文章编号:1001-7402(2017)06-0082-05

因素空间背景基的信息压缩算法*

吕金辉[1],刘海涛[1],郭芳芳[2],郭嗣琮[1],汪培庄[1]

(1. 辽宁工程技术大学 智能工程与数学研究院,辽宁 阜新 123000;
2. 江西科技学院 财经学院,江西 南昌 330098)

摘 要:用因素空间处理大数据的核心思想,是把数据库设置成为背景关系 R 的样本 S,数据的培植目标是要使样本的闭包 $[S]$ 逼近背景关系 R,这样就能对数据进行概念生成,规则归纳和逻辑推理等认知加工。要实现这一思想,最关键的问题是如何进行有效的信息压缩。汪培庄教授提出了背景基的概念并提出了钝角删除策略,十分简捷,但却是一种近似算法。本文加了一点不太强的条件,把这一策略上升成为一种精确算法,保证了因素空间对大数据处理的有效性。

关键词:因素空间; 背景关系;背景基

中图分类号:C934 **文献标识码**:A

1 引言

因素空间是事物描述和认知操作的普适性框架[1-3]. 因素空间处理大数据的核心思想[4-6]强调背景关系 R 是概念生成,规则归纳和逻辑推理的基础. 因素数据库的主要任务是要培植数据:每张因素表所给出的就是背景关系 R 的一组样本点集 S,数据培植的目标就是要实现样本 S 对母体 R 的逼近,一旦获得 R,就获得了相关知识单元的全部知识 $[S]=[S*]$. 然而,背景关系所包含的数据量太大,要想实现这一目标,就必须进行有效的信息压缩. 给定论域 U 上一组因素 $F=\{f_1,\cdots,f_n\}$,背景关系 $R=F(U)$就是 U 在 $X=X(f_1)\times\cdots\times X(f_n))$中的相. 假定 $X(f_j)$都是定量的相空间,或者虽是定性但却能用数来表示,则 R 可以一般地看作是 X 中的凸集. 利用凸集对非负系数加权求和的封闭性,S 的凸闭包$[S]$也全在 R 之中,它是对 R 的进一步逼近,这里,

$$[S]=\{\lambda_1x_1+\cdots+\lambda_kx_k \mid k>0;\ \lambda_1,\cdots,\lambda_k\geq 0;\ x_1,\cdots,x_k\in S\}. \tag{2.1}$$

S 比$[S]$所含数据个数少很多,还能不能再少呢?这就是信息压缩的问题. S 的子集 $S*$ 叫做 S 或 R 的一个基点集,如果它能生成与 S 相同的凸闭包:$[S]=[S*]$。这里的基点其实就是顶点. 最小的基点集就叫做 R 的背景基[3]. 通过背景基,可以将 $[S]$向 R 的逼近过程简化为 $[S*]$向 R 的逼近过程,这是实现数据培植理想的关键,是重中之重.

究竟怎样寻找 S 的基点呢?它可以归结为这样一个基本数学问题:在 n 维空间中,给定 m 个点 P_1,…, P_m 和一点 Q,试问点 Q 能否由 $P_1,\cdots,P_m$ 所生成?亦即,Q 是否在 $P_1,\cdots,P_m$ 诸点所生成成的凸

* 收稿日期:2015-12-01;修订日期:2017-05-01
基金项目:国家自然科学基金(No. 61350003),辽宁省教育厅科学技术研究一般基金资助项目(L2014133)
作者简介:吕金辉(1984-),男,博士研究生,研究领域:因素空间、模糊决策;刘海涛(1982-),男,辽宁阜新人,博士研究生,讲师,主要从事因素空间理论方面的研究;郭芳芳(1986-),女,硕士研究生,讲师,研究领域:区域经济发展;郭嗣琮(1951-),男,吉林白城人,辽宁工程技术大学教授,主要从事模糊分析学,模糊预测;汪培庄(1936-),男,教授,博士生导师,研究领域:模糊数学、智能科学与知识的数学表示理论。

闭包之中？亦即,是否存在着一组非负实数 $\lambda_1,\cdots,\lambda_m$ 使有 $\lambda_1 P_1+\cdots+\lambda_m P_m=Q$ ？如果 $m=n$,而且它们的秩 $r(P_1,\cdots,P_n)=n$,此时的回答很简单. 由等式 $\lambda_1 P_1+\cdots+\lambda_m P_m=Q$ 写出一个方程组,可解得唯一确定的解$(\lambda_1,\cdots,\lambda_m)$, 问题得到肯定回答当且仅当此解非负。但是,当秩 $r(P_1,\cdots,P_m)<n$ 时,方程有无穷多组解,要从无穷多组解中判断是否存在着一组非负解,随着 n 的增大,这是一个 N-hard 问题。同样,当 $m>n$ 时,判断是否存在着一组非负解,随着 n 和 m 的增大,同样具有 N-hard 的复杂性,这是一个老大难问题。

汪培庄教授提出了一个近似的简捷的算法[4], 用一个具体的例子来说明:

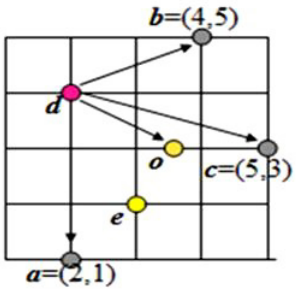

图 1　基点判别准则

Fig1　Basic point criterion

例 1　在图 1 中, S 包含 $a=(2,1)$、$b=(4,5)$、$c=(5,3)$三点,试问 $d=(2,4)$能由 S 生成吗？$e=(3,2)$呢？

解　先把 S 的中心 o 找出来: $o=(a+b+c)/3=(11/3,3)$,从 d 点出发向中心 o 连一射线,若此射线与从 d 点向 S 的其它点所连射线均交成锐角,则可近似地认为 d 不能由 S 生成,作为新的背景样本点, d 不能随意删除;否则,便可近似地认为 d 能由 S 生成,可以将它删除. 对任意两个向量 $x=(x_1,\cdots,x_n)$, $y=(y_1,\cdots,y_n)$,以$(x,y)=x_1y_1+\cdots+x_ny_n$ 表示它们的内积. 有

$$(o-d,a-d)=(5/3,-1)(0,-3)=3>0;$$

$$(o-d,b-d)=(5/3,-1)(2,1)=7/3>0;$$

$$(o-d,c-d)=(5/3,-1)(3,-1)=6>0。$$

都是正的,交成锐角, d 被认为不能由 S 生成,不得删除.

$$(o-e,a-e)=(11/3,1)(-1,-1)=-14/3<0;$$

一旦出现负数就交成钝角, e 被认为可由 S 生成,它可以从表中删除.

判断一点是否被删除的计算复杂度是 $O(nk)$ (n 是因素个数, k 是基点个数),这是近似算法但却很快捷。当相空间是以整数值为代表的托架空间时,近似的程度很高。这样一种近似算法,称为夹角判定法.

本文的目的,是要把这种近似算法上升为一种精确算法,只须增加一个不太繁杂的条件。第 2 节提出精确的夹角判定准则,第 3 节介绍背景基的提取算法,第四节是简短的结论.

2　夹角判定准则

如图 2 所示,每一个黑点代表背景关系 R 的一个样本点,形成样本集 S. S 由它的基点集 $S*=\{A,B,C,D\}$所生成. 四点所围成的白色四边形就是 $S*$ 的凸闭包$[S*]$. O 是 $S*$ 的中心。

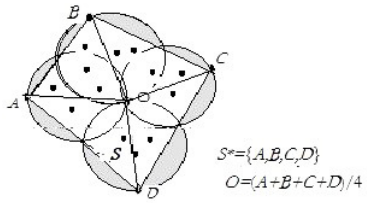

Fig 2 Circle union of S*

图 2 用诸球之并来逼近 R

图中四个圆的直径分别是线段 OA，OB，OC 和 OD．对于平面上任意一点 P，我们要考察由 P 向中心所连射线与它向某一基点（例如 A）所连射线之间的夹角．根据几何学中直径所对的圆周角是直角的定理，若夹角是直角，则 P 点必在以 OA 为直径的圆周上；若所夹的是锐角，P 点必在该圆的内部；若所夹的是钝角，P 点必在该圆的外部．对以 PA 为直径的圆如此，对以 PB，PC，PD 为直径的圆也如此。所以，夹角判定法的实质就是以这四个圆的并集逼近 R 。这一逼近的思想不限于二维平面，在多维线性空间里，以一个线段为直径，便可得到球心和半径，从而决定一个球，夹角判别法的几何意义是用诸球之并来逼近背景关系 R．

诸球之并必覆盖 R，何以见得？只要考察球心 O 在二基点所连直线上的垂足，就能明白．如图 3 所示，$O*$ 是中心 O 在直线 PQ 上的垂足。$O*$ 可以落在 P，Q 两点之间，也可以落在 PQ 线段的两侧．图 3 分别画了两种情况．无论是哪一种情况，以 OP 和 OQ 为直径所画出的两个圆的圆周都要经过中心 O 。根据直径所对圆周角是直角的定理，二圆除了相交于中心以外，还相交于垂足 $O*$ ．而 $O*$ 又在直线 PQ 上，于是，$O*$ 便是两圆周与直线 PQ 的三联点．图 2 中 OA，OB 所对应的二圆交点没有画在直线 BC 上就应该被判为绘图中所出现的一个错误．

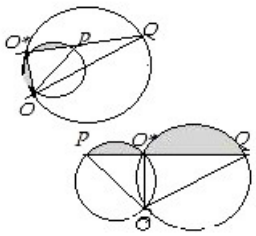

Fig 3 Padel point O*

图 3 垂足 $O*$

命题 2.1 在 $X=X(f_1)\times\cdots\times X(f_n)$ 中，任意给定不共线的三点 O，P 和 Q．O 在直线 PQ 上的垂足 $O*$ 必是以 OP 和 OQ 为直径所画出的两个圆的交点．

证明 不共线三点决定一 2 维平面。命题 2.1 可由平面几何的知识推得。□

二圆交点既在 PQ 连线之上，就不能跑到 PQ 连线之内（与 O 同侧）。这就保证了诸球之并必复盖 $S*$ 的闭包 $[S*]$．而这也就证明了下一个命题：

命题 2.2 在 $X=X(f_1)\times\cdots\times X(f_n)$ 中，给定样本点集 S 及其背景基 $S*$，对任意 $P\in X\setminus S*$，若点 P 由 $S*$ 中点所生成，则必存在一个基点 $Q\in S*$ 使有 $(Q-P,O-P)<0$．

命题 2.2 给出了点 P 是由 $S*$ 生成的一个必要条件，为了进一步导出点 P 是由 $S*$ 所生成的充分且必要的条件。需要考虑诸圆的并与 $[S*]$ 的差异。二者相差的是图中带阴影的区域。下面考虑阴影部分的数学描述，

首先计算 $O*$ 的坐标。因 $O*$ 在直线 QQ' 上,故 $O*=Q+t*(Q-Q)$,因 $OO*$ 垂直于 QQ',故有 $(Q+t*(P'-Q))-O, O-Q)=0$,故有 $(t*(Q'-Q), Q'-Q)=(O-Q, Q'-Q)$,从而得到计算公式

$$O*=Q+t*(Q'-Q),\ t*=(O-Q, Q'-Q)/(Q'-Q, Q'-Q). \tag{2.2}$$

设 Q' 是 $S*$ 中不同于 Q 的另外一个基点,直线 QQ' 连接两个顶点,必在一个边界面上. 无论空间维数如何,点 P 不在直线 QQ' 所在的边界面所界定的阴影之内当且仅当向量 OP 在 $OO*$ 方向的投影不超过边界线 QQ'. 亦即向量 OP 在 $OO*$ 方向的投影不超过线段 $OO*$ 的长度,而这当且仅当 $(P-O, O*-O)\leq(O*-O, O*-O)$. 这就证明了下一命题:

命题 2.3　在 $X=X(f_1)\times\cdots\times X(f_n)$ 中,给定背景关系 R 的一组样本点 S,对任意 $P\in S$,若 P 点由 S 中的点所生成,则必存在一个基点 Q,使对任意 $Q'\in S*\setminus\{Q\}$ 都有 $(P-O, O*-O)\leq(O*-O, O*-O)$,这里 $O*$ 是 O 在直线 QQ' 上的垂足.

夹角判别定理(背景基的信息压缩)　设 S 是一个样本集。$S*$ 是 S 的一个基点集,其中心为 O。$P\in X\setminus S*$ 是一个新的样本点. 它可被删除当且仅当存在一点 $Q\in S*$,使射线 PQ 与射线 PO 形成钝角,亦即,$(Q-P, O-P)<0$. 而且,对于任意 $Q'\in S*\setminus\{Q\}$,都有 $(P-O, O*-O)\leq(O*-O, O*-O)$,这里,$O*$ 是 O 在直线 QQ' 上的垂足.

证明　命题 2.2 保证了点 P 在诸球的覆盖域中. 命题 2.3 又保证 P 不在诸球的阴影域中,故 P 在 $[S*]$ 中,得证.

将判别定理应用到例 1,进一步肯定点 d 可被删除而点 e 是新的基点. 新旧算法一致. 但为了说明可能的不一致,我们在例 1 中增加一个点 $P=(2, 2)$,因 $(a-P, o-P)=((0, -1), (5/3, 2))=-1<0$,按原有的近似判别法则,点 P 应被删除. 但按现在的判别定理,因点 P 在以 oa 为直径的球内,取 $Q=a$, $Q'=b$,中心 O 在直线 QQ' 上的垂足参数是 $t*=(O-Q, Q'-Q)/(Q'-Q, Q'-Q)=34/60$, $O*=(-8/15, 4/15)$,算得 $(P-O, O*-O)=28/45$, $(O*-O, O*-O)=80/225$,结果是 $(P-O, O*-O)>(O*-O, O*-O)$ 按照夹角判别定理,点 P 应添加到 $S*$ 中去. 经过资格重审,新的背景基改为 $S*=\{a, b, c, P\}$.

3　背景基 $S*$ 的提取算法

给定 $S=\{x_i=(x_{i1}, \cdots, x_{in}) \mid i=1, \cdots, m\}$; $S*:=\varnothing$;

步骤 1　就每个分量 j 找出 S 的上下两个极点

$$P_j^+=argimax_i\{x_{ij} \mid x_{ij}\in S\};\ P_j^-=argimin_i\{x_{ij} \mid x_{ij}\in S\}\quad (j=1, \cdots, n); \tag{3.1}$$

如果上(下)极点有很多,上下各挑选一个,这样共有至多 $2n$ 个极点,将它们输入 $S*$,作为启动的基点。

步骤 2　计算 $S*$ 的中心 O;对每一 $P\in S\setminus S*$ 判断 P 是否被诸球所覆盖,亦即:

对每一 $Q\in S*$,判断 $(Q-P, O-P)<0$?若不是,则换 Q;若是,则转入步骤 3

步骤 3　(计算量:$4m'n$)对每一 $Q'\in S*\setminus\{Q\}$,找 O 在直线 QQ' 上的垂足 $O*$;

判断 $(Q'-Q, O*-O)<(O*-O, O*-O)$?若是,则删除 Q;若不是,则将 Q 输入 $S*$;

步骤 4　每向 $S*$ 引入一个新成员,都要对旧成员的资格进行再审查:先要更新 $S*$ 的中心 O,

再对 $S*$ 的每一旧成员 Q,判断 $(P-Q, O-Q)<0$?若不是,则换 Q;若是,则对每一

$Q'\in S*\setminus\{Q, P\}$,找 O 在直线 QQ' 上的垂足 $O*$;判断 $(O*-O, Q'-Q)<(O*-O, O*-O)$?若是,则删除 Q;若不是,则换 Q;

如此重复步骤 2,3,直到 S 中的所有点都被核查。

以上算法可以用于新样本集的开始,也可用于旧样本集的吐故纳新。

步骤 1 的计算量不多于 $4mn$,4 着眼于+,×,'比较'或其它运算。步骤 2 的计算量不多于 $4(m-m')$

$m'n$，m' 是 $S*$ 的点数；步骤 3 的计算量不多于 $4m'n$，步骤 4 的计算量不多于 $3m'(m'-1)n$。假定所有 S 的点都要经历步骤 3 和步骤 4，则总计算量不多于 $3mn+3m'(m'-1)n+m\times[4m'n]+m\times[3m'(m'-1)n]$. 最后，总计算量不多于 $4m(m')^2n$. 这里 m' 是变动的，以其最大值来估计. 面对大数据，真正有威胁的参数是 m 和 n，整个背景基的提取复杂性不超过 $O(mn)$（但有界量含 $(m')^2$），这是很理想的信息压缩算法. 背景基的在线压缩更加快捷.

4 结论

本文把钝角删除法上升为精确算法，这就为大数据的因素空间处理提供了实现的可能性. 面对着海量数据，背景基始终保持在非海量的级别，在网上吐故纳新，开展人机认知的互动.

参考文献：

[1] 汪培庄，Sugeno M. 因素场与模糊集的背景结构[J]. 模糊数学与系统，1982，(2)：45～54.
[2] 汪培庄，李洪兴. 知识表述的数学理论[M]. 天津：天津科学技术出版社，1994.
[3] 汪培庄. 因素空间与数据科学[J]. 辽宁工程技术大学学报：自然科学版，2015，34(2)：273～280.
[4] 汪培庄，因素空间与因素库简介(特约报告)，智能科学与数学论坛，2014 年 5 月葫芦岛.
[5] 刘海涛，刘增良，何华灿，何平. 因素空间发展评述，模糊系统与数学，(已录用).
[6] 曾繁慧，郑莉. 因素分析法的样本培育[J]. 辽宁工程技术大学学报：自然科学版 (已录用).

The algorithm of extraction of background bases

LV Jin-hui[1]，LIU Hai-tao[1]，GUO Fang-fang[2]，GUO Si-zong[1]，WANG Pei-zhuang[1]
(1. Institute of Intelligence Engineering and Mathematics，Liaoning Technical University，Fuxin 123000，China；
2. Institute of Finance and Economics，Jiangxi University Of Technology，Nanchang 330098，China)

Abstract：The core idea of big data processing based on factor space is placing data as sampling S of background relation R，and taking the target of data-cultivation as the approximation of R by the enclosure $[S]$. And then the thinking functions of concept generation，rules extraction and logical inference can be done on data. To realize such strategy，the key point is information compression. For this reason，Prof. P Z Wang has presented the concept of background bases and an algorithm named angle criterion，which is simple and fast，but is approximation. This paper will promote it to be an accurate algorithm by adding a rather weaker condition. The algorithm enables that factor space theory can handle big data effectively.

Key words：factor space；background relation；background bases