Visualization of Chemical Space using Kernel Based Principal Component Research

B.Firdaus Begam, J.Rajeswari

Abstract— Principal Component analysis (PCA) is one of the important and popular multivariate statistical methods applied over various data modeling applications. Traditional PCA handles linear variance in molecular descriptors or features. Handling complicated data by standard PCA will not be very helpful. This drawback can be handled by introducing kernel matrix over PCA. Kernel Principal Component Analysis (KPCA) is an extension of conventional PCA which handles non-linear hidden patterns exists in variables. It results in computational efficiency for data analysis and data visualization. In this paper, KPCA has been applied over dug-likeness dataset for visualization of non-linear relations exists in variables.

Keywords: Principal Component Analysis, Kernel Principal Component Analysis, Visualization.

I. INTRODUCTION

Cheminformatics is a distinct research area of computational molecular biology. It is an interdisciplinary field which sandwiches various disciplines like mathematics, physics, chemistry, statistics, biochemistry and informatics.

According to M.Hann [1] the term cheminformatics can be referred as "new name to old problem". It deals with molecules to understand its structure, physical/biological properties and reaction with other molecules [2-3]. Molecular data collected or acquired through observation and experiments are stored, retrieved and analyzed by means of informatics.Features calculated based on molecular structures and representations are known as molecular descriptors. Calculated molecular descriptors of each molecule are stored in multidimensional space, termed as chemical space. Chemical data areanalyzed by various exploratory and confirmatory methods along with dimensionality reduction methods which can more impact on performance analysis [4-5]. A subset which contains few number of features which can define a large set of data, known as dimensionality reduction.

Visualization of data from higher dimensional space to lower dimensional space results in identification of hidden patterns of data and irrelevant data which affects the efficiency of data [6]. The issues arise due to higher dimension was referred as "Curse of Dimensionality"[7].

Data visualization through dimensionality reduction methods can be grouped under linear and non-linear methods. Linear combinations of variables are used to transform higher space to lower space. Principal Component Analysis (PCA)[8-9] and Multi-Dimensional Scaling (MDS) are linear methods. PCA is one of the most important algorithms in multivariate analysis. Nonlinear relationship in dataset are identified by various nonlinear dimensionality reduction [10] methods like, Isometric Feature Mapping (Isomap) [11], Local Tangent Space Alignment (LTSA) [12], Kernel Principal Component Analysis (Kernel PCA) [13]. In this paper, Kernel PCA are used over QSPR drug-likeness dataset to visualize chemical space over lower dimensional.

II. KERNEL FUNCTION

Kernel function which calculates dot product of mapped instances in chemical spaces allows identifying hidden pattern, feature extraction and dimensionality reduction. Kernel function is a symmetric function where, $K(X_iX_j) = K(X_jX_i)$for all X_iX_j.Various types of kernel functions [14-15] are linear, polynomial and Gaussian Radial Basis Function are shown in equ (1) (2) and (3),

$$K(X,Y) = X.Y \quad (1)$$

$$K(X,y) = (1 + X.Y)^p \quad (2)$$

$$K(X,Y) = e^{\frac{-|X-Y|^2}{2\sigma^2}} \quad (3)$$

Linear kernel function simply dependents on input chemical space. Polynomial kernel function maps input chemical space to a chemical space of dimensionality$O(D)^p$, where p represents degree of polynomial. Gaussian RBF method maps input chemical space to indefinite dimensionality sphere.

III.MATERIALS AND METHODS

Physical properties of chemical compounds like melting point, boiling point, density, drug-likeness and other properties that are closely related to physical structure of molecule. Chemical and biological reactivity/activity can be determined by molecular structure. Mathematically it can be term physical structure of a molecule as independent variable and physical properties, chemical and biological activities as dependent variables. Relationship between molecular structures and property/activity are known as Quantitative Structure Property Relationship (QSPR) / Quantitative Structure Activity Relationship (QSAR).

A. Data Set

This paper focuses on visualizing drug-likeness of molecules from Maybridge database. The dataset contains

Revised Manuscript Received on September 10, 2019.

Dr.B.FirdausBegam, Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India.
(E-mail: firdhz.2002@gmail.com)
Dr.J.Rajeswari, Assistant Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India.
(E-mail: rajeswarikrishna82@gmail.com)

14,400 molecules. Over 2000 molecules are taken to identify the hidden patterns present in drug-likeness chemical space. The property of drug-likeness is calculated mainly based on Lipinski Rule of Five (RO5) [17].

Lipinski RO5 mainly deals with molecular properties or descriptors like, molecular weight of molecules which are less than 500 dalton , calculated logP descriptors which are not more than 5 number, hydrogen donors of molecules which have maximum 5 donor bonds and hydrogen acceptors of molecules which have maximum 10 acceptor bonds [18]. The properties of Lipinski rules are of multiples of 5, it is referred as Rule of Five Number of rotatable bonds, flexibility of molecules and polar surface area (PSA) descriptors are used along RO5 descriptors to visualize the molecules using kernel PCA which shows non-linear structure of data present in dataset.

B.Principal Component Analysis (PCA)

Conventional principal component analysis is a multivariate applied over correlated variables. It gives a subset of variables (molecular descriptors) from a larger set of variables which shows or contains information as in original set of variables. It can be termed as; converting correlated set of variables into smaller or less number of uncorrelated variables using linear combination of variables with maximum variance is called as principal components.

For the given data matrix$m \times n$ where m represents number of molecules and n represent number of descriptor generated based on structure of molecule. Variability of data set is preserved in principal component analysis, which considers linear combination of variables. Dependence of molecular descriptors is represented by covariance matrix as in Equ (4).

$$C = \frac{1}{n-1} X^T X \qquad (4)$$

Covariance matrix is symmetric and variances among descriptors are represented in diagonal of symmetric matrix as in Equ.(5).

$$C_{ij} = \frac{1}{n-1} \sum_{i}^{n} (X_{ij} - m_j)^2 = v_i \qquad (5)$$

Common and unique variances among descriptors are principal components or known as Eigen vectors. Characteristic roots of covariance matrix are Eigen values which measures the variances in all variables (molecular descriptors). From covariance symmetric matrix Eigen values and Eigen vectors are calculated as in equ (6),

$$[EigVal, EigVec] = Eig(C) \qquad (6)$$

Eigen values represent principal components and Eigen vectors provide weights to calculate the uncorrelated principal components based on variance existing between the descriptors. The Principal components are ordered based on maximum total variance of variables corresponding to its Eigen vectors. The projection matrix is calculated as in equ (7), which project data over new orthogonal axis.

$$Proj.mat. = X \times EigValues \qquad (7)$$

C. Kernel Principal Componenet Analysis (KPCA)

Extension of PCA using kernel trick for nonlinear transforms are known as Kernel PCA (KPCA).High Dimensional data precisely increases computational time. For the given data matrixX_k (as in Fig.1) where k=1,….l, $X_k \in R^N$,$\sum_{k=1}^{n} X_k = 0$. Computation of molecular descriptor (feature)$\emptyset(X_1 \ldots . X_n)$ is done over chemical space F, $\emptyset: R^N \to F, x \to X$.

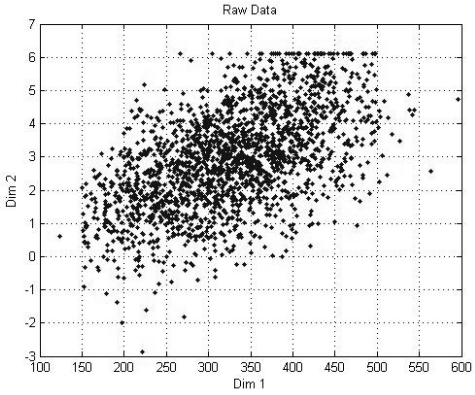

Fig 1: Visualization of Raw Data

In kernel PCA principal components of F are computed to map $\emptyset(X)$ to implicitly defined feature space computed based in inner product which is known as Gaussian kernel function as in equ (4).

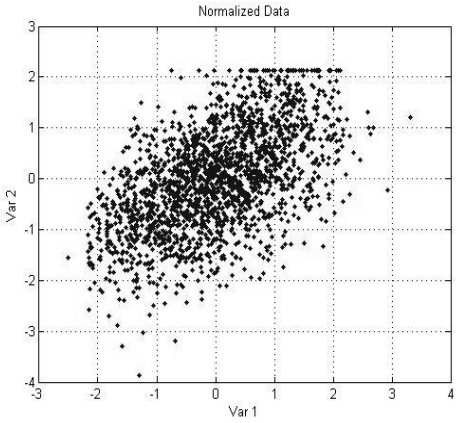

Fig 2: Preprocessing – Normalization of data

It is important for kernel matrix to undergo preprocessing in order to standardize or normalize matrix (equ (8)). The features or molecular descriptors are centralized by means of mean-centering approach (Fig.2).In this approach, mean of data is calculated and it is subtracted from each column of data.

$$\emptyset_i \leftarrow \emptyset_i - \frac{1}{N} \sum_{j=1}^{j=n} \emptyset_{ij} \quad (8)$$

Inner products are recomputed from centered matrix, $K_{ij} = \emptyset(X_i).\emptyset(X_j)$are calculated as in equ (9),

$$K_{ij} \leftarrow K_{ij} - \frac{2}{N} \sum_{k} K_{kj} + \frac{1}{N^2} \sum_{kl} K_{kl} \qquad (9)$$

After centering data, $\sum_{i=1}^{n} \emptyset(x_k) = 0$,calculating covariance matrix over chemical space F as in equ (10),

$C = \frac{1}{n} \sum_{j=1}^{n} \emptyset(x_j) \emptyset(x_j)^T$ or

$$C = \frac{1}{n} \sum_{j=1}^{n} \emptyset^T . \emptyset \qquad (10)$$

To represent data in lower dimensional space, dominant Eigen vectors of kernel matrix to be computed, $K_{ij} = \emptyset(X_i).\emptyset(X_j)$. Eigen values ($\lambda_\alpha$) and Eigen vectors ($V_\alpha$) are calculated as in equ (11),

$$K = \sum_{\alpha} \lambda_\alpha V_\alpha V_\alpha{}^T \qquad (11)$$

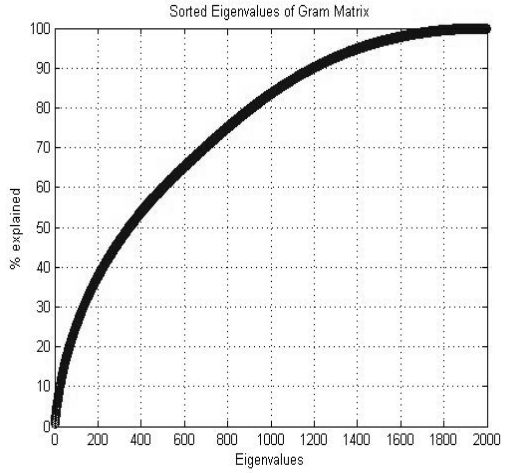

Fig: 3 Sorted Eigen Values of Gram (Kernel) Matrix

Most dominant Eigen value of kernel matrix measures the principal component which has more variance in chemical or feature space (Fig.3). Based on Eigen values and Eigen Vectors, kernel principal components are calculated by applying Eq (13),

$$y_k = \emptyset(X)^T V_k = \sum_{i=1}^{N} a_{ki} K(X, X_i) \qquad (13)$$

where $a_k = [a_{k1} a_{k2} \ldots . a_{kN}]^T$, a_k is the N dimensional column vector of a_{ki}. The gram matrix is represented as in equ(14),

$$\tilde{K} = K - 1_N K - K 1_N + 1_N K 1_N \qquad (14)$$

Where 1_N is the $N \times N$ matrix with all elements equal to $1/N$.

IV.RESULTS

The calculated KPCA components are projected on score plot which represents non-linear patterns present in data variables Principal components (PC) are calculated based on total variance existing between the variables in dataset. First PC (PC1) represents the maximum variance present in variables. Second PC (PC2) represents the next maximum variance exists in variables. The variance present in PC1 is not accounted for calculating PC2. Principal components with minimum variance will have minimum or least effect on analysis; such components can be excluded for further analysis, which reduces the dimension of the dataset.

The principal components calculated based on Gauss Radial function are geometrically represented in score plot. Variables which show greater variance in non-linear combinations deviate from axis origins which are considered as untidy data can be removed from chemical space which results in dimensionality reduction.

The plot in Fig.4(a), Fig.4(b) and Fig. 5(a), Fig. 5(b), represents two and three dimensional representation of data in new orthogonal axis

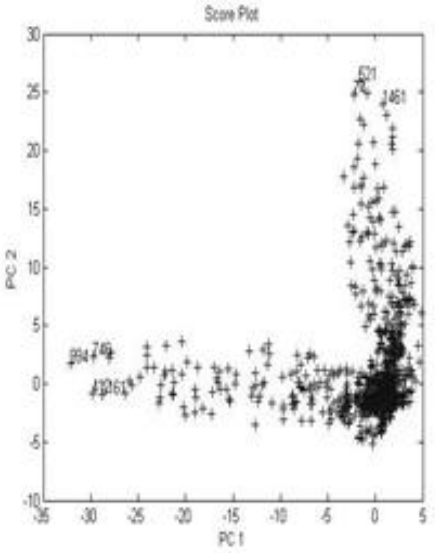

Fig. 4(a). KPCA Projection 2D(PC1,PC2)

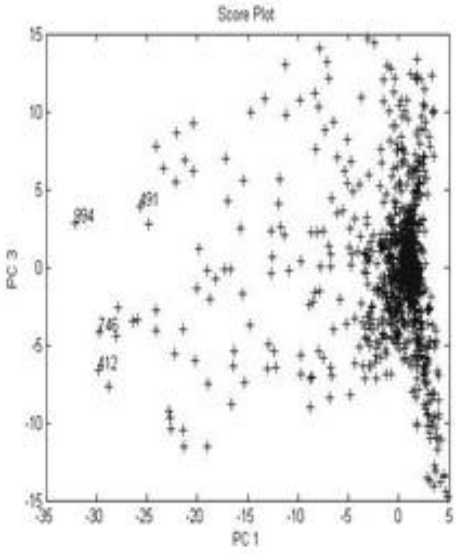

Fig. 4(a). KPCA Projection 2D (PC1,PC3)

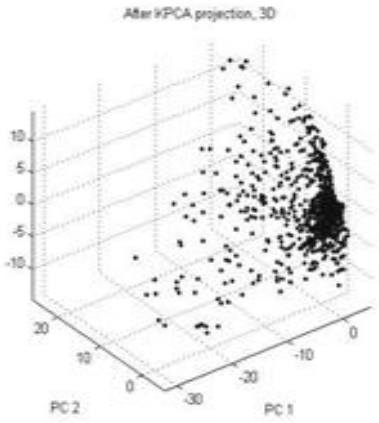

Fig. 5. KPCA Projection 3D (PC1,PC2)

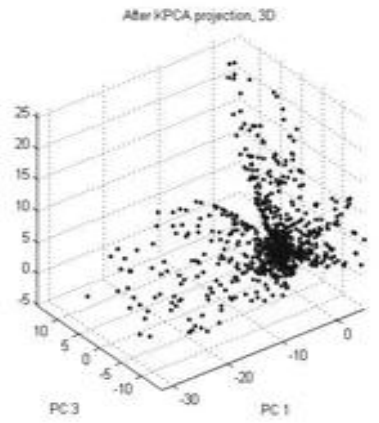

Fig. 5 (a). KPCA Projection 3D (PC1,PC3)

V. CONCLUSION

Principal component analysis extended from linear to non-linear form as Kernel Principal Component Analysis shows its significance over data visualization of drug-likeness dataset. It shows the non-linear structure pattern of data present in the dataset. KPCA preprocess the data by mapping the data to gram matrix and then perform linear functions of PCA. Along with visualization of dataset it helps in identifying the outliers present in dataset and results in dimensionality reduction.

VI. REFERENCES

1. Hann, M., Green, R. Chemoinformaticss A new name for an old problem. Curr. Opin. Chem. Biol.; 1999, 3, 379-383.
2. Thomas Engel. Basic Overview of Chemoinformatics, *J.* Chem. Inf. Model, 2006, 46, 2267-2277.
3. Nathan Brown. "Chemoinformatics- An Introduction to computer Scientists", ACM Computing survey, vol 41, Np. 2, Article 8, Publication date: Feburary 2009.
4. B.FirdausBegam and J.Satheesh Kumar, 2012. A Study on Cheminformatics and its Applications on Modern Drug Discovery, Procedia Engineering 38: 1264 – 1275.
5. Matthias Brustle, Bernd Beck, Torsten Schindler, William King, Timothy Mitchell, and Timothy Clark, Descriptors, Physical Properties, and Drug-Likeness, J. Med. Chem. 2002, 45, 3345-3355.
6. F.S. Tsai, Comparative Study of Dimensionality Reduction Techniques for Data Visualization. Journal of Artificial Intelligence, 3: 119-134, 2010.
7. R. E. Bellman, Adaptive control processes: A guided tour, ed., Princeton University Press, 1961.
8. K. Pearson, 1901. On lines and planes of closest fit to systems of points in space, Philosophical Magazine, (6) 2: 559-572.
9. H. Hotelling, 1933. Analysis of a complex of statistical variables into principal moments, Journal of Education Psychology, 24: 417-441
10. Tenenbaum, J.B., V. de Silva and J.C. Langford,. A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319-2323, 2000.
11. Zhang, Z. and H. Zha, Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J. Sci. Comput., 26: 313-338, 2005.
12. Scholkopf, B., A.J. Smola and K.R. Muller, Nonlinear component analysis as a kemel eigenvalue problem. Neural Comput., 10: 1299-1319, 1998.
13. Klaus-Robert Muller, Gunnar Ratsch, Soren Sonnenburg, Sebastian Mika, Michael Grimm, and Nikolaus Heinrich, Classifying 'Drug-likeness' with Kernel-Based Learning Methods, J. Chem. Inf. Model. 2005, 45, 249-253.
14. Kilian Q. Weinberger, FeiSha, Lawrence K. Saul, Learning a Kernel Matrix for Nonlinear Dimensionality Reduction, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004
15. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290, 2323–2326.
16. B. FirdausBegam and J. SatheeshKumar, Computer Assisted QSAR/QSPR Approaches – A Review, Indian Journal of Science and Technology, Vol 9(8), DOI: 10.17485/ijst/2016/v9i8/87901, February 2016
17. Christopher A. Lipinski, 2004. Lead- and drug-like compounds: the rule-of-five revolution, Drug Discovery Today: Technologies| Lead Profiling, (1) 4.
18. Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy and Paul J. Feeney, 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Advanced Drug Delivery Reviews 46: 3–26.

VII. AUTHORS PROFILE

Dr. B. Firdaus Begam is with Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, India. She has received her Master Degree in Software Science from Periyar University, Salem during 1998-2003. She has successfully completed Master of Philology in Computer Science during 2005 from Bharathiar University. She has completed her Ph.D in Computer Science in the Department of Computer Applications, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India during 2016. Her area of interestinclude Image processing and its applications on cheminformatics domain and Data mining and Machine Learning. (E-mail:firdh_2002@yahoo.com).

Dr. J. Rajeswari, working as Assistant Professor, in Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore. She has successfully completed her Ph.D from Karpagam Academy of Higher Education. Her area of interest is Data mining.
(E-mail: rajeswarikrishna82@gmail.com)