the t-SNE algorithm are given below as in [7].

T- SNE Algorithm

```
$\boldsymbol{Input}$: $A \in R^{N\,x\,D}$, perplexity $Prep$, number of iterations $h$, learning rate $\eta$, momentum $\alpha$

1) Calculate join probability using $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$, with $p_{j|i} = \frac{\exp(-dist^2(a_i,a_j)/2\sigma_i^2)}{\sum_{k\neq i}\exp(-dist^2(a_k,a_i)/2\sigma_i^2)}$
2) Set initial samples $B^0 = \{b_1, b_2, \ldots, b_N\}$ from $\mathbb{N}(0, 10^{-4}I)$
$For$ h=1 to $\boldsymbol{do}$:
    a) compute low dimensional joint probability
       $q_{ij} = \frac{(1+ ||b_i - b_j||^2)^{-1}}{\sum_{k\neq l}(1+ ||b_k - b_l||^2)^{-1}}$
       $q_{ij}$ between data $b_i$ and $b_j$ in the low dimensional space.
    b) Compute gradient
       $\frac{\delta C}{\delta\, bi} = 4 \sum_j (p_{ij} - q_{ij})\left(1 + \left\|b_i - b_j\right\|^2\right)^{-1}\left(b_i - b_j\right)$
       where
       $C = \sum_i KL(P||Q) = \sum_{i\neq j} p_{ij} \log(\frac{p_{ij}}{q_{ij}})$
       is the objective function of Kullback-Leibler (KL) divergence.
    c) Minimize the objective function using the gradient descent formula
       $\boldsymbol{b^{(h)} = b^{(h-1)} + \eta\frac{\delta C}{\delta b} + \alpha(h)(b^{(h-1)} - b^{(h-2)})}$

$\boldsymbol{Output}$: Embedded data representation $B \in R^{N\,x\,d}$ where $d$ can be two or three dimensional.
```

The t-SNE addresses this issue by using a student t-distribution in the low dimensional space beside Gaussian distribution employed in the SNE [7]. The student t-distribution has a longer tail compared with the Gaussian distribution. This fact helps to map faraway points of high-dimensional space as faraway points in the embedded space as well. Another major difference between the t-SNE and the SNE is that the t-SNE uses a symmetric version of the SNE [7]. The cost function of the symmetric SNE is

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

where $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$.

3.2 Proposed Approach

In the proposed S-tSNE approach, the class information guides the embedding process using the dissimilarity measure instead of the Euclidean distance due to the robustness that this metric offers [9,11]. The dissimilarity measure is calculated as in Eq.(5) which is the dissimilarity measure used in $S-Isomap$ in Eq. (2).

$$D(a_i,a_j)=\begin{cases}\sqrt{1-e^{\frac{-dist^2(a_i,a_j)}{\beta}}} & t_i = t_j \\ \sqrt{e^{\frac{dist^2(a_i,a_j)}{\beta}}} - \alpha & t_i \neq t_j\end{cases} \qquad (5)$$

The conditional probability

$$p_{j|i}=\frac{\exp(-D(a_i,a_j)/2\sigma_i^2)}{\sum_{k\neq i}\exp(-D(a_k,a_i)/2\sigma_i^2)}$$

will be calculated using the dissimilarity measure $D(a_i, a_j)$ instead of Euclidean distance $dist(a_i, a_j)$. The output of $S-tSNE$ can be used for classification problems and for visualizing data with high dimensions.

4. EXPERIMENT AND DISCUSSION

4.1 Data

To evaluate the performance of S-tSNE, three datasets with different dimensions, namely MNIST, SEER breast cancer datasets (Surveillance Epidemiology and End Results) and chest x-rays dataset ChesRay8 were used. MNIST database [16] contains 70,000 handwritten digit images, each consisting of 784 features (28 by 28 pixels). In this experiment, a sample with 5000 images (500 digits of each class) was used. Chest X-ray data [17] contains 108,948 chest X-ray images each consisting of 10000 features (100 by 100 pixels). SEER Breast Cancer dataset [18] contains around 700,000 samples. The dataset contains inconsistent and incomplete data and after pre-processing, the number of samples was reduced to 200,000 and the number of features was increased up to 960.

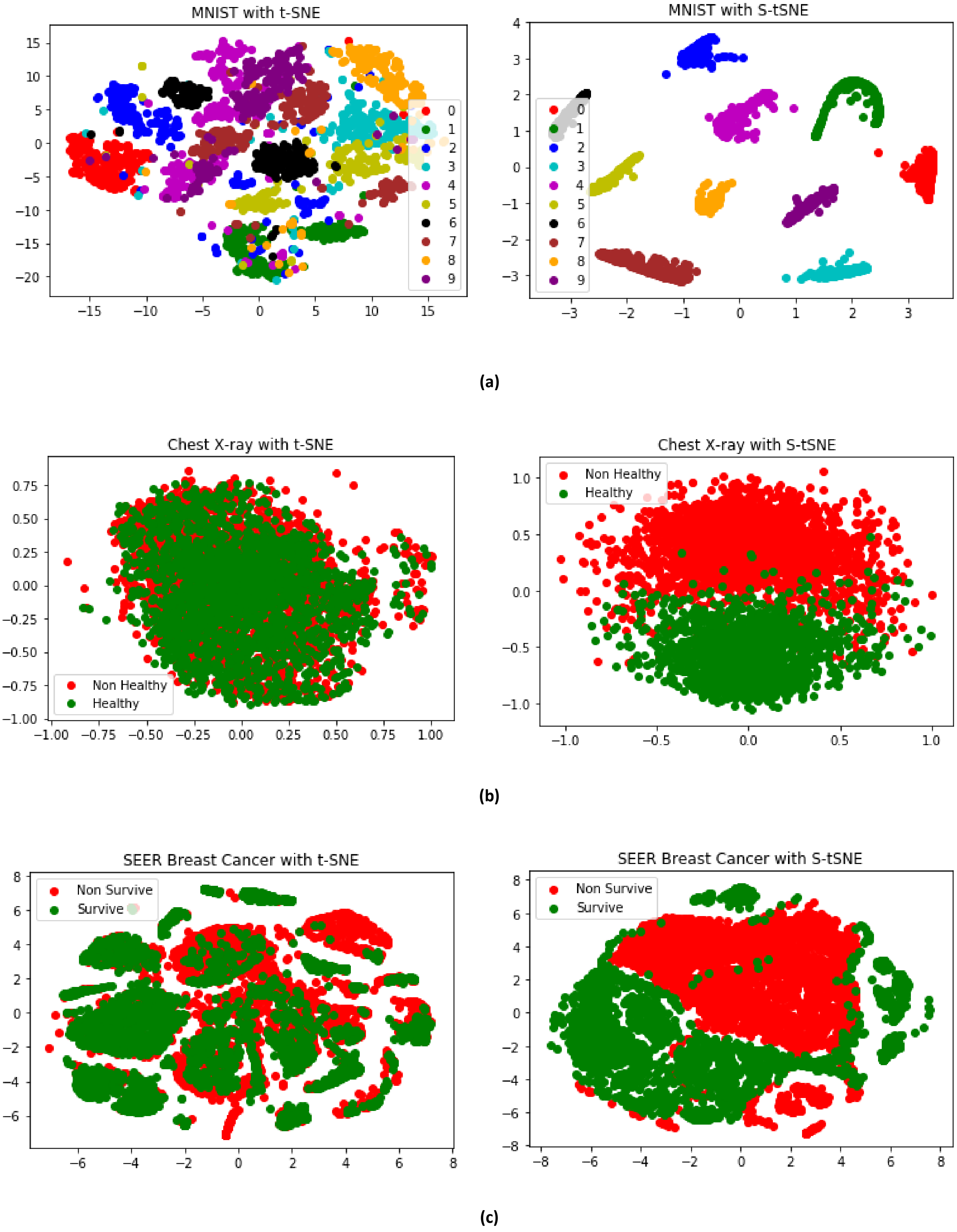
4.2 Experiment

t-SNE and S-tSNE require a lot of computational time to generate the results. This complexity increases drastically with the number of records used. Although the proposed algorithm guarantees to converge to the same result regardless of the size of the sample, 5000 samples were randomly selected from the MNIST and Chest X-ray datasets and 12,000 from SEER Breast cancer dataset for the sole purpose of shortening the convergence time required for the simulations. However, the number of samples for each class is kept approximately the same.

For both methods t-SNE and S-tSNE the perplexity selected was 30 with momentum being equal to 0.5 and the learning rate equal to 100. The number of iterations for MNIST and Chest X-ray was 100 while 1000 iterations were used for SEER Breast Cancer. The figures below show the visualization comparison of the different models.

4.3 Results and Discussions

Both the t-SNE and the S-tSNE methods require time to converge. Both can be used as feature extraction for classification tasks. However, the original t-SNE could lend low discriminant information of the data due to the fact that it is an unsupervised method.



(a)

(b)

(c)

Figure 2. Visualization of two dimensional data generated by t-SNE and S-tSNE in (a) MNIST, (b) Chest X-ray and SEER Breast Cancer datasets

The original data space and the extracted features from both the t-SNE and the S-tSNE were used as an input for k-NN (k=5) classification method. The classification accuracies are shown in table 1.

Table 1. The accuracy of k-NN classification model in high dimensional and low dimensional space using t-SNE and the proposed S-tSNE

Dataset	Original Data	t-SNE	S-tSNE
MNIST(5,000)	95.30 %	74.00 %	100 %
Chest X-ray(5,614)	59.10 %	57.27 %	96.32 %
SEER Breast Cancer(12,000)	81.86 %	77.25 %	98.81%

5. CONCLUSION AND FUTURE RESEARCH

A new supervised S-tSNE was proposed using a dissimilarity measure instead of Euclidean distance. The S-tSNE has been applied to three different datasets including MNIST, Chest X-ray and SEER Breast Cancer. All datasets have a large number of features, ranging from 784 to 10000 pixels.

The S-tSNE has shown great improvement in terms of feature extraction for classification tasks and dimensionality reduction for data visualization, plus it outperformed the original unsupervised t-SNE in both cases. However, both the t-SNE and the S-tSNE require a long convergence time. As a result, accelerating the converging process could be a potential future work.

6. Acknowledgement

This study is partially supported by Acctive Systems Ltd (http://www.acctive.co.uk/) jointly with London South Bank University via scholarship scheme.

7. REFERENCES

[1] Liu, R. and Gillies, D.F., 2016. Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognition, 53, pp.73-86.

[2] Aggarwal, C.C., Hinneburg, A. and Keim, D.A., 2001, January. On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory (pp. 420-434). Springer, Berlin, Heidelberg.

[3] Jimenez, L.O. and Landgrebe, D.A., 1998. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 28(1), pp.39-54.

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

[5] Roweis, S.T. and Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500), pp.2323-2326.

[6] Belkin, M. and Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems (pp. 585-591).

[7] Maaten, L.V.D. and Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), pp.2579-2605.

[8] Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G. and Koudas, N., 2002, July. Non-linear dimensionality reduction techniques for classification and visualization. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 645-651). ACM.

[9] Geng, X., Zhan, D.C. and Zhou, Z.H., 2005. Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35(6), pp.1098-1107.

[10] De Ridder, D. and Duin, R.P., 2002. Locally linear embedding for classification. Pattern Recognition Group,Dept.of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands, Tech. Rep. PH-2002-01, pp.1-12.

[11] Zhang, S.Q., 2009. Enhanced supervised locally linear embedding. Pattern Recognition Letters, 30(13), pp.1208-1218.

[12] Yu,M.,Zhang,S.,Zhao,L.andKuang,G.,2017,April.Deepsupervised t-SNE for SAR target recognition. In Frontiers of Sensors Technologies (ICFST), 2017 2nd International Conference on (pp. 265-269). IEEE.

[13] Cheng, J., Liu, H., Wang, F., Li, H. and Zhu, C., 2015. Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning. Ieee transactions on image processing, 24(10), pp.3203-3217.

[14] Ribeiro, B., Vieira, A. and das Neves, J. C. (2008) Supervised Isomap with dissimilarity measures in embedding learning, in: Iberoamerican Congress on Pattern Recognition, Springer, pp. 389-396.

[15] Hinton, G.E. and Roweis, S.T., 2003. Stochastic neighbor embedding. In Advances in neural information processing systems (pp. 857-864).

[16] LeCun, Y., Cortes, C. and Burges, C. J. (2010) No title, Mnist Handwritten Digit Database.AT&T Labs, .

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. Summers, ChestXray8: Hospital-scale Chest X-ray Database and Benchmarks on WeaklySupervisedClassificationandLocalizationofCommonThoraxDiseases, IEEE CVPR, pp. 3462-3471,2017

[18] Surveillance, Epidemiology, and End Results Program: Overview of the SEER Program (1973-2014). http://seer.cancer.gov/about/overview.html. Accessed 2 August 2017.