IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 58-63
www.iosrjournals.org

Dimensionality Reduction Evolution and Validation

Mohammed Hussein Shukur
(Cihan University, Erbil, Iraq

Abstract: *In this paper, proposing visualized and quantitative evaluation methods for validation dimensionality reduction techniques performance. Four well known techniques for dimensionality reduction evaluated, verify the capacity of generating a lower dimensional chemical space with minimum information error. Real chemical database used to generate a sample with specific structure as an input to evolution process. The evaluation is performed in two ways: first visually checking the sample structure, second measuring the local and global distance based error rate that are trained on the low-dimensional data representation. All the evaluation have been done for the four dimensional reduction techniques then result shows weather "trustable" to technique that likely preserves properties with low generalization error and "Un-trustable" which does not preserve the properties.*

Keywords: *Chemoinformatics, Chemical Spaces, Bioinformatics dimensionality reduction, Principal Components Analysis, Kernel PCA, Diffusion Maps, Neighbourhood Components Analysis, evaluation, distance base error .*

I. Introduction

Transforming chemical data in high-dimensional space to a space of fewer dimensions always has been as problematic to preserve structure of the original space[2].Main challenge is to select proper dimensionality reduction that has the ability to reduce the descriptor space significantly but at the same time retain the chemical space local and global structures.

Dimensionality reduction techniques transform high-dimensional data into a meaningful representation of reduced dimensionality. The reduced representation has a dimensionality that corresponds to the intrinsic dimensionality of the data. The intrinsic dimensionality of data has the minimum number of parameters needed to account for the observed properties of the data [1]. There are several techniques to estimate intrinsic dimensionality of the high dimensionality space in lower representative space. The lower representative space selected to be three dimensional spaces. Three dimensional space has more depth than two dimensional space, which it has more distraction. In recent years, a large number of new techniques for dimensionality reduction have been proposed with few number of evolution techniques. Most of the evolution techniques does not necessarily reflect the dimensionality reduction performance, such as Reconstruction Errors [8]. Selecting a proper dimensionality reduction technique that preserves local and global structure features in the lower representative space required evaluation methods that compare distance relationship in the high dimensional and low dimensional spaces.

The outline of the remainder of this paper is as follows. In section 2, dimensionality reduction techniques are reviewed; section 3 describes the datasets used; section 4 describes sampling generation; section 5 validation techniques; and finally in section 6 conclusions are presented.

II. Dimensionality reduction techniques Existing

Main requirement for the dimensionality reduction techniques is the ability to embed the high-dimensional data points into an existing low-dimensional data representation that preserve the local the global structure features. There are several techniques. Techniques fill in General categories into those that are able to detect linear structures, and those that are not. In this paper focus on four well known techniques: principal components analysis(PCA), Kernel principal components analysis(KPCA), Diffusion maps and neighbourhood components analysis(NCA).

1.1 Principal Components Analysis (PCA)

Principle components analysis is a linear transfers a high dimensional space to lower representative space that captures most of the variability in the data, by finding alinear transformation T that maximizes $T^T cov_{x-\bar{x}}{}^T$ TT, where $cov_{x-\bar{x}}$ is the covariance matrix of the zero mean data X.[3]

1.2 Kernel PCA

Kernel PCA (KPCA) is a linear technique same as PCA, but KPCA using Kernel where the linear operations replaced by a reproducing kernel Hilbert space with a non-linear mapping .PCA computes covariance matrix while Kernel PCA computes the principal eigenvectors of the kernel matrix.[12]

1.3 Diffusion Maps

The diffusion maps (DM) is a non-linear technique based on defining a Markov random walk on the graph of the data. Deploying the random walk for a number of time steps, a measure Euclidean distance between points approximates the diffusion distance, and then dimension of the diffusion space is determined by the geometric structure.[5][7][10]

1.4 Neighbourhood Components Analysis

The neighbourhood components analysis is a local linear supervised distance metric learning technique based on learning a Mahalnobis distance measure for k-nearest neighbours. NCA extend the nearest neighbour classifier toward metric learning. By projection of vectors into a space that distance metric optimizes criterion related to the leave-one-out accuracy of a nearest neighbour classifier on a training set.[4]

In this paper, dimensionality reduction techniques deployed by using Matlab Toolbox for Dimensionality Reduction (v0.7.2)[9], parameter settings described in Table 1.

No	Technique	Parameters
1	PCA	None
2	Kernel PCA	K(,,.)
3	Diffusion maps	σ = 1 t = 1
4	NCA	None

Table 1: Dimensionality Reduction Techniques parameter settings.

III. Data set

The chemical dataset are collected from ZINC website [15]. 498,711 molecules are obtained after ensuring that there are no duplicates and also by deploying the Lipinski's Rules [6]. Molecular descriptors [13] are generated for each molecule by using CDK library. Data pre-processing on the molecular descriptors is accomplished to avoid distortion of the reference space because of over-representation and overlapping nature of the data; by eliminating descriptors with more than 80% zero values, and select only representative descriptors from the set of correlated descriptors by using reverse nearest neighbor (RNN) clustering procedure [11]. Then select set of molecular descriptors that have intrinsic dimensionality estimation near to three. The data set ended up with three different datasets, organized based on molecular descriptor categories, namely, 2D descriptors (41 numbers), 3D descriptors (40 numbers), and E-state descriptors (24 numbers). Each of the datasets had a size amounting to 498,711 molecules.

IV. Sampling Generation

Sampling is the process of selecting a set molecular that contain some specific descriptors details. The purpose of these realistic samples is to validate dimension reduction techniques. Each sample collected by finding five different types of molecules groups that has specific properties of real chemical space, instead of the entire population, using the samples to have more clarity about which dimension reduction technique preserves the four chemical space properties. Sampling size is 100 molecules that cover the following

1. Ten molecules represent a lower band in the original chemical space
2. Ten molecules represent a upper bound in the original chemical space
3. Three different groups, each group containing 6 to 7 similar molecules (similarity more than 80% measured by Tanimoto coefficient [14]).
4. Eleven molecules represent a diagonal line, containing information itself
5. Fifty molecules as Random different molecules.

Selecting the specific molecules that contain specific descriptors value range; by using Fingerprint representation and similarity coefficient. Fingerprints are the bit vector encodings of a molecule. It is a length of bit string depended on how many descriptors used like for each division of molecular descriptors and each position corresponding to one molecular descriptor same as key-type fingerprints. First apply scale normalization to the data sets to be of between (0 to 1) then round up the numbers to take one decimal place value such as 0.1, 0.2, 0.3…. 1. 0

For generating the lower bound group done by finding the molecules that has all their descriptors values for normalized data between (0 to 0.3), While the upper bound group contains molecules their descriptors values for normalized data as between (0.7, 1).

For generating the three different groups each group contains similar molecules, for each group, first select one molecule then generate fingerprint representation of it, then use Tanimoto coefficient through the database to find 5 other molecules which are similar to it with 90% similarity.

For generating Diagonal group, which contain molecules their descriptors values are the same.

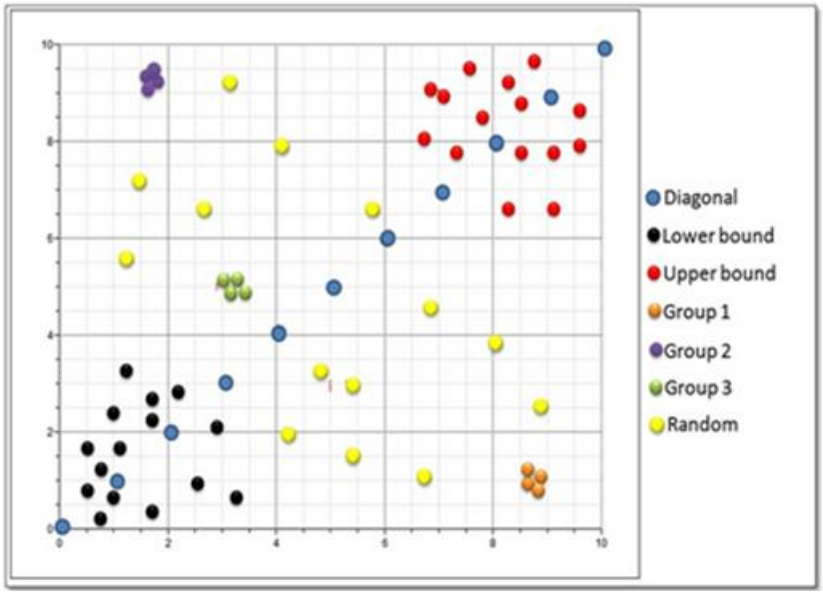


Figure 1: Sample Structure

V. Evolution and Validation techniques

1.5 Visualization

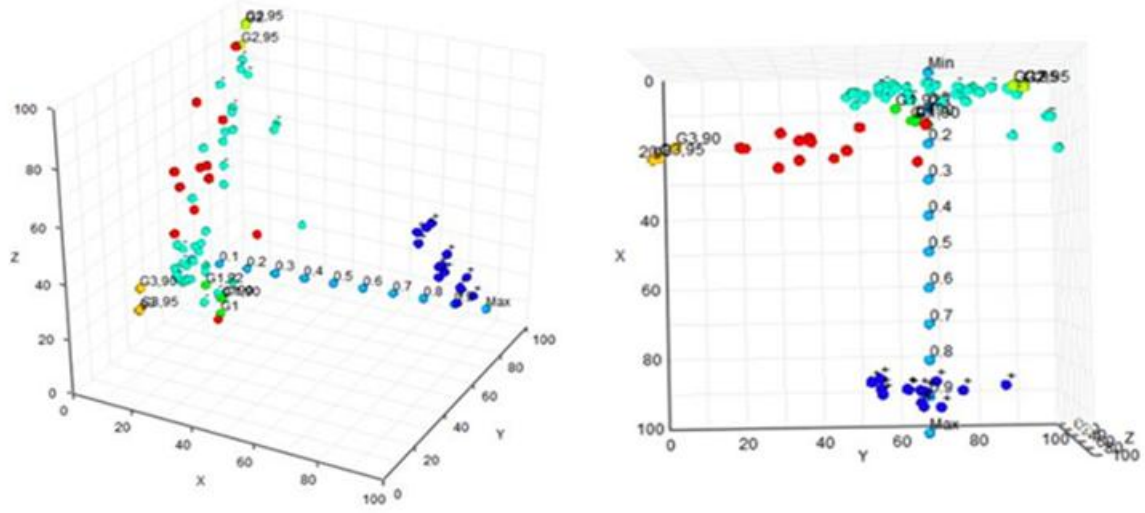
The first impulse of the human is to evaluate the dimensional reduction visually; this kind of validation has cons and pros. Its positive aspect is that it is natural and therefore an easy form of measurement, its limitation however is that validation is accuracy relative to the individual (subjective) and consequently not a precise process. Beginning with building a sample containing 6 different groups of which each group has its own similarity i.e. the lower bound has to be in the lowest of the reduced chemical space, it is supposed to contain the entire group and is located at the lower bound of the Diagonal, while the other groups are relatively at a higher bound. They should, therefore be in the opposite side of the chemical space: a diagonal group of molecules; it should be seen as a sloping line in the reduced chemical space or 3 different groups of the molecules; each group has high similarity to one centralized molecule, with some extra random molecules (see table 2).

Figure 2, shows the side view and top view of chemical space reduced by NCA. While figure 3, shows the side view and top view of chemical space reduced by diffusion maps.

No	Tech./ Features	Diagonal	upper	Lower	Similar Group
1	PCA	✓	✓	✓	✓
2	KernalPCA	X	X	X	X
3	DiffusionMaps	X	✓	✓	✓
4	NCA	✓	✓	✓	✓

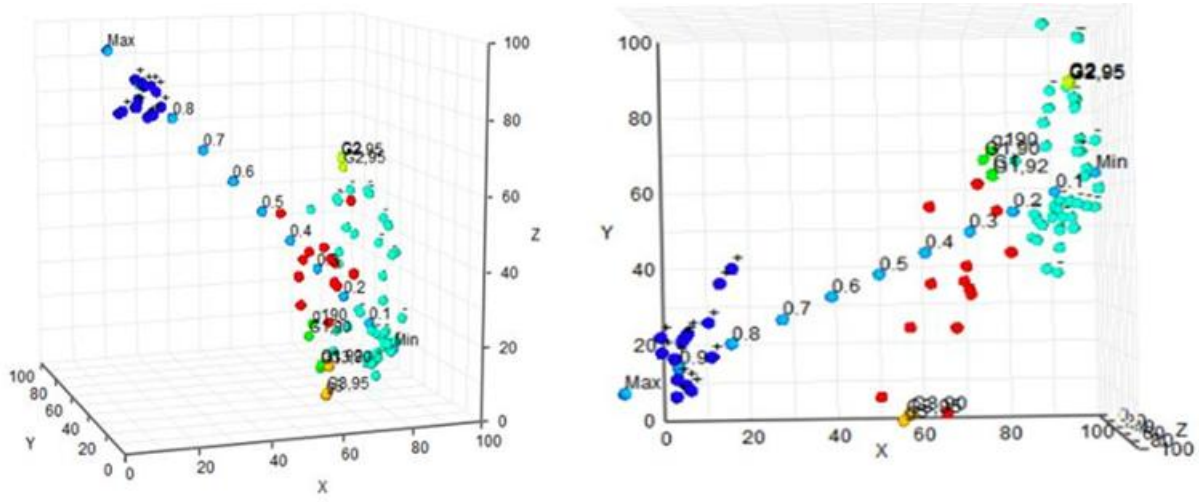
Table 2: Result of the Visualize validation

✓ Indicates feature likely to be preserve while X Indicate not.



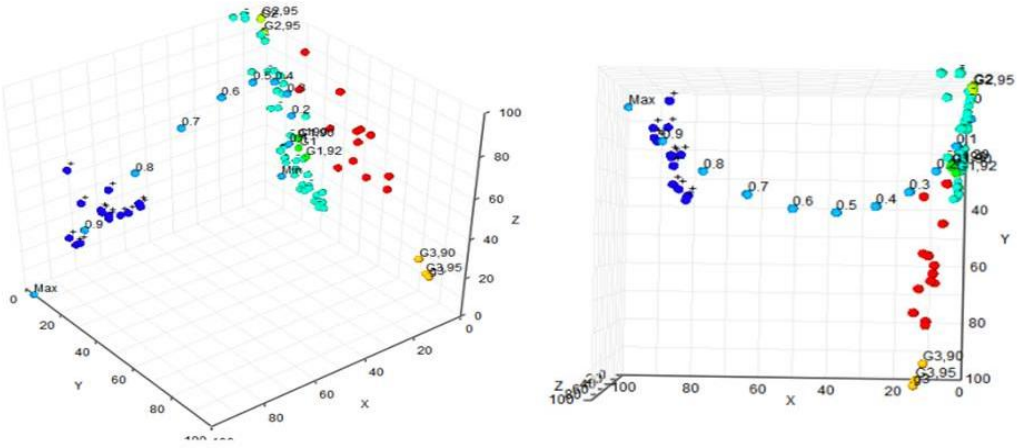
A B

Figure. 2. A Side view and B Top View of chemical space implemented by PCA



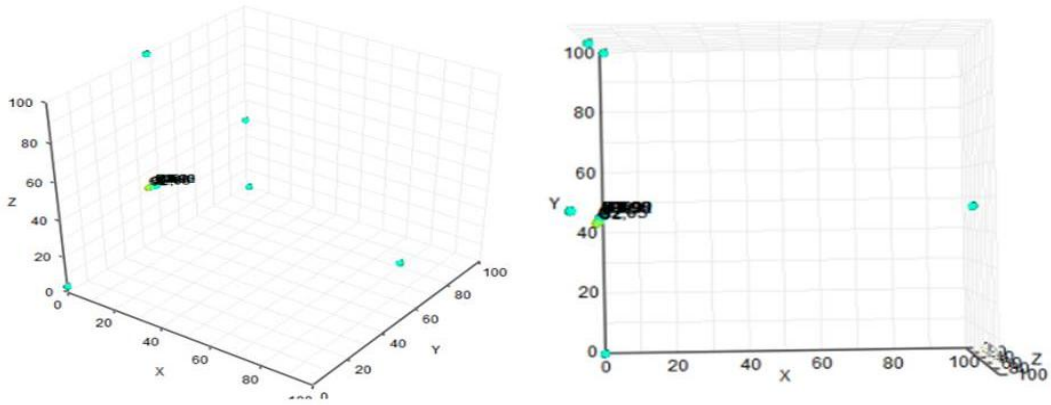
A B

Figure. 3. A Side view and B Top View of chemical space implemented by NCA



A B

Figure 4. A Side view and B Top View of chemical space implemented by Diffusion maps



A B

Figure. 5. A Side view and B Top View of chemical space implemented by KPCA

1.6 Distance Based Error Rate

Distance based error rate is a quantitative evaluation of the local structure of the data in order for successful projection of data, local and global structures need to be retained. An evaluation of the quality is based on Global distance based error rate (GErr) and Local distance based error rate (LErr).
The basic equation for both the Global and Local distance error rate:

$$Err = \frac{\sqrt{(x_{ij} - \bar{x}_{ij})}}{} \times 100 \qquad (1)$$

Where $\bar{x}_{ij}$ is the distance between data points i and j in the low dimensional space, x_{ij} is the distance in the high dimensional space between two data points i and j, where data point i is the reference.
The nuanced differences within the equation in relation to the Global and Local Error rates are explored as following.

Global Distance Based Error Rate

Presenting the Global distance based error rate on the low-dimensional data representations obtained from the dimensionality reduction techniques. The Euclidean distance of two points in scale-normalized high dimensional space is a contribution of all the dimensions, such as 2D Molecular descriptors (41), which translates to the longest distance between two points; the initial point and the highest point, are at a distance of $\sqrt{41}$=6.403. On other side in the scale-normalized low dimensional space the longest distance between two points is √3=1.732. As the ratio between distance in high Dimensional space and distance in low Dimensional space is 3.697. After obtaining the Euclidean distance for a reference point and other points in both the high and low dimensional spaces, divided result distance from the high dimensional space by 3.697.

Deploy the four dimensionality reduction techniques for all parameter settings described in Table1, and for each technique, report the Max. and Min. global distance based error rate of all runs in Table 3. The best performing technique for each dataset is shown in boldface.

	PCA	NCA	KPCA	Diffusion maps
Max	41.09%	**13.82%**	81.47%	41.54%
Min	41.09%	**13.82%**	23.27%	1.35%
Average	41.09%	**13.82%**	60.12%	22.16%

Table 2. Global Distance Based Error Rate (smaller numbers are better)

Local Distance Based Error Rate

Presenting the local distance based error rate on the low-dimensional data representations obtained from the dimensionality reduction techniques. Local distance based error rate has the same equation used in global distance based error rate as well as the Euclidean distance for measuring the distance between the reference point and other points in the both high and low dimensional space, but scale-normalize the distances for each technique result according to the longest distance among the group. i.e NCA group contains distance of 10 molecules with reference to one molecule, Max distance in that group is 1.466041, Min is 0. After normalization max distance will be 1. For each technique result, has been done the same as well as the original data. Then we compute the Error rate.

Report the Max., Min. and Average Local distance based error rate of all runs in Table 4.The best performing technique for each dataset is shown in boldface.

	PCA	**NCA**	**KPCA**	**Diffusion maps**
Max	0.0030%	0.0042%	166.2%	87.87%
Min	0.0000%	0.0000%	0.65%	0.0 %
Average	**0.0006%**	**0.0006%**	43.68%	37.45%

Table 3. Local Distance Based Error Rate (smaller numbers are better)

VI. Conclusions

This paper proposing two ways of validation and evaluation dimensionality techniques performance, using well-structured sample data with size of 100 molecules as input for four well known dimensional techniques, then evaluation is performed in two ways: first visually checking the sample structure, second measuring the local and global distance based error rate that are trained on the low-dimensional data representation. All the evaluation have been done for the four dimensional reduction techniques then result shows weather "trustable" to technique that likely preserves properties with low generalization error among others and "Un-trustable" which does not preserve the properties.

These results have potential implication for various critical chemoinformatic issues such as diversity, coverage, representative subset selection for lead compound identification, etc. Use of a lower dimensional representative space can speed up the virtual screening throughput of large chemical libraries. Virtual screening is a crucial step in drug discovery applications. The current proposal is suitable for integration into any cell-based approaches that are employed in design of targeted leads or diverse chemical subset selection applications in chemoinformatics

References

[1]. K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press Professional, Inc., San Diego, CA, USA, 1990.
[2]. J. Gastring (2003). Handbook of Chemoinformatics Vol(1): pp 231-251.
[3]. H. Hotelling (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24,pp417–441.
[4]. J. Goldberger, S. Roweis, G. Hinton and R. Salakhutdinov, "Neighbourhood Components Analysis," Advances in Neural Information Processing Systems, vol. 17, pp. 513-520, 2005.
[5]. S. Lafon, A.B. Lee (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9), pp1393–1403.
[6]. C. Lipinski, A. Hopkins (2004). Chemical space and biology Nature, 432, 855-861.
[7]. M.H. Law and A.K. Jain. Incremental nonlinear dimensionality reduction by manifold learning. IEEE Transactions of Pattern Analysis and Machine Intelligence, 28(3):377–391, 2006.
[8]. L. Maaten, E. Postma and J. Herik (2009). Dimensionality Reduction: A ComparativeReview
[9]. L. Maaten (2007). An Introduction to Dimensionality Reduction Using Matlab, Technical Report MICCIKAT 07-07.
[10]. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems, volume 14, pages 849–856, Cambridge, MA, USA, 2001. The MIT Press.
[11]. Singh, H. Ferhatosmanoglu and A. Saman (2003). High Dimensional Reverse nearest Neighbour Queries. " Proc. Conf. Information and Knowledge Management (CIKM).
[12]. Shawe-Taylor and N. Christianini. Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK, 2004.
[13]. R. Todeschini and V. Consonni (2009). Molecular Descriptors for Chemoinformatics. Wiley-VCH; 2nd, Revised and Enlarged Edition edition, Vol(1)(2). pp 39-77
[14]. Y. Wang, J. Bajorath (2009). Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching. J. Chem. Inf. Model., 49: 1369 -1376.
[15]. ZINC Database: http://zinc.docking.org/