DIFFUSION MAPS

Ivanovskaya Y.M.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Skobtsov V.Y. – PhD in engineering, Rogachevskaya A.I. – Senior Lecturer

Annotation. This paper aims to introduce one of the methods of learning a manifold, called the diffusion map. This method allows us to understand the underlying geometric structure of multidimensional data, as well as reduce dimensions if necessary, by carefully capturing the non-linear relationships between the original dimensions.

Keywords. Diffusion maps; Markov chains; Connectivity; Dimensionality reduction; Eigenmaps; Gaussian kernel

The "Curse of dimensionality" is a well-known problem in the modern scientific world. This term describes the reduction of high-dimensional data to low-dimensional data in order to optimize work with them: increase productivity, get more accurate results, and simplify their analysis.

The new representation must correctly describe the data, for example, by preserving some interesting quantities, such as local mutual distances. One of the problems of dimensionality reduction is the problem of finding meaningful structures in data sets, because it is quite difficult to extract the functions that led to obtaining this data [5].

To solve this task, structures such as graphs (combined with Markov chains), "kernel eigenmap methods", Laplacian eigenmaps, Hessian eigenmaps and local tangent space alignment are used. However, in this paper we describe Diffusion maps as a general structure of the above-mentioned particular methods [1-4].

Diffusion maps are a non-linear technique. It transforms data to a lower-dimensional space, so that the Euclidean distance between points approximates the diffusion distance in the original feature space. The dimension of the diffusion space is determined by the geometric structure underlying the data, and the accuracy by which the diffusion distance is approximated [2-3].

Diffusion maps represent the relationship between heat diffusion and random Markov chain walk. The basic observation is that if we take a random walk on the data, walking to a nearby data-point is more likely than walking to another that is far away.

The connectivity between two data points, x, and y, is defined as the probability of jumping from x to y in one step of the random walk and is

$$\text{connectivity}(x,y) = p(x,y), \tag{1}$$

Usually, this probability is specified in terms of a kernel function of the two points. For example, the popular Gaussian kernel:

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

Now we define a row-normalized diffusion matrix, P. Mathematically, this is equivalent to the transition matrix in the Markov chain. While P denotes the probability (or connectivity in this case) of single hopping from point x to point y, P^2 denotes the probability of reaching y from x in two hops and so on. As we increase the number of hops or P^t for increasing values of t, we observe that the diffusion process runs forward. Or in other words, the probability of following the geometric structure increases.

The diffusion maps allow to achieve dimensionality reduction, and the dimension of the embedding depends on both the geometry and the topology of the data set. In particular, if X is a discretized submanifold, the dimension of the embedding can be different from that of the submanifold.

References:

1. Zucker, S.W. Diffusion maps and geometric harmonics for automatic target recognition / S.W. Zucker, R. Coifman. – New Haven : Yale University, 2004. - 66 p.

2. Ihler. A. Nonlinear Manifold Learning (MIT 6.454 Summary) / A. Ihler. – Massachusetts : Massachusetts Institute of Technology, 2003. – 5 p.

3. Roweis, S.T. Nonlinear dimensionality reduction by locally linear embedding / S.T. Roweis, L.K. Saul. – London : University College London, 2000. – P. 2323-2326.

4. Belkin, M. Laplacian eigenmaps for dimensionality reduction and data representation / M. Belkin, P. Niyogi. – Chicago : University of Chicago, 2003. – P. 1373-1397.

5. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps / R.R. Coifman [et al.], - Washington : Proc. Natl. Acad. Sci. 2004 – P. 7426–7431.