一种利用确定性退火技术的聚类模型与算法研究*

杨广文[1] 王鼎兴[1] 郑纬民[1] 李晓明[2]

[1](清华大学计算机科学与技术系 北京 100084)

[2](北京大学计算机科学系 北京 100871)

摘要 针对传统聚类模型的缺陷,文章利用确定性退火技术,提出一种聚类模型及聚类算法.该模型考虑了聚类的交互作用,以前提出的一些聚类模型是它的特例.引入温度参数,把聚类问题看成一个物理系统,把求解聚类问题的最优解转化为模拟随温度变化的物理系统的平衡态.通过求解一系列随温度变化的物理系统的自由能函数的局部极小来模拟物理系统的平衡态,最终达到物理系统的基态,即聚类问题的最优解.

关键词 确定性退火技术,聚类,自由能,极大熵原理.

中图法分类号 TP18

聚类分析是进行数据分析的重要技术,并被广泛地应用于许多工程和科学技术领域.聚类分析根据数据的内在性质将数据分成一些聚合类,每一聚合类中的元素尽可能地具有相同的特性.聚类分析是在对数据不作任何假设的条件下进行分析的工具.在人工智能和模式识别中,聚类分析亦称为"无先验学习",是机器学习中知识获取的重要环节.

划分聚类方法是通过对每个模式的标记,将模式组织成一些聚合类,它仅依赖于模式矩阵.划分聚类方法可以形式地描述为:给定 n 维模式空间的 N 个模式 $X=\{x_1,x_2,\ldots,x_N\}$,将 N 个模式分成 M 个模式子集 $C_1,C_2,\ldots,C_M$,使得在一个聚合类中的模式彼此十分类似,一般有 $C_i\cap C_j=\varnothing\ (i\neq j)$,且 $\bigcup_{i=1}^{M}C_i=X$.划分聚类方法基于一聚类准则:全局准则将每一聚合类描述为一代表元,对于每一模式,依据它与某一聚合类的代表元的相似程度决定它到底属于哪一个聚合类;而局部准则利用数据的局部结构将模式聚类,如,可通过模式空间中的高密度区域形成聚合类等,或将一个模式与它最邻近的若干个相邻模式作为一个聚合类.当给定聚类准则时,求类算法就是确定一种方法,在所有可能的模式的聚类中,求使聚类准则最优的聚类.

求解聚类问题的方法随着问题的不同而有所差异.如在信息论中,有对标量量化问题的 Lloyd 算法[1],以后又对该方法推广,得到求解向量量化问题的 GLA 算法[2].在模式识别中,有 ISODATA 算法[3]以及后来利用 Fuzzy 技术发展的一些算法[4,5].

传统的聚类模型,我们可归结为求解极小化问题:

$$\min D=\sum_{j}\sum_{x_i\in C_j}(d(x_i,y_j))^2. \tag{1}$$

由该优化问题的全局最优点 $y_1,y_2,\ldots,y_M$ 作为类心,可将 X 聚类.但是,由于式(1)为一个非凸优化问题,目前还没有十分有效的办法来求全局最优解.

传统聚类模型与聚类算法具有如下缺陷[6]:(1) 聚类结果对初始条件极为敏感,这为选择合适的聚类结果设置了障碍;(2) 目前的聚类算法所求得的聚类结果往往不是最优的;(3) 聚类模型往往不考虑模式间的交互作

---

* 作者杨广文,1963年生,博士,副教授,主要研究领域为并行计算,聚类分析.王鼎兴,1937年生,教授,博士生导师,主要研究领域为并行/分布计算.郑纬民,1946年生,教授,博士生导师,主要研究领域为并行/分布计算.李晓明,1957年生,博士,教授,博士生导师,主要研究领域为并行/分布计算.

本文通讯联系人:杨广文,北京 100084,清华大学计算机科学与技术系

本文1998-04-21收到原稿,1998-07-02收到修改稿

用;(4) 若数据包含交叠的类,传统的方法则无能为力;(5) 无法判断类的真实性与合法性.

聚类分析的研究吸引了众多学者,如何给出合理的聚类模型及聚类算法,是一个重要课题.

目前,国际上出现了一些"模拟大自然的某种客观规律来设计求解一些复杂系统相关问题"的算法,即"按自然法则计算"(physical computation). 它首先是由 G. C. Fox 提出的[7]. Fox 的定义是,"将大量的自然科学领域的思想与方法用于其传统应用领域之外的其他领域,将原思想与方法的本质提取出来,用于解决新领域中的问题". 确定性退火技术是美国加州理工学院 K. Rose 博士于 1990 年首先提出的[8],它是按自然法则计算的一个重要分支[7]. 文献[9]详细讨论了确定性退火技术,并得到了一些比较满意的理论结果.

1 确定性退火技术

在退火过程中,系统在每一温度下达到平衡态的过程都应遵循自由能减少定律,系统状态的自发变化总是朝着自由能减少的方向进行. 当自由能达到最小值时,系统达到平衡态.

确定性退火技术,正是基于上述思想. 对于求解极小化问题,

$$\min E=E(x). \tag{2}$$

这里,x 可以是连续的、离散的或混合的,$E(x)$被看做是某一系统的能量. 目前还没有十分有效的方法来求解式(2). 将极小化问题(式(2))看做是求解一物理系统能量极小的状态,首先构造一自由能函数 $F(x,T)$. 由以上分析可知,在某一温度下,系统状态的变化总是朝着自由能减少的方向进行,当系统达到平衡态时,自由能函数达到极小. 文献[9]证明了当 $F(x,T)$为连续映射时,它的全局极小点 $x_{\min}(T)$为 T 的连续映射. 设 $T=\infty$时,$F(x,T)$的全局最优点极易求出,而 $F(x,0)=E(x)$. 确定性退火技术,就是在每一温度 T,以系统在 $T=T+\Delta T$ 时自由能函数极小的状态 $x_{\min}(T+\Delta T)$作为初始点,通过求解 $\min F(x,T)$的极小点来模拟系统达到平衡态的过程. 随着 T 的减小,$F(x,T)$的全局极小点不断变化,当 T 的变化 ΔT 很小时,可认为 $x_{\min}(T)$位于 $x_{\min}(T+\Delta T)$所在的局部极小区域内,故在温度 T,可以用 $x_{\min}(T+\Delta T)$作为初始点求解 $F(x,T)$的极小值. 当 T 连续减小速度合理时,有理由认为 $\lim\limits_{T\to 0}x_{\min}(T)$为式(2)的问题的全局极小点.

2 聚类模型的一般形式

对于给定的 $y_1,y_2,\ldots,y_M$,$j_i\in\{1,2,\ldots,M\}$,$[x_i\in C_{j_i}]$表示 $x_i\in C_{j_i}$这一事件,$\bigcap\limits_i[x_i\in C_{j_i}]$的概率为 $p(y_{j_1},y_{j_2},\ldots,y_{j_N})=p(x_1\in C_{j_1},x_2\in C_{j_2},\ldots,x_n\in C_{j_n})$. 聚类问题可描述为求解 $y_1,y_2,\ldots,y_M$ 及 $p(y_{j_1},y_{j_2},\ldots,y_{j_N})(j_i=1,2,\ldots,M,i=1,2,\ldots,N)$,使得

$$\sum_{j_1,j_2,\ldots,j_N}p(y_{j_1},y_{j_2},\ldots,y_{j_N})D(y_{j_1},y_{j_2},\ldots,y_{j_N}) \tag{3}$$

极小. 其中 $D(y_{j_1},y_{j_2},\ldots,y_{j_N})$为某一度量描述(已知),是当 $x_i\in C_{j_i}(\forall\, i)$时的损失函数.

$D(y_{j_1},y_{j_2},\ldots,y_{j_N})$的形式可根据问题的实际背景来选取,为一个满足某些特性的一般函数. 可取 $D(y_{j_1},y_{j_2},\ldots,y_{j_N})=d(y_{j_1},y_{j_2},\ldots,y_{j_N})^TAd(y_{j_1},y_{j_2},\ldots,y_{j_N})$,式中 A 为 $y_{j_1},y_{j_2},\ldots,y_{j_N}$的一个 $N\times N$ 矩阵函数,$d(y_{j_1},y_{j_2},\ldots,y_{j_N})$为 N 维向量函数,取 $d(y_{j_1},y_{j_2},\ldots,y_{j_N})=(d(x_1,y_{j_1}),d(x_2,y_{j_2}),\ldots,d(x_N,y_{j_N}))^T$. A 的选取,可以包含 x 中点与点、点与类及类与类之间的一些交互作用. 特别是,若取 $A=[a_{ij}]_{N\times N}$,则有 $D(y_{j_1},y_{j_2},\ldots,y_{j_N})=\sum\limits_{i=1}^{N}\sum\limits_{k=1}^{N}a_{ik}d(x_i,y_{j_i}).d(x_k,y_{j_k})$,此式可理解为当$[x_i\in C_j](i=1,2,\ldots,N,j=1,2,\ldots,M)$不独立时损失函数的情形,式中包含了一些交互项.

若取 $a_{ij}=0(i\neq j)$,则此时所有$[x_i\in C_j]$相互独立,$a_{ii}(i=1,2,\ldots,N)$的不同,表示 x 中各点的重要程度不一样(一般 $a_{ii}>0$). 此时有 $D(y_{j_1},y_{j_2},\ldots,y_{j_N})=\sum\limits_{i=1}^{N}a_{ii}(d(x_i,y_{j_i}))^2$. 由于各$[x_i\subset C_j]$相互独立,故有 $p(y_{j_1},y_{j_2},\ldots,y_{j_N})=\prod\limits_{i=1}^{N}p[x_i\in C_{j_i}]=\prod\limits_{i=1}^{N}p_{ij_i}$,因而

$$\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})D(y_{j_1},y_{j_2},\cdots,y_{j_N})=\sum_{i=1}^{N}\sum_{j_i} a_{ii}p_{ij_i}(d(x_i,y_{j_i}))^2.$$

若进一步有 $a_{ii}=1(\forall\ i)$,则此时式(3)的聚类模型即为文献[8]中提到的情形.

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

$$\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})D(y_{j_1},y_{j_2},\cdots,y_{j_N})=\sum_{i=1}^{N}\sum_{x_i\in C_{j_i}}(d(x_i,y_{j_i}))^2=\sum_{j}\sum_{x_i\in C_j}(d(x_i,y_j))^2.$$

式(3)的模型简化为式(1)的模型,即上面讨论的一般聚类模型简化为传统的聚类模型.因此,式(3)的聚类模型具有丰富的内涵,包含了一些交互作用,传统的一些模型可作为该模型的特例.

对于模型中的 $p(y_{j_1},y_{j_2},\cdots,y_{j_N})$,由于我们没有先验知识,故不能确定其具体形式,下面利用统计物理的极大熵原理来确定模型中的概率分布 $p(y_{j_1},y_{j_2},\cdots,y_{j_N})$.

对于固定的 $y_1,y_2,\cdots,y_M$,定义能量函数和熵函数分别为

$$E=\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})D(y_{j_1},y_{j_2},\cdots,y_{j_N}),$$

$$H=-\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})\ln p(y_{j_1},y_{j_2},\cdots,y_{j_N}).$$

利用极大熵原理,通过求解变分问题

$$\begin{aligned}&\max\ H=-\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})\ln p(y_{j_1},y_{j_2},\cdots,y_{j_N}),\\ &\text{s.t.}\sum_{j_1,j_2,\cdots,j_N} p(y_{j_1},y_{j_2},\cdots,y_{j_N})D(y_{j_1},y_{j_2},\cdots,y_{j_N})=E,\end{aligned}\tag{4}$$

可得到

$$p(y_{j_1},y_{j_2},\cdots,y_{j_N})=p(x_1\in C_{j_1},x_2\in C_{j_2},\cdots,x_N\in C_{j_N})=\frac{e^{-\beta D(y_{j_1},y_{j_2},\cdots,y_{j_N})}}{\sum_{j'_1,j'_2,\cdots,j'_N} e^{-\beta D(y_{j'_1},y_{j'_2},\cdots,y_{j'_N})}},\tag{5}$$

其中 β 由式(4)确定,是一个与温度成正比的参数.

3　自由能函数的确定

对于任意给定的 $Y=\{y_1,y_2,\cdots,y_M\}$ 及 $J=\{j_1,j_2,\cdots,j_N\}$,并由式(5)得到 $p(y_{j_1},y_{j_2},\cdots,y_{j_N})$,我们按照一定的策略可得到聚类问题的一个可行解.称 (Y,J) 为聚类问题的一个实例.引入温度 T,将聚类问题看做一个物理系统,系统取实例 (Y,J) 的概率为 $P(Y,J)$.

定义物理系统的能量函数和熵函数分别为

$$E=\sum_{(Y,J)} P(Y,J)D(y_{j_1},y_{j_2},\cdots,y_{j_N}),\tag{6}$$

$$H=-\sum_{(Y,J)} P(Y,J)\ln P(Y,J).\tag{7}$$

对于每一固定温度 T,利用极大熵原理,使熵 H 取极大的概率分布 $P(Y,J)$ 称为物理系统的平衡态,而使 $P(Y,J)$ 取极大的实例 (Y,J) 称为聚类系统的最可几实例.在式(6)的条件约束条件下,利用极大熵原理可得

$$P(Y,J)=\frac{e^{-\beta D(y_{j_1},y_{j_2},\cdots,y_{j_N})}}{\sum_{(Y',J')} e^{-\beta D(y'_{j_1},y'_{j_2},\cdots,y'_{j_N})}},\tag{8}$$

其中 β 由式(6)确定.事实上,$\beta\propto\frac{1}{T}$.由式(8)确定的 $P(Y,J)$ 为温度 T 时系统的平衡态.

对于聚类问题所对应的物理系统,其平衡态满足式(8).考虑其边缘分布

$$P(Y)=\sum_{J} P(Y,J)=\frac{\sum_{J} e^{-\beta D(y_{j_1},y_{j_2},\cdots,y_{j_N})}}{\sum_{(Y',J')} e^{-\beta D(y'_{j_1},y'_{j_2},\cdots,y'_{j_N})}},$$

令 $$Z(y_1,y_2,\ldots,y_M)=\sum_{J}e^{-\beta D(y_{j_1},y_{j_2},\ldots,y_{j_N})},$$

则有 $$P(Y)=\frac{Z(y_1,y_2,\ldots,y_M)}{\sum_{J'}Z(y'_1,y'_2,\ldots,y'_M)}.$$

令 $$F(y_1,y_2,\ldots,y_M,\beta)=-\frac{1}{\beta}\ln Z(y_1,y_2,\ldots,y_M),$$

从而 $$F(y_1,y_2,\ldots,y_M,\beta)=-\frac{1}{\beta}\ln\sum_{j_1,j_2,\ldots,j_N}e^{-\beta D(y_{j_1},y_{j_2},\ldots,y_{j_N})}. \tag{9}$$

由式(9)确定的函数称为聚类问题对应物理系统的自由能函数,进一步有

$$P(Y)=\frac{e^{-\beta F(y_1,y_2,\ldots,y_M)}}{\sum_{y'_1,y'_2,\ldots,y'_M}e^{-\beta F(y'_1,y'_2,\ldots,y'_M)}}. \tag{10}$$

由式(10)可知,使 $P(Y)$ 极大的 $y_1,y_2,\ldots,y_M$ 应使 $F(y_1,y_2,\ldots,y_M,\beta)$ 取极小.使概率分布极大的 $y_1,y_2,\ldots,y_M$ 称为聚类系统在温度 T 时的最可几结构(其中 $\beta\propto\frac{1}{T}$).因此,我们可以通过求解自由能函数 $F(y_1,y_2,\ldots,y_M,\beta)$ 的极小来求得聚类系统的解.

4 聚类算法

对于由式(9)定义的聚类系统的自由能函数 $F(y_1,y_2,\ldots,y_M,\beta)$,可推得

$$F(y_1,y_2,\ldots,y_M,0)=\lim_{\beta\to 0}F(y_1,y_2,\ldots,y_M,\beta)=\frac{\sum_{j_1,j_2,\ldots,j_N}D(y_{j_1},y_{j_2},\ldots,y_{j_N})}{M^N}.$$

若进一步假定 $D(z_1,z_2,\ldots,z_N)$ 关于 $z_1,z_2,\ldots,z_N$ 为凸函数,则 $F(y_1,y_2,\ldots,y_M,0)$ 为关于 $y_1,y_2,\ldots,y_M$ 的凸函数,由传统的优化方法极易求出自由能函数的极小.进一步有 $F(y_1,y_2,\ldots,y_M,\infty)=\lim_{\beta\to-\infty}F(y_1,y_2,\ldots,y_M,\beta)=\lim_{\beta\to+\infty}\sum_{j_1,j_2,\ldots,j_N}p(y_{j_1},y_{j_2},\ldots,y_{j_N})D(y_{j_1},y_{j_2},\ldots,y_{j_N})$,故

$$F(y_1,y_2,\ldots,y_M,\beta)=\begin{cases}\dfrac{\sum_{j_1,j_2,\ldots,j_N}D(y_{j_1},y_{j_2},\ldots,y_{j_N})}{M^N}, & \beta=0\\ -\dfrac{1}{\beta}\ln\sum_{j_1,j_2,\ldots,j_N}e^{-\beta D(y_{j_1},y_{j_2},\ldots,y_{j_N})}, & 0<\beta<+\infty\\ \lim\limits_{\beta\to+\infty}\sum_{j_1,j_2,\ldots,j_N}p(y_{j_1},y_{j_2},\ldots,y_{j_N})D(y_{j_1},y_{j_2},\ldots,y_{j_N}), & \beta=+\infty\end{cases}. \tag{11}$$

因此,自由能函数满足确定性退火技术的要求,可通过求解一系列随温度 T(对应于 β)变化的自由能函数的极小点来求解聚类问题.

聚类问题的求解算法如下:

(1) 取 $\beta_0=0,k=0$,求解问题 $\min F(y_1,y_2,\ldots,y_M,\beta_k)$,记最优解为 $(y_1^{(k)},y_2^{(k)},\ldots,y_M^{(k)})$;

(2) $\beta_{(k+1)}=u(\beta_k)$(u 为单增函数),以 $(y_1^{(k)},y_2^{(k)},\ldots,y_M^{(k)})$ 为初始解求解 $\min F(y_1,y_2,\ldots,y_M,\beta_{k+1})$,记最优解为 $(y_1^{(k+1)},y_2^{(k+1)},\ldots,y_M^{(k+1)})$;

(3) 判断收敛准则是否满足,若满足,则 $(y_1^{(k+1)},y_2^{(k+1)},\ldots,y_M^{(k+1)})$ 为最优聚类中心;转(5);否则,转(4);

(4) $k=k+1$,转(2);

(5) 依据最优中心 $(y_1^{(k+1)},y_2^{(k+1)},\ldots,y_M^{(k+1)})$ 及 β_{k+1},按式(8)求出概率分布 $p(y_{j_1},y_{j_2},\ldots,y_{j_N})$,并据此将 x 聚类,输出聚类结果,结束.

有几点需作说明:

(1) 算法中,收敛准则依据所考虑的问题而定,一般若类心稳定时,算法终止;

(2) 计算类数 M 的选取,一般只要使 M 足够大(至少大于真实类的数目)即可.M 越大,计算结果越好,但计算量越大;

(3) 在所得的最优中心$(y_1^{(k+1)},y_2^{(k+1)},\ldots,y_M^{(k+1)})$中,不同的$y_i^{(k+1)}$可能代表同一类,将根据不同的聚类问题及算法表现出来的特征来确定真实聚类数目以及同一类所对应的不同代表元;

(4)算法的第5步仅给出$\{x_1\in C_{j_1},x_2\in C_{j_2},\ldots,x_n\in C_{j_n}\}$的概率,若要进一步分清每一个$x_i$究竟属于哪一类,将作进一步分析.

5 算例分析与结论

对于各$[x_i\in C_j]$相互独立且取$d(x,y_j)=\|x-y_j\|=\left(\sum_{k=1}^{n}(x(k)-y_j(k))^2\right)^{\frac{1}{2}}$的情形,我们进行了大量的模拟实验.在二维空间中,随机生成不同形状、不同密度等各种各样的数据集,用本文的聚类算法进行了大量计算,得到了较好的结果.对于不相等的聚类总体、线性不可分和有桥的聚类问题,当取计算类数M足够大且同一类允许有不同的代表元时,可得到正确的结果.特别是对随机生成的多个服从正态分布的自然类的聚类问题,聚类结果令人满意,即使按传统的聚类方法以各正态分布的中心为初始点,计算结果的函数值也比利用本文提出的方法所得到的最终函数值要大.与文献[6]的工作相比,本聚类方法实用性很强,特别适用于大规模的数据处理.

本文对在模式识别等智能领域中有广泛应用的聚类问题,提出了一类一般聚类模型.该模型考虑了模式与模式、模式与类、类与类之间的交互作用.一些传统的聚类模型可作为本文的特例.文中引入温度参数,将聚类问题看成一个物理系统,通过模拟物理系统的平衡态,来得到聚类问题的解.使用取大的计算类数这一策略,可处理传统聚类算法无能为力的一些聚类问题.

参考文献

1 Lloyd S P. Least squares quantization in PCM. IEEE Transactions on Information Theory, 1982,28(1):129~137

2 Lind Y, Buzo A, Gray R M. Algorithm for vector quantization. IEEE Transactions on Communication, 1980,28(1):84~95

3 Ball G, Hall D. A clustering technique for summarizing multivariate data. Behavioral Science, 1967,12:153~155.

4 Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Penum, 1981

5 Gath I, Geva A B. Unsupervised optimal fuzzy clustering. IEEE Transactions on Pattern and Machine Intelligent, 1989, 11(7):773~781

6 Wong Yui-fai. Clustering data by melting. Neural Computation, 1993,5(1):89~104

7 Fox G C. Physical computation. Concurrency: Practice and Experience, 1991,3(6):627~653

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

9 杨广文,李晓明,王义和.确定性退火技术.计算机学报,1998,21(8):765~768
(Yang Guang-wen, Li Xiao-ming, Wang Yi-he. Deterministic annealing. Chinese Journal of Computers, 1998,21(8): 765~768)

Research of a Clustering Model and Algorithm by Use of Deterministic Annealing

YANG Guang-wen[1]　WANG Ding-xing[1]　ZHENG Wei-min[1]　LI Xiao-ming[2]

[1](*Department of Computer Science and Technology　Tsinghua University　Beijing*　100084)

[2](*Department of Computer Science　Beijing University　Beijing*　100871)

Abstract　Aiming at the defects of traditional clustering model, a kind of clustering model and algorithm are put forward and researched by use of deterministic annealing. The model takes account of the interactions of clusters, some models which were put forward previously are special cases of this one. Temperature parameter is introduced, and the clustering problem as a physical system is considered. Finding the optimal solution to clustering problem is transformed into simulating the equilibrium state of a physical system. The equilibrium state is simulated by solving a series of problems to minimize the free energy which varies with temperature, and finally, the ground state of the system is attained. That is the optimal solution of clustering problem.

Key words　Deterministic annealing, clustering, free energy, the principle of maximum entropy.