International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 6, June - 2014

Improved Dimension Reduction With Modified Diffusion Maps

Yogesh Kailas Ankurkar
Electrical Engineering Dept
VJTI, Mumbai
India

Ruta Ashok Kambli
Electrical Engineering Dept
VJTI, Mumbai
India

Ameya Vinay Mane
Electrical Engineering Dept
VJTI, Mumbai
India

***Abstract*— The task of representing the higher dimensional data into lower dimension while preserving the relative information, previously was done by principle component analysis, factor analysis, or feature selection. However if original lower dimensional data is embedded in high dimensional space, then approach based on manifold learning and graph theory allow to learn underlying geometry of data. One of such technique is Diffusion Maps. It preserves the local proximity between the data points by first constructing a representation for underlying manifold. In this paper, binary classification problem using Diffusion Map to embed the data with various kernel representations is targeted. Results show that specific kernels are well suited for Diffusion Map applications on some feature sets and in general some kernels are outperform the standard Gaussian and Polynomial kernels, on several of the higher dimensional data sets.**

Keywords—diffusion maps, dimension rectuction, diffusion kernels

I. INTRODUCTION

The trade-off between computational complexity and the resolution gained with either more features or pixels is basic and most important problem high dimensional data analysis. Hence, very first step in analyzing data is to find its lower dimensional representation and the concise description of its underlying geometry and density. To achieve this, generally the global dimension reduction techniques such as principle component analysis, multidimensional scaling is used. These techniques work well with well-behaved maximally variant data. But if the data is not locally correlated, then these techniques do not provide informative embedded data. Alternatively, graph based manifold learning techniques generally preserves the neighborhood structure. They generally preserve the neighborhood structure. Such techniques are Diffusion Maps [1] and [2], Local linear Embedding [3], Laplacian Eigenmaps [4], Hessian Eigenmaps [5], and Local Tangent Space Alignment [6].

In this paper we consider the manifold learning technique Diffusion Maps of Coifman et al. [1], [2] and analyze the neighborhood preserving effects of kernel selection on the resulting manifold for publicly available data sets. These effects are studied by looking at the classification results for each binary target data set in various embeddings.

II. DIFFUSION MAPS

A. Overview

Eigenvectors of random walk on the given dataset, giving lower dimension Euclidean space embedding of complex data, defines the Diffusion Maps. This embedding can be used for the manifold learning, dimensionality reduction, geometric analysis of complex datasets and fast simulation of stochastic dynamical systems.

By constructing a graph representation for underlying manifold Diffusion map preserves the local proximity between data points. The vertices, or nodes of this graph, represent the data points, and the edges connecting the vertices, represent the similarities between adjacent nodes. The probability for the random walk on the graph can be interpreted from the normalized edge weights. Diffusion maps transform data from higher dimensional space to lower dimensional space such as Euclidean space and hence the Euclidean distance between the points approximates the diffusion distance in the original feature space. The geometric structure of underlying data determines the dimension of diffusion space and accuracy of approximation of the diffusion distance into Euclidean distance.

B. Connectivity between data point

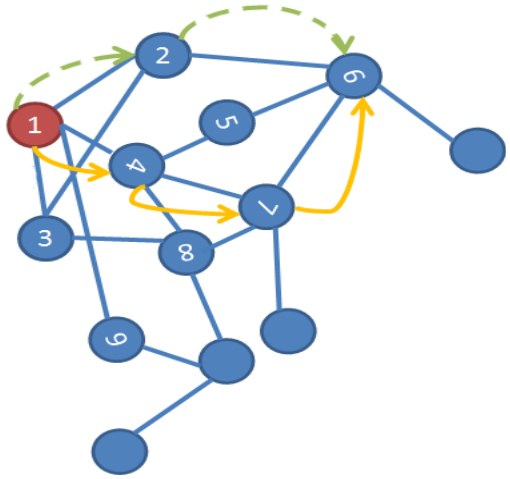
The random walk on the dataset is represented in the figure 1. The process tends to jump to the nearest point rather than farthest point. This means probability of random walk going to nearby point is more than probability of jumping to far away point. This fact helps to establish relation between feature space and probability. The connectivity between points x and y is defined as probability of jumping between them and it is given as

$$\text{Connectivity}(x, y) = p(x, y) \quad (1)$$

The connectivity is defined in terms of normalized likelihood kernel function, k and the relation is defined as bellow.

$$\text{Connectivity}(x, y) \propto k(x, y) \quad (2)$$

This kernel function, within certain neighborhood defines local measure of similarity. Kernel function goes to zero outside the neighborhood. The neighborhood is nothing but the area within which similarity measurement can be assumed to be accurate. Its size depends on the kernel parameters. For example, consider the popular Gaussian kernel,



$$k(x,y) = \exp\left(-\frac{|x-y|^2}{\alpha}\right) \qquad (3)$$

All the elements of y for which $k(x,y) \geq \varepsilon$ with $0 < \varepsilon \ll 1$, define the neighborhood of x. Here in the above kernel equation by choosing the values of α the size of neighborhood can be defined. Generally based on the data type, the size of neighborhood is chosen, for example, for sparse data, a large neighborhood is chosen and for intricate, nonlinear, lower dimensional structures, a small neighborhood is appropriate. The diffusion kernel satisfies the following properties:

1. k is symmetric: $k(x, y) = k(y, x)$
2. k is positivity preserving: $k(x,y) \geq 0$

The row –normalized matrix P is diffusion matrix with entries $P_{ij} = p(Xi, Xj)$. Each element of P matrix shows the connectivity between the two data points, Xi and Xj. This matrix provide the probabilities for a single step taken from i to j. Consider a 2×2 diffusion matrix,

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Each element, p_{ij} is the probability of jumping between data points i and j. Similarly, P^t_{ij} gives sum all paths of length t from point i to point j.

C. Diffusion Process

The data set are observed at different scales, as the value of t, in P^t, increases. The diffusion is a process, where the global connectivity of a data-set is provided by integrating local connectivity.

The data points are highly dense and connected along the geometrical structure. Hence the probability of following underlying geometrical structure increases as value of t increases. The path, diffusion process follows is made up of small-high probability jumps, rather than the long, low-probability jumps.

D. Diffusion Distance

The diffusion distance depends on the number of short, high-probability paths. The diffusion distance is small if there are t high probability paths between two points and vice-versa. The diffusion matric D, provide measure of connectivity between two points, as similarity between them in the observation space. The diffusion matric is robust to noise perturbation and it sums up all the possible paths of length t. the relation between diffusion matric and diffusion matrix is given by,

$$D_t(X_i + X_j)^2 = \sum_{u \in X} |p_t(X_j, u)|^2 = \sum_k |P^t_{ik} - P^t_{kj}|^2$$

The term $p_t(x,u)$ gives probability of jumping from x to u (for any u in the data set) in t time units, and sums the probabilities of all possible paths of length t between x and u. As explained in the previous section, this term has large values for paths along the underlying geometric structure of the data. In order for the diffusion distance to remain small, the path probabilities between x, u and u, y must be roughly equal.

E. Diffusion Map

In the previous section, a metric, the diffusion distance is defined which capable of approximating distances along this structure. The diffusion distance calculation is expensive process; hence data points are mapped into Euclidean space according to diffusion metric. In the Euclidean space, diffusion distance becomes Euclidean distance. A diffusion map, which maps coordinates between data and diffusion space, aims to re-organize data according to the diffusion metric. The diffusion map preserves a data set's intrinsic geometry, and since the mapping measures distances on a lower-dimensional structure. To find that fewer coordinates needed to represent data points in the new space examine the mapping.

$$Y_i := \begin{bmatrix} p_t(X_i, X_1) \\ \vdots \\ p_t(X_i, X_N) \end{bmatrix} = P^T_{i*} \qquad (8)$$

For this map, the Euclidean distance between two mapped points, Yi and Yj, is

$$\|Y_i - Y_j\|^2_E = \sum_{u \in X} |p_t(X_i, u) - p_t(X_j, u)|^2$$

$$= \sum_k |P^t_{ik} - P^t_{kj}|^2 = D_t(X_i, Y_j)^2$$

Which is the diffusion distance between data points *Xi* and *Xj* . This provides the re-organization according to diffusion distance. Note that no dimensionality reduction has been achieved yet, and the dimension of the mapped data is still the sample size, N. Dimensionality reduction is done by neglecting certain dimensions in the diffusion space. Take the normalized diffusion matrix,

$$P = D^{-1}K$$

where D is the diagonal matrix consisting of the row sums of K. The diffusion distances in (8) can be expressed in terms of the eigenvectors and -values of P as

$$Y_i' = \begin{bmatrix} \lambda_1^t \psi_1(i) \\ \lambda_2^t \psi_2(i) \\ \vdots \\ \lambda_n^t \psi_n(i) \end{bmatrix} \qquad (9)$$

where $\psi_1(i)$ indicates the i-th element of the first eigen vector of P. Again, the Euclidean distance between mapped points $Y_i'$and $Y_j'$is the diffusion distance. The set of orthogonal left eigenvectors of P form a basis for the diffusion space, and the associated eigenvalues λ_l indicate the importance of each dimension. Dimensionality reduction is achieved by retaining the m dimensions associated with the dominant eigenvectors, which ensures that $\|Y_i' - Y_{j'}\|$ approximates the diffusion distance, $D_t(X_i, X_j)$, best. Therefore, the diffusion map that optimally preserves the intrinsic geometry of the data is (9).

III. KERNEL FUNCTION

The kernel function captures a specific feature of data set, and also constitutes of local geometry. Hence its choice should be based on the type of application. The list of kernel functions used in this paper is given bellow:

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

- Laplacian Kernel:

$$k(x, y) = \exp(-|\|x - y\| - u|/b)/2b$$

- Gaussian Kernel:

$$k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$$

- Rayleigh Kernel:

$$k(x, y) = \frac{\|x - y\| \exp(-\|x - y\|^2/2\sigma^2)}{\sigma^2}$$

- Polynomial Kernel:

$$k(x, y) = (1 + \langle x, y \rangle)^d$$

IV. EXPERIMENTS

A. Experimental Setup

In the experiment, the effect of different kernel functions, on the total diffusion process and hence on dimension reduction is observed and studied. Each database is divided into ten groups that are as equal as possible, 10-fold cross validation. Out of ten groups nine groups are used for training and one group for testing purpose. This procedure is repeated until all groups have represented as testing set.

The average performance overall 10-folds is presented as the probability of classification (PC), or sensitivity, and the probability of false alarm (PFA), or specificity. This is done to demonstrate the trade-off between correctly classifying true cases versus incorrectly classifying false cases. Each kernel uses the same groups for each data set so that the possibility of poor individual performance due to the distribution of the draw is eliminated. In addition, each experiment is done ten times and the results are averaged over these runs.

B. Data set

The experiment discussed above tests the kernels and their embeddings for classification enhancement on the resulting Diffusion Maps over eight publically available data sets [8]:

- Pima Indian: Pima Indian Diabetes
- Sonar1: Connectionist Bench Sonar
- WDBC: Wisconsin Diagnostic Breast cancer
- WPBC: Wisconsin Prognostic Breast Cancer
- Heart: Heart Disease Data Set, Cleveland

The classification is done using linear support vector machine [10]. Let M m-dimensional training inputs x_i (i = 1, . . . , M) belong to Class 1 or 2 and the associated labels be y_i = 1 for Class 1 and −1 for Class 2. If these data are linearly separable, the decision function is given by:

$$y_i\left(w^T x_i + b\right) \geq 1 \; for \; i = 1, \ldots\ldots, M \qquad (10)$$

Where W is an m-dimensional vector, b is a bias term[11].

V. RESULTS AND CONCLUSION

The experimental results for the kernel effects on the resultant diffusion maps are shown below in Table 2 through Table 9. As the experiments demonstrate, the choice of kernel effects the resultant diffusion map. Overall, the Laplacian and Rayleigh kernels outperformed the standard Polynomial and Gaussian kernels on all of these databases, with a few exceptions such as the Pima Indian. BC datasets. It appears that the Laplacian and Rayleigh kernels perform best on the higher dimensional non-Gaussian datasets and the standard kernels work well with lower-dimensional data. Therefore, for enhanced target recognition capability and an acceptable PFA the Rayleigh kernel appears the appropriate choice to best capture the embedding distribution to enhance the diffusion map process.

REFERENCES

[1] R. Coifman; S. Lafon, "Diffusion Maps," Applied and Computational Harmonic Analysis, special issue on diffusion maps and wavelets, vol. 21, pp. 5-30, July 2006.

[2] R. Coifman; S. Lafon; A. Lee; M. Maggioni; B. Nadler; F. Warner; S. Zucker, "Geometric Diffusions as a Tool for Harmonics Analysis and Structure Definition of Data: Multiscale Methods," Proc. Nat"l Academy of Sciences, vol. 102, no. 21, pp. 7432-7437, May 2005.

[3] S. Roweis; L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, pp. 2323-2326, 2000.

[4] M. Belkin; P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," Neural Computation, v.15 n.6, p.1373-1396, June 2003.

[5] D. Donoho; C. Grimes, "Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-Dimensional Data," Proc. Nat'l Academy of Sciences, vol. 100, no. 10, pp. 5591-5596, May 2003.

[6] Z. Zhang; H. Zha, "Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment," Technical Report CSE-02-019, Dept. of Computer Science and Eng., Pennsylvania State Univ., 2002.

[7] Isaacs, J.C.; Foo, S.Y.; Meyer-Baese, A., "Novel Kernels and Kernel PCA for Pattern Recognition," Computational Intelligence in Robotics and Automation, 2007. CIRA 2007. International Symposium on , vol., no., pp.438-443, 20-23 June 2007

[8] A. Asuncion; D.J. Newman; (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[9] G. Sammelmann; J. Christoff; J. Lathrop; "Synthetic Images of Proud Targets", Proc. IEEE/MTS OCEANS 2004, pp. 266-271

[10] Vapnik, V. (1995). "Supportvector networks". *Machine Learning* **20** (3): 273.doi:10.1007/BF00994018

[11] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, B. P. (2007). "Section 16.5. Support Vector Machines"