Clustering on High Dimensional Data that Reduces Dimensionality Using Dimension Reduction Techniques

K.S.GOWRILAKSSHMI
Lecturer Department of computer science
gowrisasi2@gmail.com

Sri Ramakrishna college of Arts and Science for Women, Coimbatore, India

***Abstract* – Dimensionality reduction is the search of small set of features to describe a large set of observed dimensions. The purpose of dimensionality reduction is to transform a high dimensional data set in to low dimensional space using clustering techniques of k-means. Recently new non linear methods introduced for reducing the dimensionality data called Locally Linear Embedding (LLE).LLE combined with K-means clustering in to coherent frame work to adaptively select the most discriminant subspace. K-means clustering use to generate class labels and use LLE to do subspace selection.**

***Keywords* -Clustering, High Dimension Data, Locally Linear Embedding, k-means clustering, Principal Component Analysis**

I. INTRODUCTION

High dimensional datasets present many mathematical challenges as well as some opportunities to bound to give rise to new theoretical development [23]. The problem is especially severe when large databases with many features are searched for patterns without filtering of important features based on prior knowledge. The growing importance of knowledge discovery and data mining methods in practical applications has made the feature selection/extraction problem

II. BACKGROUND STUDY

Developing effective clustering methods for high dimensional datasets is a challenging problem due to the *curse of dimensionality.* There are many approaches to address the problem of curse of dimensionality. The simplest approach of dimension reduction techniques [6], principal component analysis (PCA) (Duda et al., 2000; Jolliffe, 2002) and random projections (Dasgupta, 2000). In PCA method, dimension reduction is carried out as a pre-processing step and is decoupled from the clustering process. Once the subspace dimensions are selected, they stay fixed during the clustering process. This is main drawback of PCA [11] [27]. An extension of this approach is Locally Linear Embedding Techniques.

III. PROPOSED WORK OF LLE

The proposed work is carried out by collecting a large volume of high dimensional unsupervised gene datasets, by using dimension reduction techniques such as, LLE describe locally linear embedding (LLE) an unsupervised learning algorithm that computes low dimensional and neighbourhood preserving embeddings of high dimensional data. Generating highly nonlinear embeddings—do not involve local minima. The work is implemented in Mat Lab 7.0. The reduction can be done with the combination of LLE + k-means clustering.

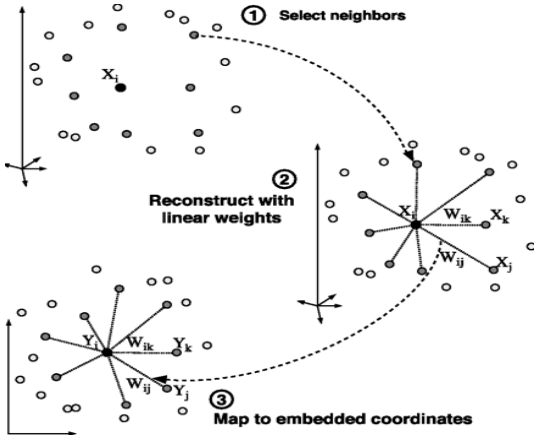
A. LLE

LLE is commonly called as Locally Linear Embedding; it is one of the dimension reduction techniques [23]. The *LLE* algorithm of Rowe's and Saul (2000) it involves mapping high-dimensional inputs into a low dimensional "description".

B. Steps of LLE Working

1) Assign neighbours to each data point W *X*i (for example by using the *K* nearest neighbours).
2) Compute the weights *W*ij that best linearly reconstruct Xi from its neighbours, solving the constrained least-squares problem. (3) Compute the low-dimensional embedding vectors Yi best reconstructed by *W*ij, minimizing the smallest eigenmodes of the sparse symmetric matrix in Although the weights *W*ij and vectors *Y*i are computed by methods in linear algebra, the constraint that points

3.3LLE Algorithms



Reconstruction errors are measured by the cost function.

$$\varepsilon(W) = \sum_i \left| \vec{X}_i - \Sigma_j W_{ij} \vec{X}_j \right|^2 \qquad (1)$$

The weights *W*ij summarize the contribution of the *j*th data point to the *i*th reconstruction. To compute the weights *W*ij, we minimize the cost space with as many function subject to two constraints: first, that each data point Xi is reconstructed only from its neighbours. The optimal weights *W*ij subject to these constraints are found by solving a least-squares problem.

$$\Phi(Y) = \sum_i \left| \vec{Y}_i - \Sigma_j W_{ij} \vec{Y}_j \right|^2 \qquad (2)$$

This cost function, like the previous one, is based on locally linear reconstruction errors, but here we fix the weights *W*ij while optimizing the coordinates Yi. The embedding cost in Eq. 2 defines a quadratic form in the vectors W *Y*i. Subject to constraints that make the problem well-posed, it can be minimized by solving a sparse *N* * *N* eigenvalue problem[36]. In this work experiments, data points were reconstructed from their *K* nearest neighbors, as measured by Euclidean distance.

IV. LLE+K-MEANS

Roweis and Saul (2000) have recently proposed [36]an unsupervised learning algorithm of low dimensional manifolds, whose goal is to recover the non linear structure of high dimensional data. The idea behind their local linear embedding algorithm is that nearby points in the high dimensional space remain nearby and similarly co-located in the low dimensional embedding. Starting from this intuition each data point x*i* (*i* = 1*; n*) is approximated by a weighted linear combination of its neighbours (from the nature of these *local* and *linear* reconstructions the algorithm derives its name). In its base formulation, the LLE algorithm finds the linear coefficients w_{ij} by minimizing the reconstruction

$$\varepsilon(\mathbf{w}) = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{n} w_{ij}\mathbf{x}_j||^2,$$

The same weights that characterize the local geometry in the original data space are supposed to reconstruct the data points in the low dimensional embedding. The *n* coordinates are then estimated by minimizing the cost function: for reconstructing it by its *k* neighbours, as in the first step of LLE. LLE is an unsupervised, non-iterative method, which avoids the local minima problems plaguing many competing methods (*e.g.* those based on the EM algorithm)

$$\Phi(\mathbf{y}) = \sum_{i=1}^{n} ||\mathbf{y}_i - \sum_{j=1}^{n} w_{ij}\mathbf{y}_j||^2.$$

This formula is applied in this proposal on five available data sets: namely Iris, Wine, Zoo, Cancer, and Drosophila. Consider a set of input data vectors *X* =(**x**1, ⋯ ,**x***n*) in high dimensional space. For simplicity, the data is centered in the preprocessing step, so that $\bar{\mathbf{x}} = \sum_i \mathbf{x}_i/n = 0.$

the standard *K*-means clustering is to minimize the clustering objective function [11].

$$\min_H J_K,\ J_K = \sum_k \sum_{i \in C_k} ||\mathbf{x}_i - \mathbf{m}_k||^2$$

Where the matrix $H = \{0,1\}^{n \times K}$ is the cluster indicator: $H_{ik} = 0$ if **x***i* belongs to the *k*-th cluster, $H_{ik} = 1$.

Suppose the data consist of real-valued vectors $\vec{X}_i$, each of dimensionality D, sampled from some smooth underlying manifold. It characterizes the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbors. In the simplest formulation of LLE, one identifies nearest neighbors per data point, as measured by Euclidean distance.

$$\mathcal{E}(W) = \sum_i \left| \vec{X}_i - \sum_j W_{ij} \vec{X}_j \right|^2,$$

1. Compute the neighbours of each data point Xi
2. Compute the Weights W_{ij} that best reconstruct each data point Xi from its neighbours minimizing the cost by constraints linear fits.
3. Compare the vectors Y_i best reconstructed by the weights W_{ij} .Minimizing the quadratic form by its bottom nonzero Eigen vectors.

V. EXPERIMENT AND RESULT

In this section it describes about the experiments evaluate the Performance of LLE and LLE +k -means in

Dimension reduction and compare with LDA+K-means, used widely in gene data sets.

VI. Dataset Description

The proposed work has been implemented and tested on five public available different gene data sets namely Iris, Breast Cancer, Drosophila, Wine, and Zoo.
The descriptions of these datasets are as follows.
Five datasets including Iris, Wine, Drosophila, cancer and Zoo

- Zoo gene describes the animal's hair, feathers, eggs, milk, airborne, back bone, fins, tails etc.
- Iris gene describes the flowers sepal length, sepal width, petal length, petal width it's a multivariate. The attribute used is real.
- Drosophila describes concise genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes.
- Breast Cancer describes tumour size, Class: no-recurrence-events, recurrence-events, age, menopause, and inv-nodes6. Node-caps, Breast-quad: left-up, left-low, right-up, right-low, central.

Wine data describes the attribute about Alcohol, Malic acid, Ash Alkalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoidphenols, Proanthocyanins.

TABLE I
DATA SET DESCRIPTION

Datasets	#Samples	#Dimensions	#Class
Iris	150	4	3
Wine	178	13	3
Zoo	101	18	7
Breast Cancer	286	9	4
Drosophila	163	4	3

TABLE 2
CLUSTERING ACCURACY TABLE ON UCI DATASET

Data set	K Means	PCA + KM	LLE + KM
Iris	0.886667	0.9643	0.98
Wine	0.646667	0.7909	0.85
Breast Cancer	0.893333	0.9645	0.99
Zoo	0.653333	0.7909	0.87
Drosophila	0.663	0.845	0.84

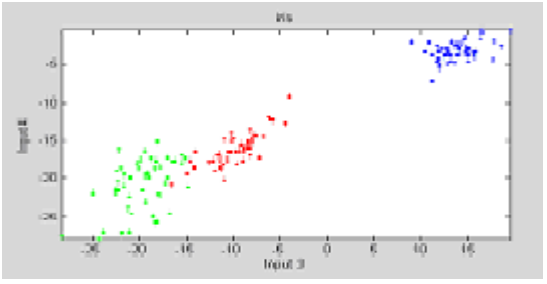


Fig2: k-means clustering with Iris

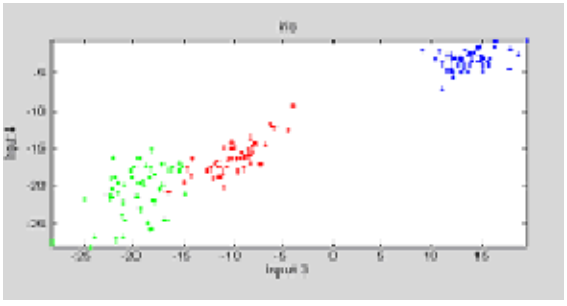


Fig 3: LDA+ k-Means with Iris Data set

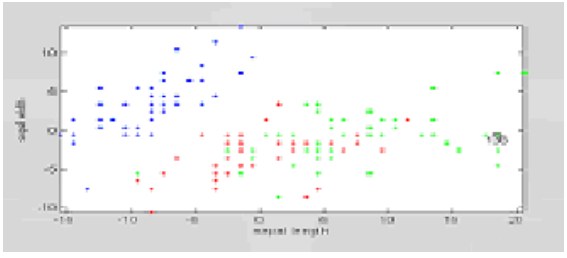


Fig 4: LLE+ k-means clustering

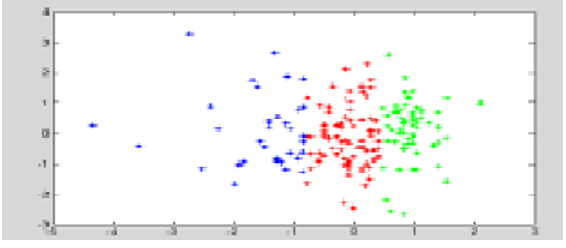


Fig 5: K-Means clustering with wine Data set With Iris Dataset

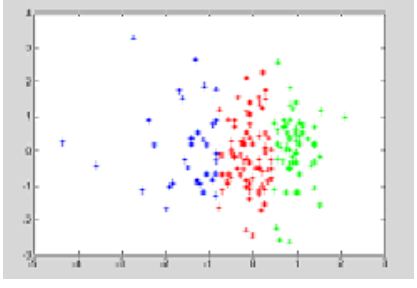


Fig 6: LDA-k means clustering

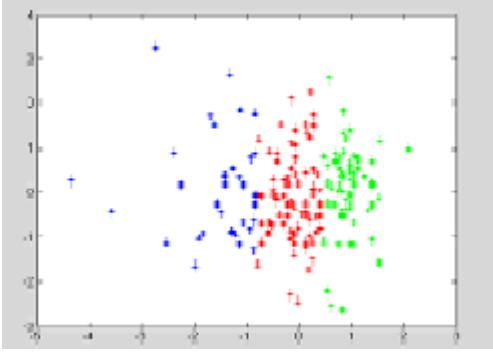


Fig7: LLE+K means clustering with wine Dataset

VII. RESULT ANALYSIS

All the above datasets have labels. View all the labels of the datasets as the objective knowledge on the structure of the datasets and use accuracy as the clustering performance measure. Accuracy discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contains data points from the corresponding class and it has been used as performance measures for clustering analysis. Accuracy can be described

$$Accuracy = Max(\sum_{C_k, L_m} T(C_k, L_m))/n,$$

Where *n* is the number of data points, *Ck* denotes the *k*-the cluster, and *Lm* is the *m*-the class. $T(C_k, L_m)$ Is the number of data points that belong to class *m* are assigned to cluster *k*. Accuracy is then computed as the maximum sum of $T(C_k, L_m)$ for all pairs of clusters and classes, and these pairs have no overlaps. On the five datasets data repository, it compares the LLE-Km algorithm with standard K-means algorithm. This work is also comparing it with PCA-based clustering algorithm: clustering. This shows that LLE-Km clustering is viable and competitive. The subspace clustering is able to perform the subspace selection and data reduction. Note that LLE-Km is also able to discover clusters in the low dimensional subspace to overcome the curse of dimensionality.

VIII. CONCLUSION

The performance of dimension reduction techniques with k-means clustering method using real different gene data sets was compared. In this thesis, considered various properties of data sets and data pre-processing procedures, and have confirmed the importance of pre-processing data prior to performing core cluster analysis. All clustering methods were affected by data dimension and data characteristics, such as the overlapping between clusters and the presence of noise. In particular the LLE + k m which are commonly used to reduce the dimension of data, The future extension of this work can be done via some other pre-processing techniques.

REFERENCES

[1] Alone et alum Broad Pattern of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleatide arrays. PNAS, 96(12): 6745-6750, 1999.

[2] Anton Schwaighofer. Matlab interface to svmlight.In (2004)

[3] Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., & Modha, D. (2004). A generalized maximum entropy approach to bregman co-clustering and matrix approximation. Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD).

[4] Barbara.D," An Introduction to Cluster Analysis for Data mining",

[5] Broom Head .D.S, Kirby, A new approach to dimensionality reduction:Theory and algorithms, SIAM journal of applied Mathematics 60 (6) (2000)2114-2142.

[6] Carreira-Perpinan.M.A, Areview of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield, 1997.

[7] Cheng, Y., & Church, G. (2000). Biclustering of expression data. Proc. Int'l Symp. Mol. Bio (ISMB), 93–103.

[8] Chrisding et al "K-means clustering Via PCA" 21st International Conference on Machine Learning, Canada-2004.

[9] Diamantras. K.I and Kung.S-Y.Principal Component Neural Networks. Theory and Applications. John Wiley & Sons, New York, Londin, Sydney, 1996.

[10] Ding, C., & He, X. (2004). K-means clustering and principal component analysis. Int'l Conf. Machine Learning (ICML)

[11] Ding, C., He, X., Zha, H., & Simon, H. (2002). Adaptive dimension reduction for clustering high dimensional data. Proc. IEEE Int'l Conf. Data Mining

[12] De la Torre, F., & Kanade, T. (2006). Discriminative cluster analysis. Proc. Int'l Conf. Machine Learning.

[13] D. DeMers and G.W. Cottrell. Non-linear dimensionality reduction. In C.L. Giles, S.J. Hanson, and J.D. Cowan, editors, Advances in Neural Information Processing Systems 5, pages 580–587. Morgan Kaufmann, San Mateo, CA, 1993.

[14] Fukunaga .K, Olsen .D.R, An Algorithm for Finding intrinsic dimensionality of data, IEEE Transactions on Computers 20 (2) (1976) 165-171.
[15] Halkidi.M, Batistakis.Y."On Clustering Validation Techniques", 2001.
[16] Han.J and Kamber.M. "Data Mining Concepts and Techniques", the Morgan Kaufmann Publisher, August 2000. ISBN 1-55860-489-8.
[17] Hartigan .J.A, Wang.M.A, "A K-Means Clustering Algorithm", Appl.Stat,Vol.28, 1979, pp.100-108.
[18] Hartuv.E, Schmitt. A, Lange.J, "An Algorithm for Clustering cDNAS for Gene Expression Analysis", Third International Conference on Computational Molecular Biology (RECOMB) 1999.
[19] Hotelling.H. "Analysis of a complex of statistical variables into principal components". Journal of Educational Psychology, 24:417–441, 1933
[20] Jackson .J.E. "A User's Guide to Principal Components". New York: John Wiley and Sons, 1991.
[21] Jain .A.K, Dubes .R.C Algorithm for clustering Data, Prentice Hall, 1988.
[22] Jain .A.K, Murty.M.N, Flynn.P.J. "Data Clustering: A Review" .ACM Computer Surveys, Vol.31, No.3, 1999.
[23] John stone I.M(2009) "Non-Linear dimensionality reduction by LLE".
[24] Jolliffe, I. (2002). "Principal component analysis". Springer. 2nd Edition.
[25] Jolliffe. I.T. "Principal Component Analysis". Springer-Verlag, 1986.
[26] Jolliffe I.T. "Principal Component Analysis" (Springer-Verlag, New York, 1989).
[27] Kambhatla.N and Leen. T. K. "Dimension reduction by local principal component analysis". Neural Computation **9**, 1493–1516 (1997).
[28] Kaski. S. "Dimensionality reduction by random mapping: fast similarity computation for clustering". Proc. IEEE International Joint Conference on Neural Networks, 1:413-418, 1998.
[29] Kaufman.L, Rousseeuw.P.J, "Finding Groups in Data: An Introduction to Cluster Analysis". John Wiley, New York, 1990.
[30] Kirby.M, Geometric Data Analysis: "An Empirical Approach to Dimensionality Reduction n and the Study of Patterns", John Wiley and Sons, 2001. Kohavi.R and John. G.The wrapper approach. In H. Liu and H. Motoda, editors, Feature Extraction, Construction and Selection: "Data Mining Perspective". Springer Verlag, 1998.
[31] O. Kouropteva, O. Okun, and M. Pietikäinen. Selection of the optimal parameter value for the locally linear embedding algorithm. 2002. Submitted to 1st International Conference on Fuzzy Systems and Knowledge Discovery.
[32] M. Kramer. Nonlinear principal component analysis using auto associative neural networks. AIChE Journal **37**, 233–243 (1991).
[33] Lee.Y and Lee.C.K, "Classification on multiple cancer types by multicategory support vector machines using gene expression data," Bioinformatics, vol.19, pp.1132-1139, 2003
[34] Li. K.-C. High dimensional data analysis via the SIR/PHD approach., April 2000. Lecture notes in progress.
[35] Li, T., Ma, S., & Ogihara, M. (2004). Document clustering via Adaptive subspace iteration. Proc. conf. Research and development in IR (SIRGIR) (pp. 218–225).
[36] Wolfe P.J and Belabbas .A (2008)"Hessian eigen maps: Locally Linear techniques for high dimensional data.
[37] Lunan .Y, Li.H, "Clustering of time-course gene expression data using a mixed effects model with B-splines, "Bioinformatics Vol.19, 2003, pp.474-482.
[38] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323-2326 (2000).
[39] Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for K-means clustering. Advances in Neural Information Processing Systems 14 (NIPS'01), 1057–1064.