A Statistical Clustering Model and Algorithm

Yang Guangwen Zheng Weimin Wang Dingxing

Department of Computer Science and Technology
Tsinghua University, Beijing, P.R.China

Abstract

In this paper, a statistical clustering model and algorithm have been discussed. Finding the optimal solution to clustering problem is transformed into simulating the equilibrium state of a physical system., and then the equilibrium state of physical system is simulated by solving a series of problems to minimize the free energy which varies with temperature and attains the ground state of the system. Moreover, a great number of simulating examples make it clear that the clustering algorithm can be widely used, especially for the problem to which the traditional clustering algorithms are helpless.

Key Words Clustering, Deterministic Annealing, Free Energy, The Principle of Maximum Entropy

1. Introduction

Clustering is an important problem which can be found in many applications where a priori knowledge about the distribution of the observed data is not available. It is being applied in a variety of engineering and scientific disciplines such as pattern recognition, source coding, image and signal processing, computer vision, a machine learning and remote sensing.

The traditional clustering model and algorithms have some defects. For example, the clustering results are highly sensitive to the initialization; they perform poorly if the data contains overlapping clusters; no interactions among clusters are taken into consideration; the most urgent problem is the lack of cluster validity criteria; all the algorithms tend to create clusters even when no nature clusters exist in the data; the obtained results are not global optimal. Therefore, it is necessary to find a new clustering model and algorithm.

The problem of traditional partitional clustering can be formally stated as follows[1]. Given N patterns $X = \{x_1, x_2, \ldots x_N\}$ in a n-dimensional metric space, determine a partition of the patterns into L groups, or clusters $C_1, C_2, \ldots, C_L$ (in

general $C_i \cap C_j = \varnothing$ (i≠j), and $\sum_{i=1}^{M} C_i = X$), such that the patterns in a cluster are more similar to each other than to patterns in different clusters. The value of L may or may not be specified. A criterion must be adopted. The traditional clustering model solves the following optimization problem:

$$\min D = \sum_{j} \sum_{x_i \in C_j} (d(x_i, y_j))^2 \qquad (1.1)$$

where y_j is the center of cluster (representative) of C_j (j=1,2,...,L). $x \in C_j$ if and only if $d(x, y_j) \le d(x, y_i)$ $(\forall i)$(if $d(x, y_j) = d(x, y_i)$ and $i<j$, $x \in C_i$). ($d(x, y_j)$ is a distortion measure of x and y_j) .

The objective function of problem (1.1) is a non-convex optimization problem, there is no efficient algorithms for it at present. Many researchers have studied the method for non-convex optimization problem.

"Physical Computation" is proposed by G.C.Fox[2] . It encompasses a variety of ideas that can be loosely classified as the use of physical analogies or methods from the physical science to problems outside their normal domain of applicability. We can view physical computation as the use of physical methods to describe general complex system. Physical computation centers on computation and computer science, seeking the natural law and succeeding in an attempt at physical model of problem. Now it has been widely used in artificial intelligence and many other fields. Neural network, genetic algorithms , deterministic annealing and simulated annealing are the branches of physical computation. Some traditional methods for discrete optimization have time complexities that scale exponentially in problem size while physical computation is often essentially linear.

Using the annealing process in statistical physics, we shall discuss a statistical clustering model and algorithm.

2. Deterministic Annealing Method

Deterministic annealing is proposed by K. Rose[3]. Using the processes of statistical physics, deterministic annealing considers an optimization problem as a physical system . Finding the optimal solution to optimization problem is transformed into solving a series of problems to minimize the free energy and obtain the global minimum which varies with temperature. The properties of solution to deterministic annealing have been discussed in detail in paper [4].

For a minimum problem:

$$\min E(x) \qquad (2.1)$$

where x may be continuous, discrite or mixture. $E(x)$ is considered as an energy function of a physical system. Since the non-convexity of $E(x)$, there is no effective algorithm to find the global optimal solution. We take problem (2.1) as the one to

find the state at which the energy function of physical system is minimum. From statistical physics, we know that the state of system varies along the direction at which free energy decreases and it will reach minimum when the system is at equilibrium state. Let $F(x,T)$ be the free energy of system under temperature T. The method of deterministic annealing is a process which simulates the equilibrium state of system under temperature T by solving the problem $\min F(x,T)$ which uses $x_{\min}(T+\Delta T)$ (the minimum solution to $\min F(x,T+\Delta T)$) as the initial point of algorithm. We have already proved in paper [4] that the global optimal solution $x_{\min}(T)$ to $\min F(x,T)$ is a continuous map about T under certain conditions. With the decrease of T, the global optimal solution to $\min F(x,T)$ varies contineously, and $x_{\min}(T)$ is located in the local minimum interval of $\min F(x,T)$ in which $x_{\min}(T+\Delta T)$ is located. So we can guarantee that $\lim_{T\to 0} x_{\min}(T)$ is the global optimal solution to problem (2.1) when the decreasing speed of T is reasonable.

We suppose further that the global optimal solution to $\min F(x,\infty)$ can be found easily and $F(x,0)=E(x)$.

Deterministic annealing is an annealing process, a global optimal solution may be obtained, but it is different from the simulated annealing.

3. A Statistical Clustering Model

For a clustering problem, we do not know the number of clusters. So we introduce the concept of "computation number" of clusters (denoted by M), which is greater than the real number of clusters. For given representatives $y_1, y_2, \ldots, y_M$ and $j_i \in \{1,2,\ldots,M\}$, $[x_i \in C_{j_i}]$ represents the probability event of $x_i \in C_{j_i}$, the probability of event $\bigcap_i [x_i \in C_{j_i}]$ is:

$$p(y_{j_1}, y_{j_2}, \ldots, y_{j_N}) = p(x_1 \in C_{j_1}, x_2 \in C_{j_2}, \ldots, x_n \in C_{j_n}) \qquad (3.1)$$

Clustering problem can be described as the one to find $y_1, y_2, \ldots, y_M$ and $p(y_{j_1}, y_{j_2}, \ldots, y_{j_N})$ $(j_i = 1,2,\ldots,M,\ i = 1,2,\ldots,N)$, such that

$$\sum_{j_1 j_2 \ldots j_N} p(y_{j_1}, y_{j_2}, \ldots, y_{j_N}) D(y_{j_1}, y_{j_2}, \ldots, y_{j_N}) \qquad (3.2)$$

is minimum. Where $D(y_{j_1}, y_{j_2}, \ldots, y_{j_N})$ is a distortion measure of $x_i \in C_{j_i}$ (i=1,2,...,N), and it depends on the background of clustering problem. We select the form of $D(y_{j_1}, y_{j_2}, \ldots, y_{j_N})$ as follows for the sake of convenience:

$$D(y_{j_1}, y_{j_2}, \ldots, y_{j_N}) = d(y_{j_1}, y_{j_2}, \ldots, y_{j_N})^T A d(y_{j_1}, y_{j_2}, \ldots, y_{j_N})$$

where $A=[a_{kl}(y_{J_1},y_{J_2},\ldots y_{J_N})]_{N\times N}$ is a $N\times N$ matrix map of $y_{J_1},y_{J_2},\ldots,y_{J_N}$, $d(y_{J_1},y_{J_2},\ldots,y_{J_N})$ is a N-dimensional vector map of $y_{J_1},y_{J_2},\ldots,y_{J_N}$

$$d(y_{J_1},y_{J_2},\ldots,y_{J_N})=(d(x_1,y_{j_1}),d(x_2,y_{J_2}),\ldots,d(x_N,y_{J_N}))^T$$

The selection of A contains some interactions among patterns and clusters.

If all the $[x_i\in C_j]$ are independent, let the probability of event $[x_i\in C_j]$ be $p(x_i\in C_j)$. $p(y_{J_1},y_{J_2},\ldots,y_{J_N})=\prod_{i=1}^{N}p(x_i\in C_j)$, $=\prod_{i=1}^{N}p_{ij_i}$, we select

$$A=\begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{NN} \end{bmatrix}$$

In general, $a_{ii}>0$. The difference of $a_{ii}(i=1,2,\ldots,N)$ indicates the difference of significance to each pattern in X. In this case, we have

$$D(y_{J_1},y_{J_2},\ldots,y_{J_N})=\sum_{i=1}^{N}a_{ii}(d(x_i,y_{j_i}))^2$$

and

$$\sum_{J_1,J_2,\ldots,J_N}p(y_{J_1},y_{J_2},\ldots,y_{J_N})D(y_{J_1},y_{J_2},\ldots,y_{J_N})=\sum_{i=1}^{N}\sum_{j_i}a_{ii}p_{ij_i}(d(x_i,y_{j_i}))^2$$

Let $a_{ii}=1\quad(\forall i)$, then clustering model (3.2) is the case of Rose [3]. If

$$p_{ij_i}=p(x_i\in C_{j_i})=\begin{cases}1 & x_i\in C_{j_i}\\ 0 & otherwise\end{cases}$$

then

$$\sum_{J_1,J_2,\cdots J_N}p(y_{J_1},y_{J_2},\cdots,y_{J_N})D(y_{J_1},y_{J_2},\cdots,y_{J_N})=\sum_{j}\sum_{x_i\in C_j}(d(x_i,y_j))^2$$

clustering model (3.2) is model (1.1), that is to say the clustering model (3.2) is simplified as the traditional model (1.1). Model (3.2) includes the interactions of clusters and patterns. Some traditional models are the special case of model (3.2).

4. Clustering Algorithm Using Deterministic Annealing

4.1 Definition of Free Energy

If all the $[x_i\in C_j]$ are independent, We consider the clustering problem as a physical system, T is the temperature, and the probability of event $[x_i\in C_j]$ be $p(x_i\in C_j)$.

Define the energy function and entropy function of system as follows:

$$E = \sum_{i=1}^{N}\sum_{j=1}^{M} a_{ii} p(x_i \in C_j)(d(x_i, y_{j_i}))^2$$

$$H = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i \in C_j)\ln p(x_i \in C_j)$$

For fixed temperature T, using the principle of maximum entropy, we know that the probability which maximizes the entropy H is the equilibrium state of physical system, and Y is the most probable configuration of clustering system.

For the problem:

$$\max \quad H$$

$$s.t. \quad \sum_{i=1}^{N}\sum_{j=1}^{M} a_{ii} p_i (d(x_i, y_j))^2 = E \tag{4.1}$$

Using the principle of variation, we get

$$p(x_i \in C_j) = \frac{e^{-\beta a_{ii}(d(x_i, y_j))^2}}{\sum_{k=1}^{M} e^{-\beta a_{ii}(d(x_i, y_k))^2}} \tag{4.2}$$

where β is determined by (4.1) and $\beta \propto \frac{1}{T}$.

Define the free energy function

$$F(y_1, y_2, \cdots y_M, \beta) = \begin{cases} \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M} a_{ii}(d(x_i, y_j))^2 & \beta = 0 \\ -\frac{1}{\beta}\sum_{i=1}^{N} \ln \sum_{j=1}^{M} e^{-\beta a_{ii}(d(x_i, y_j))^2} & 0 < \beta < +\infty \\ \sum_{i=1}^{N}\sum_{j=1}^{M} a_{ii}(d(x_i, y_j))^2 & \beta = +\infty \end{cases} \tag{4.3}$$

Let $d(x, y_j) = \|x - y_j\| = (\sum_{k=1}^{n}(x(k) - y_j(k))^2)^{\frac{1}{2}}$, $(d(x, y_j))^2$ be differential convex function. The free energy function satisfies the requests of deterministic annealing.

By the definition of $F(y_1, y_2, \ldots, y_M, \beta)$, $F(y_1, y_2, \cdots, y_M, 0)$ is a continuous differential convex function about $(y_1, y_2, \ldots, y_M)$, we can find the global optimal solution easily with the help of traditional optimization algorithm. From paper [4] the global optimal solution to $F(y_1, y_2, \ldots, y_M, \beta)$ is a continuous map about $(y_1, y_2, \ldots, y_M)$ when

$\beta \in [0,+\infty)$. At the minimum point of $F(y_1, y_2, \ldots, y_M, \beta)$, $\frac{\partial F(y_1, y_2, \cdots, y_M, \beta)}{\partial y_j} = 0\ (\forall j)$, and we get:

$$y_j = \frac{\sum_{i=1}^{N} a_{ii} x_i p(x_i \in C_j)}{\sum_{i=1}^{N} a_{ii} p(x_i \in C_j)}$$

where $p(x_i \in C_j)$ is determined by (4.2).

When $\beta=0, p(x_i \in C_j) = \frac{e^{-\beta a_{ii} d(x, y_j)}}{\sum_{k=1}^{M} e^{-\beta a_{ii} d(x_i, y_k)}} = \frac{1}{M}$, and $y_1 = y_2 = \cdots = y_M = \frac{\sum_{i=1}^{N} a_{ii} x_i}{\sum_{i=1}^{N} a_{ii}}$, it is the global optimal solution to $F(y_1, y_2, \ldots, y_M, \beta)$.

4.2 A Stochastic classification method

For some β, we obtain $y_1, y_2, \cdots, y_M$ and $p_j = p(x \in C_j)$ (j=1,2,...,M), and then how to determine which cluster each $x \in X$ belongs to. We will depend on the probability $p_j = p(x \in C_j)$ to classify X.

For each $x \in X$, divide intervals [0,1] into M small interval according to $p_j = p(x \in C_j)$ (j=1,2,...,M):

$$[0, p_1), [p_1, p_1 + p_2), [p_1 + p_2, p_1 + p_2 + p_3), \cdots, [\sum_{i=1}^{M-2} p_i, \sum_{i=1}^{M-1} p_i), [\sum_{i=1}^{M-1} p_i, \sum_{i=1}^{M} p_i]$$

Select a random variable ξ which is an uniform distribution on [0,1]. Define a random variable η: $[\eta = k]$ if and only if $\xi \in [\sum_{i=1}^{k-1} p_i, \sum_{i=1}^{k} p_i)$. If $\eta = k$, we classify x into C_k, this process is going on until all the elements in X are classified into each cluster. This method is supported by the conclusion $p(\eta = k) = p_k$, since

$$p(\eta = k) = \int_{I_k} 1dx = \int_{\sum_{i=1}^{k-1} p_i}^{\sum_{i=1}^{k} p_i} 1dx = \sum_{i=1}^{k} p_i - \sum_{i=1}^{k-1} p_i = p_k .$$

This classification method is superior, especially concerning clustering problem which involves interactions.

4.3 A Clustering Algorithm

Let $d(x, y_j) = \|x - y_j\| = (\sum_{k=1}^{n} (x(k) - y_j(k))^2)^{\frac{1}{2}}$, the algorithm is:

(1) let $\beta_0 = 0, y_1^{(k)} = y_2^{(k)} = \cdots = y_M^{(k)} = \frac{\sum_{i=1}^{N} a_{ii} x_i}{\sum_{i=1}^{N} a_{ii}}$,k=0.

(2)u is a monotonous increase function , $\beta_{k+1} = u(\beta_k)$. $(\bar{y}_1^{(0)}, \bar{y}_2^{(0)}, \ldots, \bar{y}_M^{(0)}) = (y_1^{(k)}, y_2^{(k)}, \ldots, y_M^{(k)})$, using the formula

$$\bar{y}_j^{(s+1)} = \frac{\sum_{i=1}^{N} x_i a_{ii} p(x_i \in C_j^{(s)})}{\sum_{i=1}^{N} a_{ii} p(x \in C_j^{(s)})} \qquad (s=1,2,3,\ldots)$$

where $p[x_i \in C_j^{(s)}] = \frac{e^{-\beta a_{ii} (d(x_i, \bar{y}_j^{(s)}))^2}}{\sum_{k=1}^{M} e^{-\beta a_{ii} (d(x_i, \bar{y}_k^{(s)}))^2}}$. Let $(y_1^{(k+1)}, y_2^{(k+1)}, \ldots, y_M^{(k+1)})$ be the convergence point of the interation.

(3) If the stopping criterion is satisfied, $(y_1^{(k+1)}, y_2^{(k+1)}, \cdots, y_M^{(k+1)})$ is the optimal clustering center. Go to (5), otherwise go to (4).

(4) k=k+1, go to (2).

(5) Classify all the elements of X according to the method stated in **Section 4.2**, and then stop.

Notes:

(1) The selection of M. If we know the cluster number, and the diameters of each cluster are almost the same , we select that M be the cluster number, otherwise we selecte M which is large enough (at least larger than the real cluster number).

Clustering problems which traditional clustering algorithms are helpless can be solved easily by selecting large M (e.g. the problem having bridges, being linear indivisible, having unknown cluster numbers, having large difference of geometric size among clusters, etc.).

(2) In algorithm, the selection of stopping criterion depends on the real clustering problem. In general, if the clustering center is stable, the algorithm stops;

(3) In general, M is only the computing cluster number, and we can determine the real cluster number by using the message of algorithm in the computing process.

(4) When the real cluster number is determined, we can use some methods to classify all $y_1, y_2, \cdots, y_M$ to all clusters. Different y_j may represent the same cluster.

5. Conclusion

In this paper, we put forward a statistical clustering model and algorithm. Temperature parameter is introduced, the clustering problem as a physical system is considered, and the equilibrium state of the system is obtained under different temperatures by using the principle of maximum entropy. A free energy of clustering system is constructed and the clustering algorithm is introduced by using deterministic annealing . Finding the optimal solution to clustering problem is transformed into simulating the equilibrium state of a physical system., and then the equilibrium state of physical system is simulated by solving a series of problems to minimize the free energy which varies with temperature and attains the ground state of the system. Moreover, a great number of simulating examples make it clear that the clustering algorithm can be widely used, especially for the problem to which the traditional clustering algorithms are helpless.

References

[1] A.K.Jain and R.C.Dubes, Algorithms for Clustering Data, Prentice Hall, Inc.,1988.

[2]G.C.Fox, Physical computation, concurrency: practice and experience, Vol 3(6),627-653,1991.

[3]K.Rose, E.Gurewitz and G.C.Fox, Statistical mechanics and phase transition in clustering, Physical Review Letters,65:945-948,1990.

[4]Yang Guangwen et. al. Deterministic Annealing.Chinese Journal of Computer(to appear in vol.21, no.8,1998).