Pattern Recognition Letters 77 (2016) 8–13

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Analyzing documents with Quantum Clustering: A novel pattern recognition algorithm based on quantum mechanics

Ding Liu[a,b,*], Minghu Jiang[b,**], Xiaofang Yang[b], Hui Li[c]

[a] Department of Computer Science and Technology, School of Computer Science & Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Xiqing District, Tianjin 300387, China
[b] Laboratory of Computational Linguistics, School of Humanities, Tsinghua University, Haidian District, Beijing 100084, China
[c] Institute of Computer Science, Heidelberg University, Im Neuenheimer Feld 348, Heidelberg 69120, Germany

ARTICLE INFO

Article history:
Received 14 July 2015
Available online 24 March 2016

Keywords:
Quantum clustering
Text analysis
Text clustering

ABSTRACT

The article introduces Quantum Clustering, a novel pattern recognition algorithm inspired by quantum mechanics and extend it to text analysis. This novel method improves upon nonparametric density estimation (i.e. Parzen-window), and differentiates itself from it in a significant way, Quantum Clustering constructs the potential function to determine the cluster center instead of the Gaussian kernel function. Specifically, detailed comparative analysis shows that the potential function could clearly reveal the underlying structure of the data that the Gaussian kernel could not handle. Moreover, the problem of parameter estimation is solved successfully by the numerical optimization approach (i.e. Pattern Search). Afterwards, the results of detailed comparative experiments on three benchmark datasets confirms the advantage of Quantum Clustering over the Parzen-window, and the additional trial on authorship identification illustrates the wide application scope of this novel method.

1. Introduction

Text clustering (document clustering) is a process of automatic document organization which is also representative of the pattern recognition problem. It plays a crucial role in text mining and, to this point, has been subjected to intensive examination [11,27,28]. Various clustering algorithms have been developed based on different principles, such as the hierarchical and partitional algorithms. However, researchers are still interested in developing new methods to promote the research on text clustering. Recently, a novel approach known as Quantum Clustering (QC) has emerged which derives from principles of quantum mechanics [9,10]. This novel method is developed based upon the conventional approach of density estimation and is considered as a new approach to nonparametric clustering methods [18]. As a result, the researchers will ask "Whether it could be applied to text analysis? And what about its performance on real task in comparison with the classical density estimation approach?" This paper focuses on introducing this novel method into text analysis and giving a clear and detailed comparative analysis between Quantum Clustering and the conventional density estimator, in order to show the real advantage of Quantum Clustering and answer these questions.

The most important conventional density estimation is referred to as the Parzen-window estimator. It is considered the main approach in the nonparametric family. In Ref. [18], the nonparametric methods approximate the probability density function without any underlying model assumption, and usually associate a kernel function to each data sample. Typically, the commonly used Gaussian kernel is regarded as the kernel function, and this kind of kernel function depends on a single parameter (i.e. the width parameter). Based on the Parzen-window estimator, the Quantum Clustering is developed.

The thinking of Quantum Clustering is inspired by the fundamental physics principles (i.e. quantum mechanics). Different from the Parzen-window estimator, Quantum Clustering constructs the potential function to estimate the density distribution of the data points instead of the Gaussian kernel, since the potential function has the potential to reveal the underlying structure of the data. Due to these advantages, some researchers began to apply this method to image segmentation, signal processing, etc. [17,18]. Furthermore, the Dynamic Quantum Clustering (DQC) has been proposed [22–24] and successfully employed in analyzing big,

* Corresponding author at: Department of Computer Science and Technology, School of Computer Science & Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Xiqing District, Tianjin 300387, China. Tel.: +86 15822510163; fax: +86 22 58685358.
** Corresponding author. Tel.: +86 13520115507; fax: +86 10 62785736.
E-mail addresses: dingliu_thu@126.com (D. Liu), jiang.mh@mail.tsinghua.edu.cn (M. Jiang).

http://dx.doi.org/10.1016/j.patrec.2016.03.008
0167-8655/

complex, real-world datasets obtained from various fields, such as X-ray nano-chemistry, condensed matter, seismology, biology, finance [25], and information retrieval [5,6]. And also, researcher have focus on speeding it up by graphics processor [26]. Based on these works, we extended QC to text analysis in Ref. [15]. Furthermore, in this paper, detailed and full-scaled comparative analysis between QC and Parzen-window estimator were conducted. Moreover, we also compared QC with the common used DBSCAN algorithm. And, the problem of parameter estimation was solved successfully by the numerical optimization approach (i.e. Pattern Search). All of these have not been subjected to rigorous analysis in previous works. Experimental results show the QC outperformed the Parzen-window significantly.

2. Method

2.1. Principle of Quantum Clustering

The inspiration of Quantum Clustering derives from the analogy that exists between data points and particles in a certain state. According to the postulate of quantum mechanics, a quantum system evolves in space and time following the Schrödinger differential equation. The Schrödinger equation describes the evolutionary process of a quantum mechanics system and is specified by a wave function. The Schrödinger equation can be written as various formulations in different context, and the time-independent version is given by Eq. (1) [8]

$$H\psi(x) = \left(-\frac{\hbar^2}{2m}\nabla^2 + v(x)\right)\psi(x) = E\psi(x) \tag{1}$$

where H is the Hamiltonian operator, E is the eigenvalue energy level, $\hbar$ and m denote the Reduced Plank Constant and the mass of a particle respectively. The function $\psi(x)$ refers to the so-called wave function and corresponds to the eigenstate of the given quantum system. The function $v(x)$ denotes the potential function, and ∇^2 is the Laplacian. Conventionally, the potential $v(x)$ is given and the equation is solved to find solutions $\psi(x)$. Such a function $\psi(x)$, can be assimilated with the kernel-based sum which depends on the given data points.

Different from quantum physics, where we want to estimate the location of particles given their potential function $v(x)$, in Quantum Clustering we solve the inverse problem. By considering the wave function $\psi(x)$ as an known condition of the Schrödinger equation, we aim to determine the potential $v(x)$ in Quantum Clustering, which characterizes the data probability density function[18]. Similar to particles in quantum physics, data points that are located in close proximation have close potential values. In this case, we can observe the analogy between clusters of data points and an electron cloud of hydrogen atoms illustrated in Fig. 1. The atomic nucleus corresponds to the cluster center, and each position where the electron probably appears (white dot) corresponds to each data point. According to the laws of quantum mechanics, the atomic system is in the lowest energy level (ground state) when the electron stays in the lowest orbit. Conversely, if the electron transits to a higher orbit, the system will be activated to the higher energy state (excited state). In a similar way, the values of potential function $v(x)$ calculated from the highest density data points reach the bottom, and increase with the decline of density of data points. Therefore, the cluster center could be revealed by the minima of $v(x)$. That is exactly the most significant characteristic of $v(x)$ as required.

2.2. Algorithm

The essential part of the algorithm is to calculate the potential function $v(x)$ by Schrödinger equation. Given Gaussian kernel as

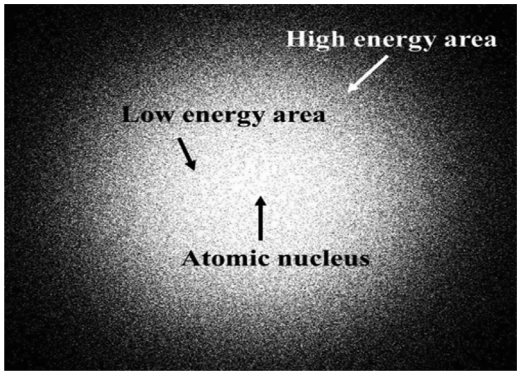

Fig. 1. Electron cloud diagram of hydrogen atoms (available at http://baike.haosou.com/doc/history/id/168383).

Eq. (2) for the wave function, and $m = \hbar^2/\sigma^2$, where σ denotes the width parameter.

$$\psi(x) = \sum_i e^{-(x-x_i)^2/2\sigma^2} \tag{2}$$

Afterwards, we solve Eq. (1) for $v(x)$ by next few steps. First, the Eq. (1) is rewritten as:

$$H\psi(x) = \left(-\frac{\sigma^2}{2}\nabla^2 + v(x)\right)\psi(x) = E\psi(x) \tag{3}$$

Then, the $v(x)$ could be solved as:

$$v(x) = E + \frac{\frac{\sigma^2}{2}\nabla^2\psi(x)}{\psi(x)} \tag{4}$$

Further, based on Eq. (1), the first-order derivative of $\psi(x)$ is:

$$\psi(x)' = \sum_i \left(e^{\frac{-(x-x_i)^2}{2\sigma^2}} \cdot -\frac{(x-x_i)}{\sigma^2}\right) \tag{5}$$

And the second-order derivative of $\psi(x)$ is:

$$\psi(x)'' = \sum_i \left(e^{\frac{-(x-x_i)^2}{2\sigma^2}} \cdot \frac{(x-x_i)^2}{\sigma^4} - e^{\frac{-(x-x_i)^2}{2\sigma^2}} \cdot \frac{1}{\sigma^2}\right) \tag{6}$$

Thus, the $v(x)$ could be solved as:

$$\begin{aligned} v(x) &= E + \frac{\sum_i (e^{\frac{-(x-x_i)^2}{2\sigma^2}} \cdot \frac{(x-x_i)^2}{2\sigma^2} - e^{\frac{-(x-x_i)^2}{2\sigma^2}} \cdot \frac{1}{2})}{\sum_i e^{\frac{-(x-x_i)^2}{2\sigma^2}}} \\ &= E - \frac{1}{2} + \frac{1}{2\sigma^2\psi(x)}\sum_i (x-x_i)^2 e^{\frac{-(x-x_i)^2}{2\sigma^2}} \\ &\approx \frac{1}{2\sigma^2\psi(x)}\sum_i (x-x_i)^2 e^{\frac{-(x-x_i)^2}{2\sigma^2}} \end{aligned} \tag{7}$$

where E is considered as constant, since they do not affect the topological structure of $v(x)$. Thus, the final version of $v(x)$ approximates to the last line in Eq. (7).

Generally, after we obtain the $v(x)$, some classic optimization approaches can be employed to deduce the clustering allocation, which is intended to locate the minima according to their topographic locations on the hypersurface of $v(x)$. In our study, we utilized the classic BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm which is a kind of Quasi-Newton methods [3] to address the problem. The details of the BFGS could be found in many academic literatures (e.g., [4,14]).

2.3. Potential function VS. Gaussian kernel

Like the Gaussian kernel function in Parzen-window estimator, the potential function formulate a hypersurface from the given