Overview of Manifold Learning and Its Application in Medical Data set

Elnaz Golchin and Keivan Maghooli

Department of Biomedical Engineering,Science and Research Branch, Islamic Azad University,Tehran,Iran

ABSTRACT:

The purpose of this study is introduction of new and efficient applications of manifold learning in medical Science and related sciences. Manifold learning is one of the most widely used methods in precise clustering of high-volume data set in data analysis science.In the first; we have a short overview on definition of manifold learning and its main algorithms. Then we describe how to use these algorithms in data mining. In its applications can be cited to database classification of tiny cancer, evaluation of biological pathways of gene ontology, predict brain tumour progression, cage base modelling for all of body movement in the biomechanical science and also processing of brain and heart images. In the following we will describe some of these applications in details.

Key Words:

Manifold learning, LLE, Isomap, High Dimensional Dataset Clustering and Image Processing.

1. Introduction

Identification and assessment of the medical high dimensional dataset such as gene ontology data, oncology[1], MRI image set, consecutive frames of echocardiography images[2] and also used for all video images n order to flow objects[3] and surveymodelling of walking cycle in order to determine motion abnormalities have clinical diagnostic and therapeutic importance. Extraction of useful information from data sets accurately and clearly, has an important role on prevention and treatment of many diseases that are a threat to humanity nowadays. Correct diagnosis of cancer[1],[4] estimation of brain tumour progression[4],[5],[6] detection of general and regional anomalies on the left ventricle muscle in echocardiographic images and ability of survey of many abnormalities in walking cycle[7] are all examples that use of manifold learning algorithms have played an important role on them. Therefore, nowadays by increasing in the rate of these diseases and also by increase in the numbers of mortality of them in all parts of the world, requires careful consideration of the data set that related to these diseases. Hence, the purposes of this article are survey manifold learning and its different algorithms and also explanation of how to use it in different collection of databases. In the first, we have a review on definition of manifold learning and also we express different algorithm of it. Afterward, we will review the latest research based on manifold learning. How to apply manifold learning in different database in the latest researches in the fields of oncology, gene ontology, and diagnosis of brain tumours in MRI

images and estimation of tumour progressions will be evaluated. Accomplished researches and its successful results and treatment of many diseases are indication of the importance of careful survey of manifold learning applications and its expansion and development.

2. Manifold learning concept

2.1. Definition:

In mathematics, a manifold is defined as the set of points that are locally behave like Euclidean spaces. The behaviour is similar to Euclidean spaces means that there is the possibility of attribution ofpeculiarities to these points [8], [9]. After that a manifold according to the local behaviour is similar to which Euclidean spaces are determined [10]. In other words, if the manifold is locally like to R^m, dimension of it defined M. Therefore, M dimensional manifold, locally needs to coordinate M for its description. The most common way to describe the manifold is showing a set of points in aR^n Euclidean space. This action is called embedding manifold in R – space.

2.2. Manifold learning purpose:

The purpose of manifold learning is allocation of manifold with lower dimension to database that the structure is high dimensional manifold. Therefore manifold learning in processing of high-dimensional data will be very effective.

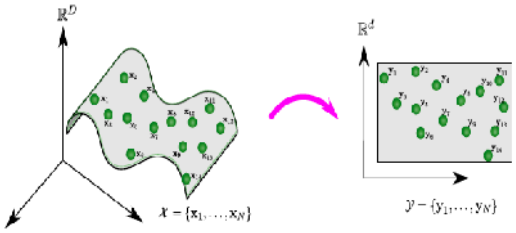

Figure1.Main purpose of manifold learning, reduction of data set [20].

3. Manifold learning algorithm

Manifold learning methods mapping dataset from high-dimensional to low-dimensional such that their intrinsic geometry is preserved.These algorithms are based on the assumption that desired database is on or near the structure of high-dimensional manifold that mapped in the low dimensional space by manifold learning algorithms. These methods used to extract and identify this manifold[11], [12], [13].Manifold learning algorithms are divided in to two global and local categories. In global methods, database mapped from high-dimensional to low-dimensional such that the database global properties are preserved. But in the local methods database are mapped to low dimensional such that local properties preserved. Various types of algorithms are shown in Table 1.

Table**1**. Various Types of Manifold Learning Algorithms

.

General Manifold Learning Types	Isomap[11]	MDS[14]	Kernel PCA[14]	FastMVU[14]
Local Manifold Learning Types	LLE(Local Linear Embedding) [12],[1]	Laplacian EigenMaps [13]	Hessian LLE [8],[16]	LTSA [15]

\

Among the major algorithms can be noted to,Isomap[11]and local linear embedding That they will described in the following.

3.1.Isomap:

Isomap algorithms is one of the first manifold learning collection algorithms that introduced in the journal of science in 2000[11]. This algorithm is among the global methods which pay to reduction of non-linear dimension by preservation of geodesic distance[17],[18]. In other word, the main properties of this algorithm is preservation of geodesic distance[19] between far points and similar points when they mapped from high-dimension to low-dimension[2].mapped sample points to lower space[20].

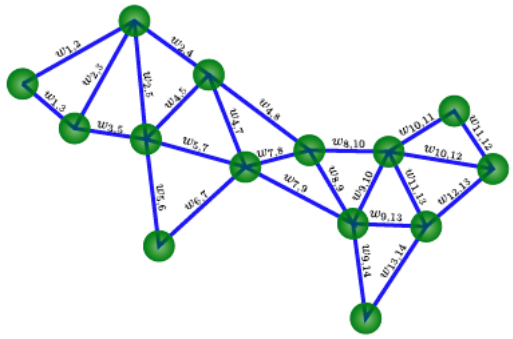

Figure2.Neighbourhood graph and Show ridge on the manifold[20]

3.2. Local linear Embedding:

LEE algorithm is one of the powerful tools of manifold learning methods for reduction of non-linear dimension [1],[10],[12]. In this algorithm dataset mapped from high-dimension to low-dimension when the local structure is preserved. The points that are around thepoint (concept of locality) from neighbourhood weighted graph for this algorithm are selected. For this reason, this algorithm has the locally name.In this method, locality properties of high-dimension manifold dataset are reconstructed by a linear combination of close neighbours and a series of coefficients are obtained. Then every high-dimensional data xi (i=1,2,……,n)constructed from weighted

combination of its k nearest neighbours(Figure3). For accomplishment of this action, E (w) Value function (reconstruction error) expressed in equation 1 must be minimized(7).

$$E(w) = \sum_{i=1}^{n} | x_i - \sum_{j=1}^{k} W_{ij} x_j |^2 \quad (1)$$

Then, in the reduction of dataset dimension are trying to preserved these coefficients(Figure4).

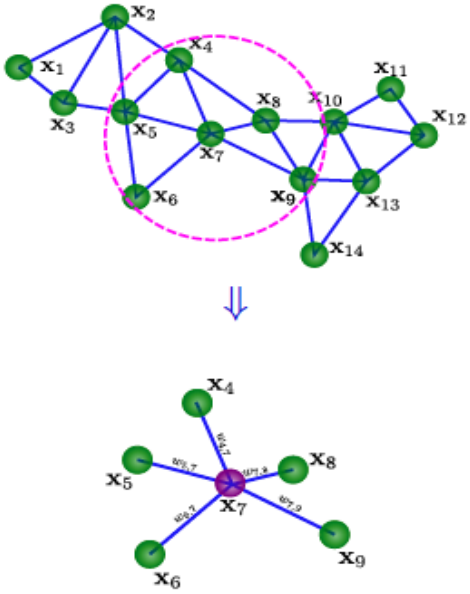

Figure3.the concept of LLE[20].

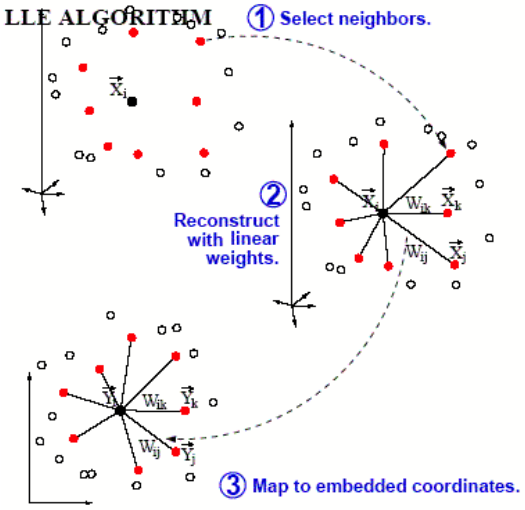

Figure 4. local linear Embedding schematic[21]

3.3. Compare benefits of manifold learning with older methods of reducing dimensions

In the Table2, Isomap and local linear Embedding algorithms are compared with older methods of reducing dimensions. For manifolds that contain curvature and corner, only new Isomap and local linear Embedding algorithms are usable. This ability is one of the most important specificities of manifold learning usage in comparison to older methods. One of the disadvantages of manifold learning algorithm is failure of local linear Embedding usage in noise-polluted database that In this case, the noise reduction pre-processor for the database is required.

Table2.comparison of algorithms[22]

	LLE	ISOMAP	PCA	MDS
Speed	FAST	Extremely Slow	Extremely Fast	Very Slow
Handle Curvature	MAYBE	YES	NO	NO
Handle Corners	YES	YES	NO	NO
Clusters	YES	YES	YES	YES
Handle Noise	NO	MAYBE	YES	YES
Sensitive to Parameters	YES	YES	NO	NO

4. Manifold learning applications

In according to , the correct clustering of dataset structures in shortest processing time are considered, hence , in the recent researches that contain high-dimensional dataset are used from these algorithms. In the following, we mention briefly some examples of these applications.

4.1. Oncology by use of manifold learning:

One of the recent researches in the field of tiny cancer dataset clustering that its results were published in 2013, by someone named Carlotta was conducted at the Polytechnic University in Italy[1]. This scientist investigated the different methods of the most useful manifold learning technique in reduction of cancerous dataset dimensions. The desired data of Carlotta et al are arrays of tiny cancer. The purpose of this research is clustering of data by use of different methods of manifold learning and compare of obtained results in different methods. As previously mentioned, employment of manifold learning methods in the field of oncology is necessary for clustering of too small sizes (at the micro level) cancerous data that has high dimensions in the space.This article in order to avoid overestimate authenticity looking for a sample free method, For this purpose, we use from kernel regression that contain multiple output. Therefore, we extend kernel regression based on regression functions. When apply this expanded method with isometric feature mapping provides an independent visualization and design from data. Mathematical calculations on very small arrays of cancerous dataset indicated usage of Isomap and LLE methods on very small cancerous arrays dataset has higher accuracy in comparison to other methods of manifold learning (Table 3).algorithms except one type of cancer with the name of CML that MVU algorithm has been successful [8].

Table3. Comparison various types of manifold Learning methods on different types of cancerous data sets[8].

Data set	Isomap	LLE	HE	LTSA
Brain Tumor	0.803 0.1 s	**0.838** **0.4 s**	0.715 25 s	0.783 0.5 s
CML	0.649 0.1 s	0.636 0.2 s	0.626 68 s	**0.642** **3.2 m**
CNS	0.700 0.1 s	**0.715** **0.3 s**	0.697 1.7 m	0.672 0.5 s
Castric	0.830 0.1 s	**0.803** **0.2 s**	0.785 19 s	0.790 0.4 s
Lung Cancer	0817 0.2 s	**0.849** **0.3 s**	0.660 1.7 s	0.784 0.6 s
Leukemia	0.956 0.2 s	**0.969** **0.4 s**	0.742 4.1 m	0.911 0.9 s
Medulloblastoma	0.698 0.1 s	**0.721** **0.1 s**	0.695 5 s	0.690 0.2 s

For assurance of correction and fairness of outcomes and also for comparability, the number of evaluation fold was considered invariant. Unlike Bartenhagen et.al claims in 2010 that announced all of data collection must cluster to two patients and healthy class, Carlottas article investigate multiple clustering based on multi-variable nature of purpose. Mean accuracy of total data collection that we repeat ten times fivefold Cross Validation was registered in the following table[1]. We find from achieved results of table 3, mean value of the authenticity of all types of cancer assign a larger number to itself in LLE, Isomap, and LE.

4.2. Evaluation of biological pathways in gene ontology by use of manifold learning:

Nowadays many diseases arise from misplaced and erroneous fluctuation in biological pathways.The main issues in this domain is identification and distinguishing of gene pathways that are currently active from data that describe genes. This study purpose is identification and characterization of active pathways that entered to cellular location ofgenes and began to cooperation. Data set of this study are collection of gene ontology dataset and database of gene descriptor. These data have high dimension therefore in the first they embedded to a form of manifold with low dimension by use of Laplacian eigenmaps[1],[13] and LLE.Then use from model based clustering for identification of active biological data describer pathways. As a result, the study that accomplished in Michigan University in 2011 indicated using of manifold embedding method in extraction of the immune system pathways of macrophage genes expresser data collection.

4.2.1 Stepping of the compound methods of manifold embedding and LaplacianEigenmaps:

1)We provide a matrix K*K(K=168) weighted dimensional matrix[6] (w) from the phrase of (’tnenopmoCralulleC‘) that there is in the collection of gene ontology data [23].2) $W_{i,J}$ weight that had for each pair (i , j) in weighted matrix in the previous step are assigned to each pair (K,2) of genes.Attention: Each weight be nearer to gene is been better.3) We find N number of neighbors

in the fPCA[1] space by use of Euclidean distance.2) We have fourth equation from Laplacian graph.3) We solve the following equations.

$$L_{i,j} = \begin{cases} d_i & if\ i = j; \\ -W_{i,j}\ if\ i\ is\ connected\ to\ j; \\ 0 & otherwise, \end{cases} \quad (2)$$

$$min_y {}^{y^T} L_y = \frac{1}{2}\sum_{i,j}(y_i - y_j)^2 W_{i,j} (3)$$

Denotes a diagonal matrix Implementation results of manifold learning embedding by use of W weighted matrix and MoG clustering methods that are shown in the following Figure[23].

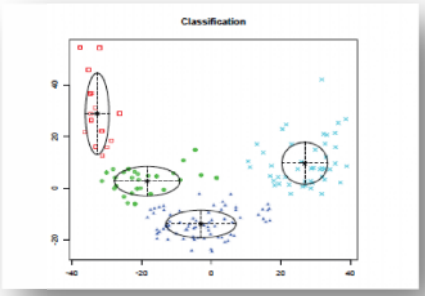

Figure 5. Clustering after dimensions embedding by method of Eigenmaps[23].

4.3. Prediction of brain tumor progression by use of manifold learning algorithm:

Medical imaging department of Texas university by cooperation of Anderson MD oncology center in Texas, presented a new initiating method in manifold learning in 2011 [5] .The purpose of this project is presentation of an appropriate method for accurate diagnosis of low dimensional manifold that related to desired data structure. Desired data collection in this project is essential data collection for diagnosis of brain tumor. Therefore, data collection contains of collections of MRI scans .In this study; they are tried to be able to find a manifold with lower dimension for tumor, recovered and healthy tissues. Moreover, Our most important goal is survey and research in the field of finding of relationship between tumorand healthy tissues. By mapping the bridge between manifolds that related to two consecutive images of MRI can found tumor progression. Hereby, helps to manage the patient's treatment plan. Achieved results from early step of this study guarantee and support the hypothesis presented in this article. Manifolds related to healthy and tumor tissues in the lower space are detachableAlso the manifold that related to tumor tissue progression is found closer to tumor tissue between these manifolds.Total diagrams of this study has been shown in the following Figure in the very simple form[24].

[1] Functional PCA

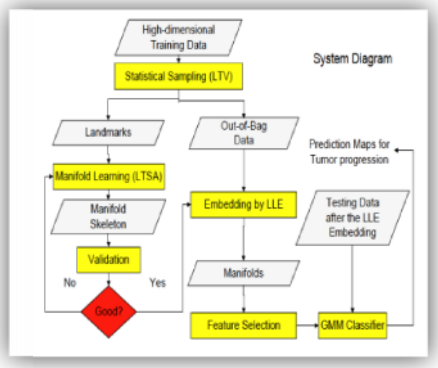

Figure 6 . Block diagram of the application of manifold learning for detection of tumor progression[10].

In this part, we explain some of block of block- diagram. In LTV part that is the stage of sampling, used from landmarks that have been placed in the correct location. A simple schema of landmarks investment are shown in the Figure6.

As you seen in Figure 7, the correct landmarks investment around A-point is more correct than B-point. Because the data changes around A-point are lower than B-point.Then we find neighbors K- nearest for each point of dataset.After this step turn reach to Eigenvectors (special vectors) characterization. We find these vectors by use of Eigenvalues that shown in Figure7 by red color. The average value of the angles of any point of data obtained by other eigenvalues k neighboring points. Validation of the Intake Manifold learning is dependent on several factors. Among these factors, that is parameters selection that selected in this study successfully. Also in this study used from LTSA for finding of correct numbers of neighbors. Moreover, two following functions presented in

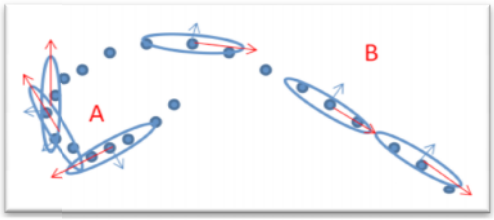

Figure 7. Using Landmarks in manifold Learning[10].

order to cost function (equation5) that in this equation d_m is indication of Euclidean distance and d_M is indication of geometric distance.

$$SF = \frac{\sum_{i,j,i\neq j}\left(d_M(i,j)-d_m(i,j)\right)^2}{\sum_{i,j,i\neq j} {d_M}^2(i,j)} \quad (4)$$

And

$$Acc = \frac{No.of\ correctly\ classified\ training\ data}{Total\ no.of\ training\ data} \quad (5)$$

4.3.1. Embedding by use of LLE:

In Figure8 red colored points are indication of landmarks that these landmarks are indication of total manifold structure. Yellow Colored Square is indication of the point that remains in the new space after implementation of LLE. In next step, we classify embedded data. For doing this action we use MRI images atlas. We put landmarks on the points that contain tumor and healthy tissues. We find general structure of manifold from these landmarks. In the first we define landmarks that indicated tumor and healthy tissues. Then improve general structure of manifold by use of a criterion

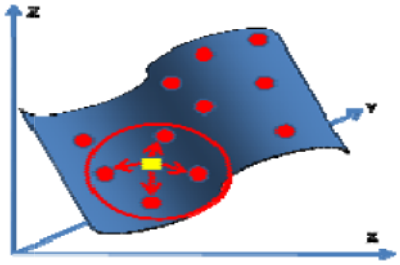

Figure8. Embedding by use of LLE

At the end by GMM model, provide an anticipation atlas by use of landmarks and implementation of trained model for all of interior skull dataset.Figure9 showed the results of this study.

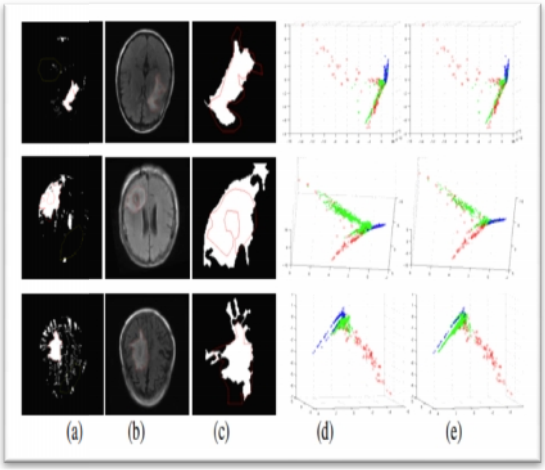

Figure 9. Characterization of tumor location in images.

a)the results of GMM prediction. b)Originalimages.c)the results after processing. d) Three-dimensional scatter plot that red colored points are related to tumor location and blue colored points are related to healthy tissue and green colored points are indication of tumor sample in outside the tumor area that marked in the first observation (in other word, it is tumor tissue that presented as healthy tissue.e)green points that indication of progressed tumor sample. In this study .we find the relation between healthy tissue and tumor tissue in high-dimensional dataset that related to MRI images by use of non-linear manifold learning[5].

5. Discussion and conclusion

Manifold learning is one of the most important and recent methods for reduction of dataset. It contains of different algorithms. In this article described two the most useful and main algorithms. Moreover, in this article described three surveys of recent accomplished studies by using of manifold learning. Then compared the successful results of this study with older methods of reducing dimensions. That is presentation of manifold learning has benefits like more accuracy, the ability of using for high-dimensional datasets domain and dataset with curvature and corner. One of the most important benefits in medical science is the ability of very small cancerous arrays and gene ontology dataset clustering

References:

[1] Carlotta Orsenigo, Carlo Vercellis, (2013), 'A comparative study of nonlinear manifold learning methods for cancer microscopy data classification' Elsevier, EXPERT SYSTEMS WITH APPLICATIONS, 2191-2197.

[2] Ahmad Shalbaaf, Hamid Behnam, Zahra AlizadehSani. (1391)'Present Appropriate Features for Identification of heart abnormalities using Echocardiographic Images'. PhD document, Iran Science and Industry University, Electrical Department, November, 2012.

[3] Pless, R. (2003)'Image/spaces and video trajectories: using Isomap to explore video sequences'. IEEE International Conference computer vision and pattern Recognition (CVPR), Nice, France, 1433-1440.

[4] Zhang, J.Li,S. and Wang, J. (2004), 'Manifold Learning and applications in Recognition', Intelligent Multimedia Processing with Soft Computing, 281-300.

[5] Loc Tran, Deb Banerjee, Xiaoyan Sun, Jihong Wang, Ashok J Kumar, David Vinning, Frederic D.McKenzie, Yaohang Li, and Jiang Li, (2011), 'A Large-Scale Manifold Learning Approach for BrainTumor Progression Predict.

[6] Zhang,Q. Souvenir, R. and Pless, R.(2005)'Segmentation Informed by Manifold Learning', 5thInternational Workshop on Energy Minimization Methods in Computer Vision and Pattern recognition (EMMCVPR), 398-413.

[7] Souvenir, R. and Pless, R. (2007)' Image distance function for manifold learning', Image vision Computer, 25,365-373.

[8] Munkres J.R., Topology, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2000.

[9] Rudin W., Principals of Mathematical Analysis, 3ed.McGraw-Hill, 1976.

[10] Saul, L. and Roweis, S. (2003) 'Think globally, fit locally: unsupervised learning of low dimensional manifolds', J Mach Learn Res, 4,119-155.

[11] Tenenbaum, J.B.S.V. and Langford, J.A.(2000) 'global geometric framework for nonlinear dimensionality reduction',Science, 290,2319-2323.

[12] Roweis, S.T. and Saul, L.K. (2000) 'Nonlinear Dimensionality Reduction by Locally Linear Embedding', Science, 290,2323-2326.

[13] Belkin, M. and Niyogi, P. (2001) 'Laplacianeigenmaps and spectral techniques for embedding and classification', Adv NearalInf Process Syst., 14,585-591.

[14] Borg, I. and Groenen, P. (1997) Modern MultidimetionalScaling: Theory and Applications. Spring. Berlin

[15] Zhang, Z. and Zha, H (2004)'Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment', SIAM J Sci Compute, 26,313-38.

[16] Maaten L.V.D. (2007) 'AnIntroduction to dimensionality reduction using .Matlab Technical Report' MICC – Maastricht University 7-9.

[17] Dijkstra, W. (1959) 'A note on two problems in connexion with graphs', Number Math, 1, 269-271.

[18] Floyd, R.W. (1962), 'Algorithm 97: Shortest Path Common', ACM, 5,345.

[19] Jie, C. Ruiping, W. Shiguang, S. Xilin, C. and Wen, G. (2006)'Isomap based on the Image Euclidean distance' 18th International Conference pattern Recognition(ICPR), Hong Kong, 1110-11130.

[20] Diana Mateus, Helmholtz Zentrum, (2011),'Introduction, Difficulties an Perspectives', 'Tutorial on Manifold Learning With Medical Images' at CAMP, TUM Munchen University.

[21] Tim Doster, John Benedetto, Wojciech Czaja, (2011)'NonLinear Dimentionality Reduction for Hyperspectral Image Classification', 'University of Maryland'.
[22] Todd Wittman, Gilad Lerman, (2005), 'Manifold Learning Techniques: So witch is the best?',Geometric Data Analysis, University of Minnesota.
[23] ArvindRao, Alfred O. Hero, (2011), 'Biological pathway inference using manifold embedding' IEEE-ICASSP 2011, 5992-5995