35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Classification of Clinical Data using Diffusion Maps: a Density Based Approach

Ko-Kung Chen, Chih-I Hung, Bing-Wen Soong, Hsiu-Mei Wu, Yu-Te Wu, Po-Shan Wang*

Abstract— **Clinical data analysis is of fundamental importance, as classifications and detailed characterizations of diseases help physicians decide suitable management for patients, individually. In our study, we adopt diffusion maps to embed the data into corresponding lower dimensional representation, which integrate the information of potentially nonlinear progressions of the diseases. To deal with nonuniformaity of the data, we also consider an alternative distance measure based on the estimated local density. Performance of this modification is assessed using artificially generated data. Another clinical dataset that comprises metabolite concentrations measured with magnetic resonance spectroscopy was also classified. The algorithm shows satisfactory results.**

I. Introduction

The rapid development of the clinical investigation modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), provides more high dimensional data on certain given disease. Analysis of clinical data is of challenging due to several reasons:

- Relatively small size of available data: recruiting patients is sometimes difficult; this is especially true if additional criteria are required for patients to meet, or the study aims to recruit patients with rare diseases.
- High dimensionality of the data: a whole gamut of clinical investigation modalities can be used to obtain the information about patients noninvasively, even a single instrument can offer a list of various features.
- High complexity of the data: The progression of many diseases and symptoms do not follow a linear trend, and sometimes there can potentially be several different developing patterns for certain disease. These characteristics result in irregular distribution, and possibly nonlinear structure of the data.
- High noise: the distribution of the features of patients can sometimes cover a wide range. Also the involuntary movements of the patients cause the unpredictable variation of the data, especially when the instruments are sensitive to motion, such as MRI and magnetic resonance spectroscopy (MRS).
- The gray zone: Sometimes it is difficult to differentiate different groups since different diseases may mimic similar, or even the same symptoms.

Exploring the behavior and patterns of clinical data is mostly done with statistics or linear analysis. But the aforementioned natures may make these approaches incapable of dealing the data effectively and efficiently. Sometimes the higher dimensional data may actually lie on a lower dimensional space. Recently a new approach, namely diffusion maps [1], had been proposed to deal with high dimensional data. Based on the stochastic process on the spectral graph theory, diffusion maps is among the most powerful spectral dimensionality reduction tool to locate intrinsic lower dimensional coordinates of a given multi-dimensional dataset [2]. Diffusion Maps has applied to diverse applications [3-6].

The diffusion maps use a distance measure that preserve local information of a given dataset. The distance between a pair of data points is short providing there exists some paths connecting them; that is, the affinity for this pair of points is high. This characteristic is ideal in clinical data analysis since the distribution patterns of the patients do not always behave in a linear sense. And intuitively, the progression of diseases and symptoms mimics the concept of local connectivity if the data exhibits certain longitudinal behavior. In addition to capable tracking down the nonlinear structure, diffusion maps also reduces the dimensionality of the data. However, the irregular distribution of the data, along with relative small population size and other factors discussed above, complicate the discerning of the data.

In this study, we use a self-tuning kernel, which is coupled with a density estimator, to adjust the bias introduced by the underlying distribution of the data. The capability of this approach lies in its ability to deal with data scattering induced by nonuniform density. An artificially generated data using Gaussian distribution is given in later section to illustrate this

This research was supported in part by the National Science Council (NSC101-2221-E-010-004-MY2, NSC101-2314-B-733-001-MY2) and Brain Research Center, National Yang-Ming University and a grant from Ministry of Education, Aim for the Top University Plan.

Ko-Kung Chen is with the Dept. of Biomedical Imaging and Radiological Sciences, Brain Research Center, National Yang-Ming University, Taipei, Taiwan (e-mail: ancient_laws@ yahoo.com.tw).

Po-Shan Wang is with Institute of brain science, Taipei Veteran General Hospital, Taipei, Taiwan, Dept. of Medicine, Municipal Gandau Hospital, Taipei, Taiwan, Dept. of Neurology, Brain Research Center, National Yang-Ming University, Taipei, Taiwan, Dept. of Neurology, Taipei Veteran General Hospital, Taipei, Taiwan (corresponding author).

Chih-I Hung is with the Dept. of Biomedical Imaging and Radiological Sciences, Brain Research Center, National Yang-Ming University, Taipei, Taiwan.

Bing-Wen Soong is with the Dept. of Neurology, Brain Research Center, National Yang-Ming University, Taipei, Taiwan, Dept. of Neurology, Taipei Veteran General Hospital, Taipei, Taiwan.

Hsiu-Mei Wu is with Dept. of Radiology, Taipei Veteran General Hospital, Taipei, Taiwan, Dept. of Radiology, National Yang-Ming University, Taipei, Taiwan.

Yu-Te Wu is with the Dept. of Biomedical Imaging and Radiological Sciences, Brain Research Center, National Yang-Ming University, Taipei, Taiwan, (phone: +886228267169; fax: +886228201095;e-mail: ytwu@ym.edu.tw).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

phenomenon. Another clinical dataset comprise spinocerebellar ataxia type 3 (SCA3) patients, multiple system atrophy (MSA) patients, and several normal subjects is classified as well.

II. Materials and Methods

A. Diffusion Maps

For a given measure (X, μ), a dataset X consists of N samples with underlying distribution μ is included. The data points may be characterized as $x_i=(y_{i1},y_{i2},\ldots,y_{il},\ldots,y_{im})$, $i=1,2,\ldots,N$, $l = 1,2,\ldots,m$, A kernel $k:X \text{ x } X\rightarrow R$ is defined to measure the pairwise similarities between every pair of data points. The kernel function is nonnegative, and it defines certain notion of connectivity between data points pairwisely. Since the design of the kernel will influence the geometry captured by diffusion maps, the choice of the kernel should be guided by the characteristics of the data or prior knowledge that one bears in mind. A popular choice for distance measure is the Gaussian kernel:

$$k(x_{i,} x_j) \equiv \exp\left(\frac{-\| x_i - x_j \|^2}{\sigma^2}\right) \tag{1}$$

Since the kernel is symmetric, the constructed weighted graph will be undirected as well. Once the choice of kernel is determined, the mass can be defined as:

$$m(x_i) \equiv \int_X k(x_{i,} x_j)\, d\mu(x_j) > 0 \tag{2}$$

Then a weighting function can be built by normalizing the kernel using mass:

$$p(x_{i,} x_j) \equiv \frac{k(x_{i,} x_j)}{m(x_i)} \tag{3}$$

Since the weighting function satisfies $\int_X p(x_{i,} x_j)d\mu(x_j) = 1$, the constructed graph can be viewed as an asymmetric Markov chain built over the data, where the $p(x_{i,}\ x_j)$ is interpreted as the probability for state x_i transits to state x_j in a single time step. A square matrix P whose elements are $p(x_{i,} x_j)$ is then constructed. Taking powers of P, which is equivalent to drive the Markov chain forward, will reveal corresponding intrinsic geometry of the data. If one allow the Markov chain running unceasingly, all the data points will be merged together and regarded as a single cluster.

As long as the matrix P is nonsingular, it can be written in quadratic form: $P^t = v\ \lambda^t\ v^{-1}$, where v is the discrete set of eigenfunctions $\{v^{(i)}:i=1,2,\ldots,N\}$ with corresponding eigenvalues $\{(\lambda^{(i)})^t:i=1,2,\ldots,N\}$. The sequence of eigenvalues has the property such that $1= |\lambda^{(1)}| \geq |\lambda^{(2)}| \geq \ldots \geq |\lambda^{(N)}| \geq 0$. Since the sequence of eigenvalues tends to zero, a few largest eigenvalues and their corresponding eigenfunctions can be used to approximate the P with minimal truncation error.

The diffusion maps is then defined as:

$$\Psi^{(t)}(x_i) \equiv \{(\lambda^{(j)})^t v^{(\mathrm{j})}\,(x_i)\},j = 1,2,\ldots,N \tag{4}$$

The dimension of the new embedding depends on only the powers of the P and no longer depends on the dimension of the original data. Coifman and Lafon [1] define the diffusion distance between state x_i and state x_j in t time steps to be:

$$D^{(t)}(x_{i,} x_j) \equiv \|\Psi^{(t)}(x_i) - \Psi^{(t)}(x_j)\|^2. \tag{5}$$

The diffusion distance computes the affinity between data points pairwisely. The diffusion distance is robust to noise, as the distance between every pair of data points depends on all existing connections between them.

Since P is asymmetric, (5) is actually built under the weighted distribution. Alternatively, another diffusion distance can be defined by using a symmetric weighting function to simply certain settings. A symmetric normalization can be obtained by defining the kernel to be:

$$k\hat{}(x_{i,} x_j) \equiv \frac{\sqrt{m(x_i)}}{\sqrt{m(x_j)}}\, p(x_{i,} x_j) = \frac{k(x_{i,} x_j)}{\sqrt{m(x_i)}\sqrt{m(x_j)}} \tag{6}$$

Setting up the matrix P with (6) as distance measure, the resulting formula for diffusion distance will be base on the underlying distribution only.

B. Self-tuning Kernel based on Density Estimation

The design of the kernel influences the resulting embedding due to the fact that the structure of the constructed Markov chain is altered. Even if the kernel is drawn from one of the known parametric family of distributions, tweaking its parameters may yield quite distinct results; this is especially true if the underlying distribution function of the data is irregular.

While a global setting captures the intrinsic geometry of the data, it would not be able to effectively address the intrinsic nonuniform density of the data. To compensate the bias and skewness introduced by the distribution of the data, we consider the local density of the data points. Density plays an important role in statistics; it conveys the distribution pattern to be drawn from the data. In our case, consider any point x_i in the original data, let the set $\xi(x_i) \equiv \{\, x_j\text{: } \| x_i - x_j \| < \varepsilon$, $\mathrm{j}=1, \ldots,n_i\}$ being its neighbor. Providing one assume that the data is drawn from the normal distribution $N(u, \tau^{\,2})$ and the variance of all dimensions are the same, then the ratio that the local variance of data points in the set $\xi(x_i)$ to the global variance $\tau^{\,2}$ should depend on the size of the local set, that is, n_i. One may further assume that if this ratio of associated with $\xi(x_i)$ surpasses certain predefined value, then data points in the $\xi(x_i)$ are actually discernable. Since the density of $\xi(x_i)$ is proportional to its sample size, we can use n_i as a density estimator.

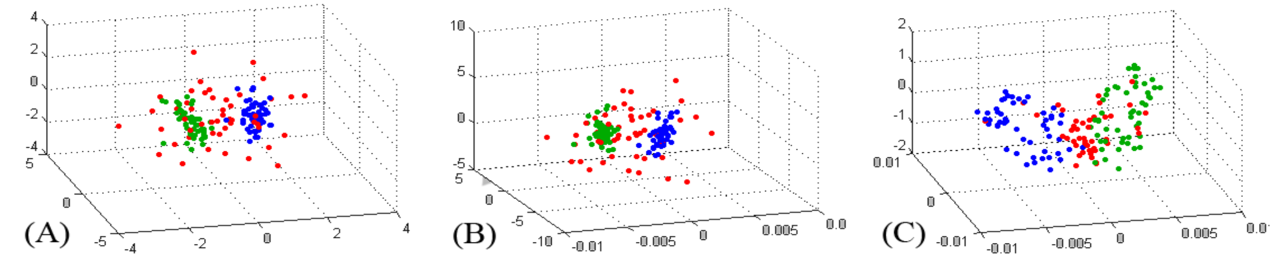
Alternatively, one can use the following integral to estimate the local density instead of n_i, as delineated by Silverman [6]:

$$d(x_i) \equiv \int_{\xi(xi)} k_d(x_{i,} x_j)\, dx \tag{7}$$

where k_d is a Gaussian function that takes the same form as (1), except for the fact that σ is replaced by σ_d (In our case, we set σ_d to be ε). For normal distribution, it can be shown that estimated $d(x_i)$ will converge to n_i as the sample size N is sufficiently large. The $d(x_i)$ is then incorporated into the (1) to form a self-tuning kernel:

$$a(x_{i,} x_j) \equiv \exp\left(\frac{-d(x_i)\| x_i - x_j \|^2}{\tau^{\,2}}\right). \tag{8}$$

Figure 1. Demonstration of two different approaches using artificially generated data. The setup for this data constitutes three Gaussian distribution: $N((1,0,0),1)$, $N((-1,0,0),1)$, and $N((0,0,0),2)$. (A) The original data displayed in 3D. The red group disperses through both the green one and the blue one. One can consider the higher density clusters as typical representative of certain disease, or the gray zone induced by overlapping interval between different groups. (B) The diffusion coordinate computed using Euclidean distance measure. The red group is incorrectly regarded as the background of the data due to its lower density. (C) Diffusion coordinate obtained with self-tuning kernel. The dispersive red cluster is glued altogether and can be easily identified.



This approach is more flexible, and can reduce the possibility that samples with different characteristics being identified as the same due to nonuniform density of the data. Furthermore, the method is nonparametric; this feature is desirable since no additional prior or background knowledge is required for clients to obtain meaningful results.

C. *Implementation*

1. For any m-dimensional data, we normalize the data such that Var[y_{il}^2] = 1, l = 1,2,…,m, respectively.

2. For any point x_i, its neighbor is defined to be the set $\xi(x_i)$; a Gaussian kernel k_d is then adopted to evaluate the local density of the data. A normalized density $d'(x_i)$ is then defined by dividing $d(x_i)$ by $\int_X d(x_i)dx_i$ such that . $\int_X d'(x_i)dx_i = 1$.

3. A density based Gaussian kernel is used to evaluate the pairwise affinity. The τ^2 is assumed to be Var[y_{il}^2]$^{1/2}$.

4. A symmetric matrix is built using aforementioned formula, with the parameter t being 2. By spectral decomposition, a lower dimensional diffusion embedding can be defined using several largest nontrivial eigenfunctions and their corresponding eigenvalues up to a predefined precision.

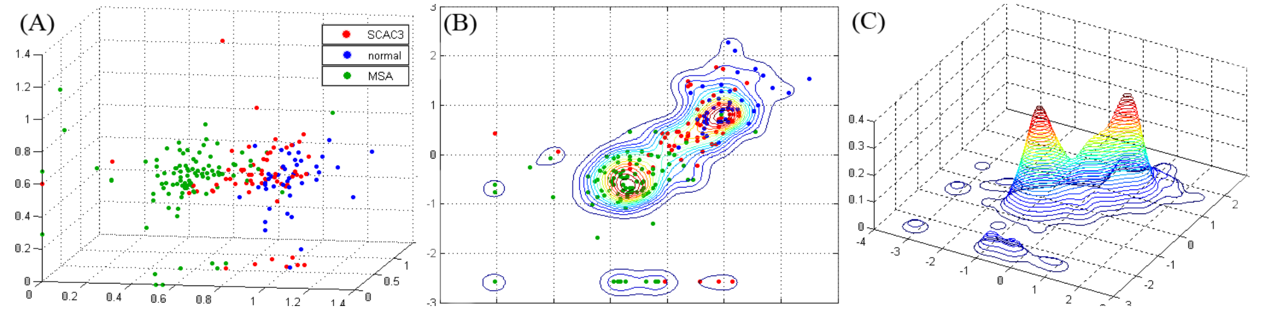
A result using artificially generated data is given in figure 1. We assume that the local density of the gray zone is generally higher than clean groups, since it is a mixture of subjects from different clusters; also, if there are multiple clusters with varied density appear at once, it is unlikely that a naïve kernel would be able to treat the data properly. Each dataset is randomly divided into training set and testing set. The training set is fed to train the support vector machine (SVM) first, and then evaluating the performance of SVM with the testing set. The results show that the overall classification ratio has been improved using the modified method.

An important characteristic of such spectral clustering techniques is that they are feasible only if the different groups can be separated in the lower dimensional representation [8]. This issue arises in our study of the clinical dataset, where the classification accuracy of the MSA groups in the diffusion embedding actually decreases in comparison to that using original data.

III. Experimental Results

The clinical dataset comprise relative metabolite concentrations measured using magnetic resonance spectroscopy (MRS). The MRS is carried out on left and right cerebellum, left and right basal ganglia, and vermis. The relative concentration of three different metabolites, namely N-acetylaspartate (NAA), Choline (Cho), and myo-inositol (mI), are measured at all five anatomies; these three concentrations have been normalized using the concentration of creatine (NAA/Cr, Cho/Cr, mI/Cr). The dataset consists of three different groups, namely the 63 SCA3 patients, 98 MSA patients, and 44 normal subjects. While the SCA and MSA share similar clinical symptoms, the MRS has been shown to be a potential modality to differentiate SCA and MSA,

Figure 2. The Intermediate result of density estimation performed on the dataset. (A) The original data in 3D view. The dimensions being chosen as coordinates are: NAA/Cr in both left and right cerebellum, and Cho/Cr in left cerebellum. (B) The contour of bivariate density estimation using NAA/Cr in both left and right cerebellum. (C) The corresponding 3D contour of the estimated density. It can be inferred that not only SCA3 has lower density, but the gray zone induced by both normal subjects and SCA3 group complicate the situation further.



particularly multiple system atrophy-cerebellar type (MSA-C) [9]. While the MSA patients are readily separable from the other two groups, classifying SCA3 patients and normal subjects is more difficult, as shown in figure 2A. Using diffusion maps, we obtain an embedding that separate the original data into three clusters, but all of them are mixture of different groups; furthermore, one may have the wrong impression that the MRS is not capable of discerning these patients. This suggests a naïve distance measure is unsuitable. Density estimation is first performed on the original data (figure 2B and 2C), then incorporated into the distance measure as self-tuning factor. Figure 3 shows embeddings obtained with simple kernel and density based kernel.

The classification is carried out on original data, principal component based representation (PCA), diffusion embedding, and diffusion embedding with density based kernel built in, respectively. The SVM is performed at least twenty times for each setting. The classification accuracy is listed in the table 1.

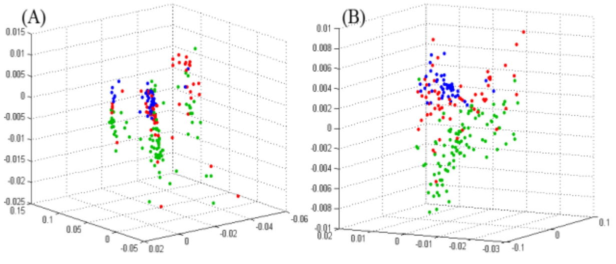
Table 1. Classification accuracy using SVM

	SVM Classification Accuracy			
	Original Space	PCA based	Naïve kernel	Density based kernel
SCA3	65% - 75%	69% – 73%	81% – 90%	87% - 95%
MSA	83% - 93%	48% - 51%	70% - 81%	72% - 79%
Normal	65% - 73%	75% - 78%	50% - 65%	78% - 85%

It is evident that the density based approach performed better than naïve kernel, especially in classifying normal subjects; and the performance of PCA based approach is inferior to that of density based kernel approach.

Also, it is interesting to note that in comparison with the classification using original data, the accuracies of MSA decreased for all three other approaches, particularly in the PCA based one. The issue can be attributed to nonuniformaity of the dataset itself. We illustrate this by considering the embedding of the simulated data in figure 1C: as one embed the red group in to a more compact form, the structures of both green and blue group are becoming dispersive, leading to complications in identifying them. Such tradeoff between data structure and scattering removal is inevitable as one applies the dimensionality reduction method given the nonuniformity induced by random samples. In other words, the information of the compact group will leak into the relative sparse group; the larger the difference of the density between these groups, the more leakage from those originally compact structure.

Figure 2. Diffusion coordinate obtained with different approaches. (A) Diffusion embedding computed with the naïve distance measure. The data roughly is divided into three distinct clusters, all of which are mixture of different groups. (B) Result obtained using self-tuning kernel. The false clusters in (A) no longer exist.



Based on such observation, we suggest that the density based kernel design is suitable for dealing data mixture where different groups are partially mixed with one another, as well as possesses nonlinearity induced by intrinsic nonuniformity. The drawback is that if the densities of distinct clusters differ significantly, the self-tuning kernel would develop a potentially false one way relation such that the affinities from sparse regions to compact groups are overly estimated, but not the contrary; this characteristic cause information leakage from originally compact and well defined structure, hence mollifying the rigidity of the original data.

IV. Conclusion

Based on the clustering properties of the diffusion maps, we analyze the clinical data in a lower dimensional space induced by distance measure of the diffusion maps. To adjust the nonuniformity introduced by the underlying distribution of the data, we estimate the local density of the data, and use it as self-tuning factor of the distance measure. This approach shows satisfactory results on both artificial data and metabolite concentrations obtained with MRS. The results also demonstrate that distance measure with scaling factor based on variance of local mean generally is more capable than a naïve kernel.

Acknowledgment

We would like to thank Bing-Wen Soong and Hsiu-Mei Wu from Taipei Veteran General Hospital for making the MRS data available to us.

References

[1] R. R. Coifman, S. Lafon, "Diffusion maps," A*ppl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp.5-30, 2006.

[2] A. Singer, R. R. Coifman, "Non-linear independent component analysis with diffusion maps," *Appl. Comput. Harmon*. Anal., vol. 25, no. 2, pp.226-239, 2008.

[3] S. Lafon, Y. Keller, R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784-97, November 2006.

[4] B. Nadler, S. Lafon, R. R. Coifman, I. G. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Plank operators," *Neural Information Process. System* 18. MIT Press, 2005, pp. 955-962.

[5] A. Singer, Y. Shkolnisky, B. Nadler, "Diffusion interpretation of nonlocal neighborhood filters for signal denosing," *SIAM Journal Imaging Science*, vol. 2, no. 1, pp. 118-139, January 2009.

[6] B. W. Silverman, "Density estimation for statistics and data analysis," *Monographs on Statistics and Applied Probability*, vol. 26. London: Chapman and Hall, 1986.

[7] P. Etyngier, S´egonne, R. Kwriven, "Shape Prior using Manifold Learning Techniques," *in Proc. IEEE International Conference on Computer Vision*, vol. 15, pp. 132-141, Oct 2007.

[8] V. Chandola, A. Banerjee, V. Kumar, "Anomaly detection: a survey," *ACM comput. Surv.*, Vol. 41, no. 3, pp. 15:1-15:58, July 2009.

[9] J. F. Liring, P. S. Wang, H. C. Chen, B. W. Soong, W. Y. Guo, "Differences between Spinocerebellar Ataxias and Multiple System Atrophy-Cerebellar Type on Proton Magnetic Resonance Spectroscopy," PLoS ONE 7(10): e47925. doi:10.1371/journal.pone.0047925.