Quantum Physics Term Paper
Quantum Clustering Algorithm

Yi-Ting Lee
B03202017, Physics,
National Taiwan University

December, 2016

Abstract

Numbers of published theories are based on the observations of nature phenomenons, so as the algorithms in data analysis. The quantum clustering algorithm attempts to provide a different solution for clustering problems in the field of data analysis by applying the concepts in quantum mechanics and the formula of Schrodinger equation. In the study, I summarize the concepts of quantum clustering algorithm, implement the methods, and compare the results to the k-means clustering algorithm.

1 Introduction

Due to the improvement in computational resource, machine learning is more widely used in data analysis in recent years. Machine learning can be divided into many aspects. Based on its purpose, it can be categorized into: regression, classification, clustering, anomaly detection, etc.

This study will be focused on the clustering algorithm. In general cases, the k-means clustering algorithm is the simplest and most common method. However, data scientists have discovered the correlations between quantum mechanics and clustering problems in several aspects, such as follows,

1. Nonuniformity

   The probability of a particle appearing in different position is different, while the distribution density of the data points is not uniform.

2. Unordered

   At a certain time, the probabilities of a particle appearing in each particular positions is not sequential, and cannot be numbered, while there is no order in the data points.

3. High-dimension

   A particle appears in a 3 dimensional space in time-independent system, and can be represented as a vector or an eigenstate. The data points with m features can be seemed as existing in a m-dimensional feature space, and one data point can also be represented as a m-dimensional vector.

In 2001, David Horn and Assaf Gottlieb applied Schrodinger equation to clustering problems in [1] , and the method is called a quantum clustering algorithm. Several advanced algorithms based on quantum clustering also published after that.

2 Quantum Clustering and Schrodinger Equation

2.1 Schrodinger Equation

In quantum mechanics, the form of time-independent, non-relativistic Schrodinger equation is:

$$H\varphi\left(\overrightarrow{x}\right)=\left(-\frac{\hbar^2}{2m}\overrightarrow{\nabla}^2+V\left(\overrightarrow{x}\right)\right)\varphi\left(\overrightarrow{x}\right)=E\varphi\left(\overrightarrow{x}\right)$$

where

- $\varphi(\overrightarrow{x})$ is an eigenstate in Schrodinger equation,
- H is Hamiltonian operator,
- $V(\overrightarrow{x})$ is the potential energy of a particular position,
- E is the energy of the system.

The eigenstate, $\varphi(\overrightarrow{x})$ can be represented as the superposition of Gaussian distribution form:

$$\varphi(\overrightarrow{x}) = \sum_{i=0}^{\infty} A_i e^{-\frac{(x-x_i)^2}{2\sigma_i^2}}$$

where

- x is a vector(position) in Hilbert space,
- x_i are the $i-th$ means or expectations of the Gaussian distribution,
- σ_i is standard variation, describing how particle spreads out,
- A_i is the amplitude of a component.

2.2 Correspondence

In order to correspond to the data analysis, the eigenstate should be modified to:

$$\varphi(\overrightarrow{x}) = \sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

without considering the amplitude A_i, for each component, since each data points in the system is equally important, where

- x is a particular position in a high dimensional feature space,
- x_i can be presented as the $i-th$ data points in the feature space,
- σ can be viewed as a width adjustment parameter, and all values of σ are same for all data points,
- n is the number of data points, not infinite anymore.

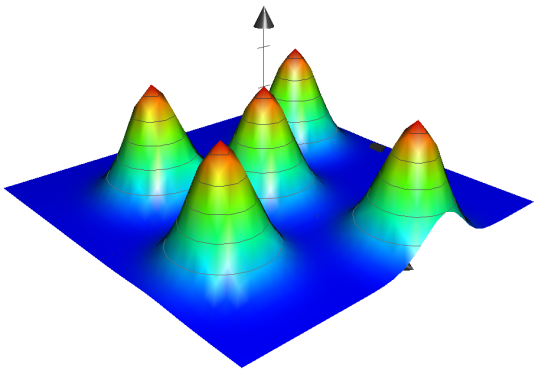

Figure 1: Data points in 2-dim feature space

Figure 1 shows 5 data points in 2-dim feature space, and each point is separable enough to others for visualization. The z-axis represents $\varphi(x, y)$. Since the amplitudes and σs are same for all data points, the altitude and shape of each point are also same.

Figure 1 also can correspond to the probability (without normalization) of a particle in 2-dim continuous space. The z-axis represents the probability. If any position, (x, y), is far from these 5 specific positions, then $\varphi(x, y)$ almost equals zero.

Since the data points are fixed, time-independent Schrodinger equation can describe the system. The potential energy, $V(\overrightarrow{x})$, won't vary over time, and energy, E, can even be seemed as a constant.

The key point to connect quantum mechanics and clustering problem is to assume that each local minimum of the potential energy is the cluster center. Since particles are more likely to appear at the regions with lower potential energy, it is similar to the situation that the data points in same cluster tend to gather in the certain region.

3 Definitions in Quantum Clustering

3.1 Data Points

In quantum clustering, there are n data points in the m-dimensional feature space. Thus, all data points can be represented by a matrix x, and its dimension is $n \times m$.

3.2 Distance Matrix

Distance matrix [2], D, with n × n dimension, describes the distance (degree of the difference) of any two data points, x_i and x_j. D obeys $0 \leq D(x_i, x_j) \leq 1$ and $D(x_i, x_i)$ always equals zero.

Definition

$$theta(x_{ip}, x_{jp}) = \begin{cases} 1 & x_{ip} = x_{jp} \\ 0 & x_{ip} \neq x_{jp} \end{cases}$$

$$Sim(x_i, x_j) = \sum_{p=1}^{m} theta(x_{ip}, x_{jp})$$

$$D(x_i, x_j) = 1 - \frac{Sim(x_i, x_j)}{m}$$

where,
x_i and x_j are the i-th and j-th data points,
x_{ip} and x_{jp} are the p-th dimension of x_i and x_j,
with $1 \leq i, j \leq n$.

β also needs to defined in advance as a threshold to determine if two data points are "similar" enough, by comparing with the indexes in distance matrix. Again, $0 \leq \beta \leq 1$.

3.3 Sigma

$$\sigma \equiv \left[\frac{4}{(m+2)\,n}\right]\left(\frac{1}{m+4}\right)$$

Sigma [2] contains the information of the data set, depending on the number of data points and the dimension, that is, n and m.

3.4 Potential Energy

$V(\overrightarrow{x})$ can be determined by the following equations.
Since

$$\left(-\frac{\sigma^2}{2}\overrightarrow{\nabla}^2 + V(\overrightarrow{x})\right)\varphi = E\varphi$$

$$V(\overrightarrow{x}) = \frac{E\varphi + \frac{\sigma^2}{2}\overrightarrow{\nabla}^2\varphi}{\varphi} = E + \frac{\sigma^2}{2}\frac{\overrightarrow{\nabla}^2\varphi}{\varphi}$$

with σ and φ are known once data set is given, and:

$$\overrightarrow{\nabla}^2\varphi = \overrightarrow{\nabla}^2 \sum_{i=0}^{n} e^{-\frac{(x-x_i)^2}{2\sigma^2}} = \overrightarrow{\nabla} \sum_{i=0}^{n} \frac{2\,(x-x_i)}{2\sigma^2} \times e^{-\frac{(x-x_i)^2}{2\sigma^2}}$$

$$= \sum_{i=0}^{n} \frac{4\,(x-x_i)^2}{4\sigma^4} \times e^{-\frac{(x-x_i)^2}{2\sigma^2}} = \frac{1}{\sigma^4} \sum_{i=0}^{n} (x-x_i)^2 \times \varphi$$

Thus,

$$V\left(\overrightarrow{x}\right) = E + \frac{\sigma^2}{2\sigma^4\varphi} \sum_{i=0}^{n} (x-x_i)^2 \times \varphi = E + \frac{1}{2\sigma^2\varphi} \sum_{i=0}^{n} (x-x_i)^2\,\varphi$$

Now, the only unknown is the energy E. However, the goal at this stage is to find the local minimum of the potential energy, and because E is a constant, it can be ignored in this equation.

Thus, all parameters are known to obtain the minimum of the potential energy, and this equation will be used in the program in the next part.

4 Implement

4.1 Problem and Data Set

Problem Definition

Use the collected data to cluster the university students to determine if any two of them are in the same grade or not.

To simplify the problem, I selected a relatively small data set with n = 83 (83 students) and m = 7 (7 features). These data, from questionnaires, contain the information of the students in the university in 4 grades (4 clusters).

The features include total credits in one semester, spending how many hours in labs per week, earning how much money per month, etc. In addition, all the features will be normalized in the beginning in the program.

4.2 Clustering and Classification

To explain more accurately for the problem, one has to understand the differences of clustering and classification problems more carefully. The classification problem is to determine "what" one data point belongs to, while

the clustering is to determine if two data points are belong to the "same" cluster.

Thus, in this problem, the goal is to determine if two students are belong to the same grade, instead of deciding which grade one is in. The evaluation function in this part is also based on this definition.

4.3 Method

I use Python to implement these two algorithms with following steps:
Step1 Import data as x.
Step2 Data normalization.
Step3
(A) K-means clustering algorithm.
(B) the quantum clustering algorithm [1][2].
Step4 Compute performance.
Step5 PCA and visualization.

The more detailed are:
(A)
Use sklearn tool kit [3] to implement k-means clustering algorithm [4][5] and divide all the data into c_1 clusters. K-means algorithm clusters data based on the euclidean distance between the data points and EM algorithm [6] (run iterations until converge).

(B)

```
compute D and set a proper value for beta
compute sigma
compute V(xi) for all data points by the equation in 3.4
set c_2 = 0
while (number of x != 0){
    find x_vmin in x satisfying V(x_vmin) is the minimum of V(xi)
    use D to find out {x_vmin_k} in x, which is in the same cluster as x_vmin
    add {x_vmin_k} into the k-th cluster
    remove {x_vimn_k} from x
    add 1 to c_2
get c_2 clusters
```

The source code and data can be seen at https://github.com/tiffany70072/Quantum-Clustering. It can also be compiled to visualize the quantum clustering part as $\beta = 0.8$.

4.4 Evaluation and Result

The performance of two algorithms is strongly based on the parameters set by the users.

In the k-means algorithm [4][5], the number of clusters, c_1, need to be set in advance. In general situations, c_1 is no need to equal to the real number of clusters, "4" (4 grades) in this problem, to get the best performance. In addition, a serious problem of k-means algorithm is that the initial conditions are random, so the performance is different each time.

On contrary, β is the corresponding parameters in quantum clustering. The bigger β is, more data points will be grouped into the same clusters, since the model can tolerate the "bigger difference" in the same cluster. Thus, c_2 will be smaller. Again, c_2 is no need to equal 4.

The result is shown in Table 1.

Algorithm	k-means clustering					
c_1	4	4	4	8	8	8
RI	0.588	0.595	0.592	0.679	0.671	0.673
Algorithm	quantum clustering					
β	0.0	0.2	0.4	0.6	0.8	1.0
c_2	56	54	54	17	6	1
RI	0.620	0.634	0.649	0.631	0.608	0.264

Table 1: Performance of two algorithms

The evaluation function is defined as Rand measure [7], (simplified as RI), which can get a lower accuracy if $c_i = 1$ or n, and get a higher performance if $1 < c_i < n$, corresponding to general situation and intuition. On contrary, F-measure[7] is not suitable at this stage.

$$RI \equiv \frac{TP + TN}{TP + TN + FP + FN}$$

where
TP is the number of true positives (x_i and x_j are in the same cluster, while the result from algorithm is "yes"),
TN is the number of true negatives,
FP is the number of false positives (x_i and x_j are "not" in the same cluster, while the result from algorithm is "yes"),
and FN is the number of false negatives.

4.5 Visualization

One can understand the differences between the algorithms better by visualizing the result.

All data points are visualized after the process of PCA (Principle Components Analysis [8]) to display on 2-dimensional space. PCA can convert a set of high-dim vectors into 2-dim vectors with losing the least information.

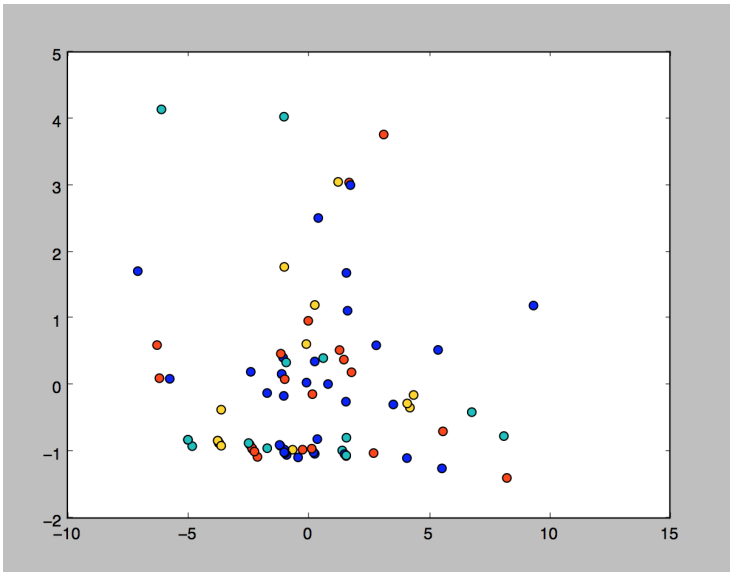

Figure 2: Visualization of real clusters

There are 4 clusters in the data set, and each cluster is colored by different colors in the figure.

Figure 3 visualizes the best result of k-means clustering when $c_1 = 8$. The method is based on the distance of each data points. Thus, the data points in similar region are more likely to be grouped as a same cluster. However, the data points has been mapped to 2-dim space by PCA after k-means clustering, so the shape won't be an ordinary circle. For instance, the area of the "yellow" points is shaped like a triangle.

Figure 4 visualizes the result of quantum clustering when $\beta = 0.8$ and $c_2 = 6$ for simplicity, instead of the best result. It computes the potential energy and find the local minimum. Although the result seems to be more different from figure 2 than k-means did, quantum clustering really "find out some connections" that k-means couldn't do by only calculating the distance. For instance, The "yellow" points in figure 4 can be compared to the "blue" points in figure 2, and the "blue" point in figure 4 can be compared to the "red" points in figure 2.

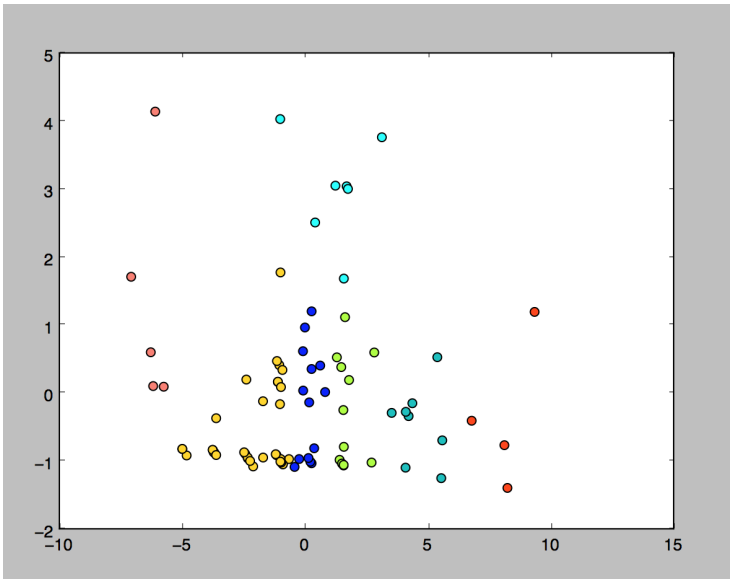

Figure 3: Visualization of k-means clustering

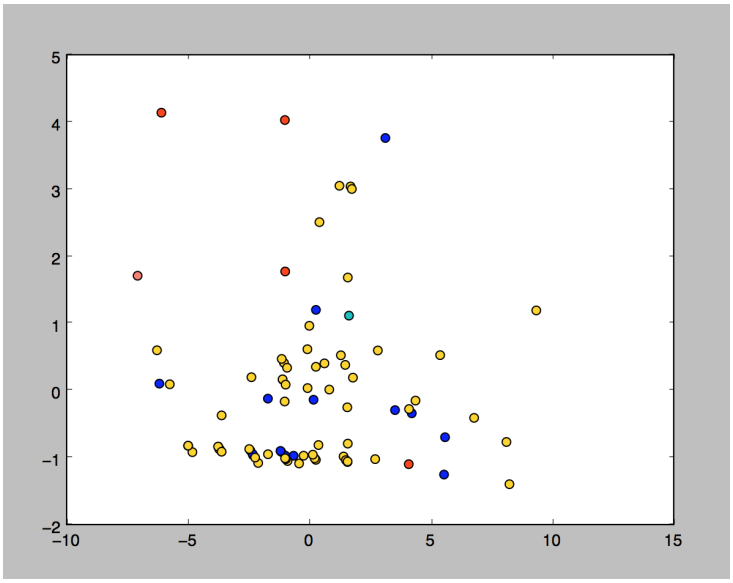

Figure 4: Visualization of quantum clustering

5 Conclusion

One can get several conclusions from the result in Table 1.

1. The result is strongly depends on c_i we chose.

2. Although the real number of clusters is 4 in this data set, c_i (the number of clusters) of the best performance of each algorithm is not equal to 4. This is corresponding to the general situation. That is, $c_1 = 8$ and $c_2 = 54$ can get a higher accuracy than others shown in the table 1, with 54 is far from 4.

3. The quantum clustering doesn't have randomness problem (, so only 1 result for each β is shown).

4. Randomness is always a problem in k-means clustering. This problem will be slighter if the program runs more iterations. I set the iterations = 500, which is a generally accepted number in relatively small data set.

   On contrary, the problem will be more obvious especially in extremely nonuniform data set. It can even result in that there isn't any data point in specific clusters, and this situation doesn't appear this time.

5. The performance of each algorithm is at about 60% to 68%, which is not very satisfying. The main reason is that the features are not good enough. The goodness of features determine the upper bound of the performance, and the goodness of the algorithms can only determine "how close" to the upper bound.

In this data set, the performance of k-means clustering (67% to 68%) is slightly better than the performance of quantum clustering (about 65%) if a proper c_i is chosen.

However, several ideas are worth trying in the future:

1. Maybe some specific data sets are more suitable for quantum clustering algorithm. That is, one can apply quantum clustering only if one can make sure that the properties of specific data set correspond well to the properties of the real particles in quantum mechanics.

2. One can modify the distance matrix, D, to avoid tricky problems, since D measures the distance only by comparing if the coordinate of each axis is "same" or not.
   Set point A = (10, 20), B = (11, 21), C = (10, 30). Then D considers

that A and C are more similar than A and B, because one of their coordinate are same.
In addition, this matrix is suitable for the data with discrete features. Assume the values of one feature belongs to $\{0, 1\}$, then it is more likely for the data points to have same value. On contrary, if another feature spreads from 0, 0.001, 0.002,... 0.999, to 1, then it is difficult for data points to have the same value.
Thus, it may cause unexpected mistakes and performance gets worse.

3. Scientists have proposed several improved theories for quantum clustering algorithm. One of them is *Modified categorical quantum clustering* [2]. It can deal with the labels which are not all independent with each other. Take our data set for example, there are 4 labels in it, corresponding to 4 grades of the students. However, the relation between grade 1 and grade 2 is different from the one between grade 1 and grade 4, and this algorithm can deal with this kind of property.

References

[1] David Horn and Assaf Gottlieb, *Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics*, (2002)

[2] Zhao Zheng-Tain, Zhao Xiao-Qiang, and Li Wei, *Improved clustering algorithm for categorical attribution data by using quantum mechanics*, (2009)

[3] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[4] https://www.coursera.org/learn/machine-learning/home/week/8.

[5] https://en.wikipedia.org/wiki/K-means_clustering.

[6] 4.3.2, 4.4.2 of X. Huang, A. Acero, H. Hon, *Spoken Language Processing,* Prentice Hall, (2001).

[7] https://en.wikipedia.org/wiki/Cluster_analysis#External_evaluation.

[8] Lecture 13 in http://speech.ee.ntu.edu.tw/DSP2016Autumn/