动态挖掘进程中参数演化与矛盾域分布规律的研究*)

杨炳儒　张　帆　韩彦岭
(北京科技大学信息工程学院　北京 100083)

摘　要　基于内在认知机理的知识发现理论研究基础，宏观地描述了其核心内容之一——信息扩张机制的内涵及主要研究内容，给出了动态挖掘进程规律的部分成果，详细阐述了参数演化规律及矛盾域的分布规律，揭示了动态(在线)挖掘进程中潜在的本质、规律与复杂性，为进一步解决海量数据、动态数据给挖掘进程带来的本质的、极富挑战性的难题奠定了较为坚实的理论基础。

关键词　知识发现，信息扩张机制，参数演化，矛盾域

Research Overview of Information Increasing Mechanism in Inner Cognitive Mechanism of Knowledge Discovery

YANG Bing-Ru　ZHANG Fan　HAN Yan-Ling
(School of Information Engineering, Bijing University of Science and Technology, Beijing 100083)

Abstract　On the basis of research on knowledge discovery theory based on inner cognition mechanism, this paper described one of the kernel contents, which was the connotation and main content of information increasing mechanism, and gave the part production of dynamic mining process regulation. The regulations of parameter evolvement and contradiction field distribution were exhausted in detail, the potential essence, regulation and complexity were revealed in the process of dynamic mining. It established the solid theory basis for solving the essential and challenging problem in the process of sponge data mining and dynamic data mining.

Keywords　Knowledge discovery, Information increasing mechanism, Parameter evolvement, Contradiction domain, Information entropy

1　引论

在科学发展的漫长历史进程中，学科发展的一般规律常常表现为：在其发展初期往往结合应用背景，从具体的技术方法入手，得到个体的演绎结果；但当它经历了一段发展的路程后，就迫切需要理论的概括与指导，另辟路径，解决发展中困难问题。作为处在国际学术前沿、多学科交叉的新边缘学科——知识发现(Knowledge Discovery)[1]历经 10 余年的发展，走到今天，同样面临着呼唤理论概括来解决主流发展中若干棘手问题的局面。本文从知识发现、认知科学与智能系统等多学科交叉的角度，以认知自主性为核心概念，将知识发现视为一个开放的和不断进化的认知系统，研究它的系统结构、方法、进化与运行机制[2]。于 1997 年创造性地提出了内在认知机理研究的新方向，进而于 2002 年完成了基于内在认知机理的知识发现理论——KDTICM[3]。

知识发现内在认知机理是 KDTICM 中最核心的内容，涵盖 3 个机制：双库协同机制(第一机制)、双机融合机制[4](第二机制)和信息扩张机制(第三机制)，它们是 KDTICM 的 3 个理论支柱，是构成整个理论体系的"生长源"。通过对内在认知机理的研究，可揭示其作为认知系统的知识发现系统潜在本质、规律与复杂性[5]。

长期以来，在知识发现的主流发展中，挖掘算法与评价方法的讨论基本上是在一个时间剖面上、相对稳定的状态下进行的，而对于动态挖掘进程、实时与在线的挖掘进程考虑得很少；动态(在线)挖掘[6]是具有相当大难度的热点问题。我们首次从认知物理学等新理念出发，发现了知识发现系统内在认知机理涵盖的信息扩张机制(第三机制)，主要研究了动态(在线)数据挖掘[7~9]过程从一个抽象级向下一个抽象级、从固有数据库(知识库)向扩展数据库(知识库)过渡时所呈现的运行规律。在动态挖掘进程中，针对规则参数的演化规律、关联规则取舍方法和可理解性问题、挖掘规则矛盾域的分布和计算、变论域下阈值综合设置等一系列富有挑战性的问题，在一定程度上给出了整体解决方案。

2　动态挖掘进程规律的描述

知识发现过程是数据库、知识库和知识获取与评价方法相互影响、相互制约的一个开放的、动态的并且不断进化的复杂过程，其本质体现为动态挖掘进程。表现在：首先，数据库可能是不断更新和不断扩充的；其次，知识库也会在量和质上不断扩充和更新；再次，随着知识质和量的演变，知识评价的主、客观标准也会发生变化；最后，信息空间的粒度在不断更新。

下面通过非形式化描述来刻画动态挖掘规律的内涵：

(1)假设规则经评价后，通过规则支持度、可信度、充分性因子等参数间的关系和它们的历史变化规律对各类知识库中的规则进行深入分析，决定规则的取舍，提高规则的可理解性。

(2)在挖掘进程逐渐深入的过程中，利用不动集簇判定算法找到数据中的"不动点"，即能保证知识稳定出现的数据簇。当这样的数据簇出现后，对应知识的出现相对而言不会因数

*)国家自然科学基金重点项目(69835001)、教育部科技重点项目(教技司[2000]175)、北京市自然科学基金项目(4022008)资助。**杨炳儒**　教授，博士生导师，主要研究方向为知识发现与智能系统、柔性建模与集成技术；**张　帆**　博士生，主要研究方向为知识发现与智能系统机制；**韩彦岭**　博士后，主要研究方向为知识发现与智能诊断。

据变化而受大的影响。

(3)根据质、量互变原理,数据量增长到一定程度时,引起客观事物、状态或进程的突变。利用突变协调算法,判断突变条件,给出突变临界值,从而解决矛盾知识的处理问题[10]。

(4)由挖掘进程与挖掘算法等表现出的知识发现系统复杂性的研究。

动态挖掘进程规律的描述涵盖了以上4方面的内容,围绕其进行的研究均是在信息量(数据量)扩张的挖掘进程中诱发出的,故亦称之为信息扩张机制。

3 动态挖掘进程规律部分研究结果

3.1 规则参数的演化规律

3.1.1 基本概念

本文中仅称通过数据挖掘[11]算法得到的规则为规则,记作 $A\Rightarrow B$。

在知识发现过程中,伴随每一次挖掘总会有一些参数存在,这些参数是依特定的、公认的定义从数据库中与规则有关的数据上实时计算得到的。它们具有特定的含义,能够从不同的角度刻划规则的特征,称此类参数为规则参数。

定义 1 在数据量不断增加的历史进程中,在给定的时间段内,可以计算得到规则同一参数的若干个值。以时间为序排列它们,得到一个参数值序列,这个序列有且仅有上升、下降、平行和波动4种演化趋势,称此为参数演化。

若描述规则的参数有 N 个,每一种参数具有上述4种演化趋势,则参数演化情况的组合数共有 4^N 种。其中出现参数波动的情况有 4^N-3^N 种,可依据参数分布信息决定规则取舍;在所余 3^N 种组合情况中,根据 N 个参数间的关系去掉现实中不可能出现的组合,降低后续分析的难度。最后,面对不同的规则形式,根据其相随参数的性质,提出决定规则取舍的依据。

3.1.2 评价关联规则所选用的参数

在知识发现主流研究中,认为规则支持度、可信度和相关性是伴随规则存在的不可缺少的参数,其中支持度和可信度是关联规则评价中最常用的两个参数。我们将用充分性因子代替相关性,因为它能区分规则 $A\Rightarrow B$ 和 $B\Rightarrow A$。除这3个参数外,规则前件支持度和后件支持度能够体现前、后件支持度的发展趋势,并且后件支持度是前3个参数不能涵盖的一个参数,所以在对关联规则的分析中本文最终选择5个参数——规则支持度(记为 S_r)、规则可信度(记为 C_r)、规则充分性因子(记为LS)、规则前件支持度(记为 S_A)、规则后件支持度(记为 S_B)。至此,5个参数全面地涵盖了规则的前件、后件、规则自身及前件与后件关系等方面的信息,并能完成规则取舍的操作。

3.1.3 意外规则参数演化定理

评价关联规则所选用的5个参数中,任何一个对于规则都有着不可缺少的重要意义。在考虑规则的历史表现时,上述的每一种参数都应考虑,每一种参数的每一种演化亦都应考虑。但如果5种参数中每一参数都考虑上升、下降、平行保持和无规律波动4种演化可能,那么上述5个参数的全部组合就为 $4^5=1024$ 种。若对其一一考察,无疑是笨拙的和难于完成的。但分析中又不应遗漏任一可能情况。我们首先考虑每个参数上升、平行、下降这3种情况,如此共有 $3^5=243$ 种组合情况。首先给出如下的定义和定理:

定义 2 我们称 $E:P\to\{-1,0,1\}$,对任意 $x\in P$

$$E(x)=\begin{cases}1, & x\text{上升}\\ 0, & x\text{平行}\\ -1, & x\text{下降}\end{cases}$$

为参数演化函数,其定义域为 $P=\{$支持度 S_r、可信度 C_r、前件支持度 S_A、后件支持度 S_B、充分性因子 LS$\}$。

定义 3 我们称实时数据库中,某一时刻由参数演化函数的值组成的五元组 $V=\langle E(S_r),E(C_r),E(S_A),E(S_B),E(\text{LS})\rangle$ 为参数演化的一个组态。由参数演化过程中所出现的全部组态构成的集合称为参数演化的组态空间 S。

定义 4 支持度低于给定阈值(即小于用户给定的最小支持度,大于用户给定的用于挖掘意外规则的最小支持度),而可信度高于指定阈值的关联规则称为意外规则。

这里意外规则本身是一个不确定性概念,其特征表现为:充分性因子>1;支持度低;可信度高。通过设定用户给定的挖掘意外规则的最小支持度值,可用于区分噪声数据。为了使我们的结果具有较为普遍的意义,我们使用语言场和语言值理论[12],来描述以上参数。即把支持度、可信度等看作语言变量,并根据一定的指标,如用户给定的挖掘意外规则最小支持度等,把二者的具体值分成"很小"、"小"、"高"、"很高"等几个语言值。我们发现了如下规律:

定理 1(意外规则参数演化定理) 在KDD的动态挖掘进程中,在对实时数据库 DB 实施分库和每种参数只考虑上升、平行、下降3种演化情况的前提下,对于定义4所论的意外规则而言,其组态空间为

S=$\{\langle 0,0,0,0,0\rangle$, $\langle 0,0,0,1,-1\rangle$,
$\langle 0,0,0,-1,1\rangle$, $\langle -1,0,-1,0,0\rangle$,
$\langle -1,0,-1,1,-1\rangle$, $\langle -1,0,-1,-1,1\rangle$,
$\langle 0,1,-1,0,1\rangle$, $\langle 0,1,-1,-1,1\rangle$,
$\langle 0,1,-1,1,0\rangle$, $\langle 0,1,-1,1,1\rangle$,
$\langle 0,1,-1,1,-1\rangle$, $\langle -1,1,-1,0,1\rangle$,
$\langle -1,1,-1,-1,1\rangle$, $\langle -1,1,-1,1,0\rangle$,
$\langle -1,1,-1,1,1\rangle,\langle -1,1,-1,1,-1\rangle\}$。 (证明见文[13])

其意义在于,将1024种组态情况划归为16种(波动型除外),并且有些组态具有某种共性,即具有相同的规则支持度和可信度,如1—3;4—6;7—11和12—16。所以又可将16种参数演化组态中具有共性的组态聚集在一个主题下,形成4类主题。经分析后,可向用户提供规则背后蕴涵的深层信息。

3.1.4 动态挖掘进程中意外规则的取舍方法和可理解性讨论

基于上述定义与定理,按4类主题及参数波动变化类的完备性表现加以讨论,给出意外规则在动态挖掘进程中取舍方法及其可理解性分析。下面仅以支持度保持平行、可信度保持平行的情形(组态空间S中的第1、2、3个组态)为例加以说明(余者类同)。

分析图1(不失一般性,以典型的商场销售记录为背景):

(Ⅰ)支持度 S_r 保持平行,可信度 C_r 保持平行,说明规则的意外性稳定出现,排除了由于误操作等因素造成的意外情况,那么只需考察规则的 LS 的变化即可决定规则的取舍。而LS的变化取决于 S_B 的变化。

(Ⅱ)支持度 S_r 保持平行,可信度 C_r 保持平行,根据二者定义与意外规则定义,所以 S_A 在较低的数值水平上保持平行。

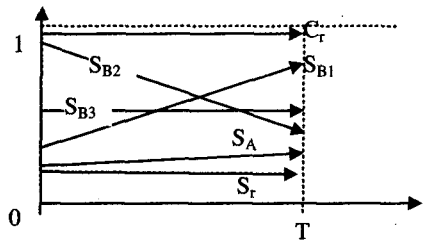

图 1　支持度保持平行、可信度保持平行

(Ⅲ)在时刻 T,$LS>1$,假设临时知识库的规则已经经过"对规则取舍算法"[14]的操作,即 $S_A<S_B$。

(Ⅳ)在 T 时刻前,当 $S_B<S_A$,或 $S_B>C_r$ 时,根据推论,规则舍弃。但如图 1 所示,当满足 S_r 平行,C_r 平行的条件时,上述情况的发生应仅为点状分布。因本文试图通过 5 参数的变化过程分析规则,故个别点的情况可以忽略,则 S_B 的变化为图示的 S_{B1}(上升)、S_{B2}(下降)、S_{B3}(平行)3 种。

(Ⅴ)根据上述分析,可见规则的意外性表现稳定,且 $LS>1$ 并始终大于其对规则,故规则应保留。

可理解性分析:当 S_{B2} 出现时 LS 上升,出现支持规则的特性,即 S_B 销售量的下降,使它的出现更加依赖 A 的出现,可充分信任规则,但需做出减少 B 的进货量的类似决策;如 S_{B1} 出现,则用户需注意,虽然事务 A 决定了事务 B,但 B 的畅销使其对 A 的依赖越来越小。S_{B3} 的出现说明各参数表现平稳,T 时刻计算出的各参数可信(其它各种组态情况分析类同,从略)。

从以上分析可看出,只有图 1 所示规则才是人们心目中真正期待的意外规则。而其它各种情况,其意外性只在 T 时刻和其有限的时间邻域内保持,但终将顺着原演化趋势发展下去,而失却了那些意外规则所具有的特征。所以,只有图 1 所示规则才能长期遵循,而其余情况只能在一个时间区间内相信之。如果图 1 中出现参数大幅的跳动,排除噪声干扰的情况后,笔者认为此时有可能发生了规则突变,这也是我们将进一步研究的内容。

3.1.5　发生参数波动变化时规则取舍的研究

规则的参数波动变化情况有 $4^4+3\times4^3+3^2\times4^2+3^3\times4+3^4=781$ 种,我们引用了认知物理学方法来处理参数波动变化的态势,并创新性地运用了信息扩散原理。

信息扩散原理是一种在样本不足的情况下,对样本应遵循的规律进行认识的模糊数据处理方法。针对自动评价方法[15]可在领域专家不介入的情况下,利用知识(规则)的可计算参数进行评价,并由信息扩散原理弥补参数相对不足的缺陷,得到规则参数的概率分布信息,据此客观地展现规则特征,从而实现规则评价。

信息扩散原理:设 $W=\{w_1,w_2,\cdots,w_m\}$ 是知识样本,V 是基础论域,设 w_j 的观测值 v_j,设 $x=\varphi(v-v_j)$,则 W 非完备时,存在函数 $\mu(x)$,使 v_j 获得的量值为 1 的信息可按 $\mu(x)$ 的量值扩散到 v 去,且扩散所得的原始信息分布

$$Q(v)=\sum_{j=1}^{m}\mu(x)=\sum_{j=1}^{m}\mu(\varphi(v-v_j))$$

能更好地反映 W 所在的总体的规律。

在此,我们设基础论域为 $U=\{u_1,u_2,\cdots,u_n\}$

式中 u_1,u_2,…,u_n 为控制点。令

$$q(u_i)=\overline{f}_m(u_i)=\frac{1}{\sqrt{2\pi}mh}\sum_{j=1}^{m}\exp\left[-\frac{(u_i-v_j)^2}{2h^2}\right]$$

则其物理意义为:m 个观测样本 $v_1,v_2,\cdots,v_m$ 将其所携带的信息扩散给 U 中的一个控制点 $u_i(i=1,2,\cdots,n)$ 的信息量总和。其中 h 为扩散系数,可根据样本集合中样本的最大值 b、最小值 a 和样本个数 m 来确定:

$$h=\begin{cases}1.4230(b-a)/(m-1) & m<10\\ 1.4208(b-a)/(m-1) & m\geqslant10\end{cases}$$

再令

$$Q=\sum_{i=1}^{n}q(u_i)$$

其物理意义是:由观测样本集合$\{y_1,y_2,\cdots,y_m\}$,经信息扩散,将信息扩散给控制点的信息总量。易知 $p(u_i)=q(u_i)/Q$ 就是样本落在 u_i 处的频率值,可以作为概率的估计值。

则 $p(u\geqslant u_i)=\sum_{k=i}^{n}p(u_k)$ 是超越 u_i 的概率值。那么通过 $p(u\geqslant u_i)$ 就可以得到大于用户给定最小参数值的规则参数的概率。

假设在基础论域内取 n 个控制点,通常由用户给定规则的最低指标值,用于规则挖掘。相应参数大于等于这一指标的规则在挖掘过程中被保留,记为 minvalue。minvalue 应该被取为控制点之一。假设 minvalue 是第 i 个控制点,那么 $P_{minvalue}=(n-i+1)/n$ 的物理意义是超载 minvalue 时至少应达到的概率值。当 $p(u\geqslant u_i)\geqslant P_{minvalue}$ 时,规则保留,否则规则舍弃。

如前所述,我们得出了揭示参数演变规律的意外规则参数演化定理,由这个定理导出了一类关联规则取舍和可理解性分析的新方法。由不同的挖掘方法得到的关联规则,大部分都存在支持度偏小而可信度较大的特点,是关联规则中具有广泛性的特征,因此定理中涉及到的具体内容,如参数个数 N,组合情况的分析和规则舍弃依据等,均具有普适性及重要的参照性。

3.2　矛盾域分布的研究

在动态知识发现进程中,时空条件有可能发生变化,数据库和知识库在不断更新和不断扩充,而且随着知识质和量的演变,知识评价的主、客观标准也会发生变化。在这种情况下,往往会挖掘出一些前件相同而后件相反或互斥的"矛盾知识"。下面我们将讨论如何处理动态知识发现进程中的这些矛盾知识,给出时间段内矛盾区间确定规律。

3.2.1　矛盾规则的概念模型

以下将对矛盾规则的概念做严格约定,组成本文研究范畴内矛盾规则的特定概念模型。

定义 4　如果在某一次数据挖掘的操作中(以产生式规则为例),同时出现两个条件(规则前件)相同但结论(规则后件)相反(互斥)的规则,则称这两个规则为矛盾规则,所对应的时间轴上的投影就是矛盾域。

我们在矛盾规则的研究中,首先研究具有支持度(S)和可信度(C)两个参数的情形,然后再把它扩展到一般的 k 维空间中去。依照矛盾规则的概念模型,我们称规则(比如产生式规则 $P\Rightarrow Q$)与其对应的矛盾规则($P\Rightarrow\neg Q$)只有同时或者说在同一次挖掘时共同出现,才能称之为矛盾。这是非常重要的一个条件,因为在不同的时空条件下,这两个对立规则可能是允许出现的。这一点,对于我们在取舍具有对立性质(比如说矛盾规则)的规则时,是极为重要的。也就是说,矛盾规则是相对的,在同一时空下矛盾的规则,在不同的时空下是可能允许共存的,不能轻易舍弃。因此,我们有必要加上时间这一参数,从而建立一个三维空间规则参数不等式(或方程),表

示在一个动态的挖掘过程中，规则的参数随时间变化的趋势。当两对立(矛盾)的规则在阈值空间中同时出现时，所对应的时间轴上的区间(或点集)，就是我们所求的问题的解。

假设在对真实数据库进行动态挖掘时，规则的两个参数(支持度和可信度)的阈值设为 c_0 和 s_0，那么在三维空间里求解规则及其对立(或称矛盾)规则同时出现的问题，就是求取满足下列两参数不等式中对 t 的区间(或点集)的解：

$$①\begin{cases}c(t)-c_0\geqslant 0\\ s(t)-s_0\geqslant 0\end{cases}$$

$$②\begin{cases}c^*(t)-c_0\geqslant 0\\ s^*(t)-s_0\geqslant 0\end{cases}$$

若：①有解，即表示该规则(比如产生式规则 $P\Rightarrow Q$)在 t_i 时被挖掘出来。

②有解，即表示矛盾规则(比如 $P\Rightarrow\neg Q$)在 t_j 时被挖掘出来。

①与②在同一时刻(或称同一次挖掘时)，即 $t_i=t_j$ 有解，则表示出现了矛盾规则。

3.2.2 矛盾域的求解

对于上述联立不等式的求解，我们可以先分别解①、②两个非线性方程组：

假如有解，那么将得到 t 的区间(或点集)，然后再对这些区间进行分析，就可以求得矛盾域了。

设已求得矛盾规则参数方程组的解共 k 个，我们将对应的时间参数 t 的值按大小作一个序，设为：$t_1<t_2<\cdots<t_k$，

然后，作 $k-1$ 个小区间，即：$[t_1, t_2]$，$[t_2, t_3]$，…，$[t_{k-1}, t_k]$

接下来在各点 $t_i(i\in(1,2,\cdots,k))$ 处求偏导数：$\frac{\partial c}{\partial t_i}$，$\frac{\partial s}{\partial t_i}$

则可确定矛盾规则参数空间曲线在各点处的单调性，然后将上述各个小区间中 t_i 处的偏导小于 0 而 t_{i+1} 处的偏导大于 0 的区间$[t_i, t_{i+1}]$全部剔除，得到剩余的各个小区间：

$$[t_1^*, t_2^*],[t_2^*, t_2^*, t_3^*],\cdots,[t_{l-1}^*, t_l^*](l\in i)$$

上式可作为不等式：

$$\begin{cases}c^*(t)-c_0\geqslant 0\\ s^*(t)-s_0\geqslant 0\end{cases}$$

的解中相应的 t 解区间，亦即矛盾规则参数空间曲线过空间直线(c_0, s_0, t)的对应 t 的区间。我们在各个小区间$[t_1^*, t_2^*]$，$[t_2^*, t_3^*]$，…，$[t_{l-1}^*, t_l^*]$里找到满足不等式组

$$\begin{cases}c(t)-c_0\geqslant 0\\ s(t)-s_0\geqslant 0\end{cases}$$

的区间

$$[T_1^*, T_2^*],[T_2^*, T_3^*],\cdots,[T_{p-1}^*, T_p^*](p\in l)$$

就是我们所求的矛盾域。即在上式的时间区间里，该动态的挖掘过程中将产生矛盾规则。

3.3 变论域下阈值综合设置的研究

3.3.1 阈值协调器的设计思想

阈值协调器的核心部分是针对用户感兴趣的规则，在求得该规则及其对应的矛盾规则的拟合曲线后，用户给出初始的阈值；协调器将计算其矛盾域或矛盾区间，如果有解，则自动对初始的阈值进行调节。继续计算矛盾域或矛盾区间，直到无解。然后得到一个阈值的取值区间，使得在这一动态挖掘过程中，用户所感兴趣的这一规则在挖掘过程中将无矛盾规则出现。协调器设计的理论基础是根据威尔斯特拉斯(Weierstrass)定理，即任何连续函数都可以用 n 次多项式做任意精确的逼近，以便求得规则及其对应矛盾规则在动态过程中的空间拟合曲线，这也是阈值协调器的设计思想。

3.3.2 变论域下阈值设置的一般方法

针对动态知识发现进程中的矛盾规则的取舍，阈值协调器输出的是一个区间而不是单一的点值。我们在阈值协调器输出的阈值区间值的基础上，结合点值型的阈值设置方法，给出一般化的点值与区间值相结合的变论域下阈值设置的方法。

在这里，主要是解决变论域下阈值设置的输出函数 $MIS(i)$ 如何确定的问题。

首先，我们已经在时空论域中，通过使用阈值协调器，得到了一个基础的“阈值区间”，可以把它作为一个基本的、客观的阈值取值区间(“阈值参照系”)。设最小支持度的阈值区间 $\aleph_s=[a_s, b_s](0\leqslant a_s\leqslant 1, 0\leqslant b_s\leqslant 1)$

由于数据库中各数据项目的重要度 $\tilde{B}_i$ 不同，为了使包含 $\tilde{B}_i$ 较大的项目的那些规则也能被挖掘出来，势必需要调整它的最小支持度阈值，即按某一个原则降低它的阈值。这样的某原则的确定，就是通过在数据属性论域里讨论的对数据属性的模糊综合评判，然后得到各项目的不同重要度 $\tilde{B}_i$，再计算它的偏离度 $p_i=1-\overline{\tilde{B}_i}$，以此为依据来调整它的最小支持度阈值。于是，我们可以确定一个阈值设置的输出函数 $MIS(i)=f(i)*(1-\overline{\tilde{B}_i})$，(满足 $p_i<\delta$))，其中：$\delta=\frac{a_s}{f(i)}$ 可称之为偏离界；$f(i)$ 表示当前项目 i 在数据库中出现的实际频度。

可以看出，当一个项目的 $\tilde{B}_i$ 越大时，相应的归一化值(即对比尺度)$\overline{\tilde{B}_i}$ 也就越大，它的 p_i 就越小，项目的最小支持度阈值的取值就越小，达到了降低该项目的最小支持度阈值、尽可能让包含重要度较大的项目(也就是偏离度较小者)的规则被挖掘出来的目的。

该阈值设置输出函数 $MIS(i)$ 的建立，是在变论域下综合各因素而确立的一个客观、合理的阈值设置方法。它的意义在于：①如果按此输出函数设置阈值，既不会因为设置过高而使重要的规则挖掘不出来；②也不会因为设置过低而出现矛盾规则和其它重复冗余的规则；③同时也摒弃了传统单一的阈值设置模式；④摒弃了主观经验主义设置阈值的不合理方法；⑤摒弃了单一“点值”的思想，提倡“点值”与“区间值”混合设置的思想。

结论 动态挖掘进程中规律性问题的研究，对知识发现——作为认知系统的潜在本质、规律与复杂性有了进一步深刻的揭示与认识；对知识发现的主流发展起着重要的驱动作用[16]；在一定程度上解决了知识发现主流发展过程中出现的一些极富挑战性的问题；进一步拓展与完善了基于内在认知机理的知识发现理论 KDTICM。应当指出：本文的研究结果，均通过了实际(例)运行实验与对比实验，以验证其有效性与先进性；另外，引伸出一些新的研究生长点。

参 考 文 献

1 Fayyad U. Knowledge Discovery and Data Mining Towards a Unifying Framework. In: KDD'96 Proc 2nd Intl Conf on KD & DM, AAAI Press, 1996

2 Yang Bing-ru, Zhang De-zheng, Tang Jing, et al. Expanded researching on knowledge discovery, Information Fusion, 2000. FUSION 2000. In: Proceedings of the Third International Conference on, July 2000, 2: 10～13

(下转第 230 页)

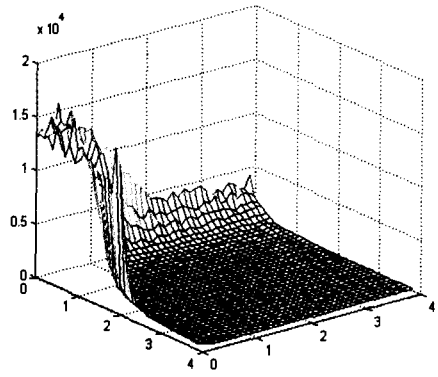

图 3 $w=1$,感知系数 c_1、c_2 与 PSO 性能曲面

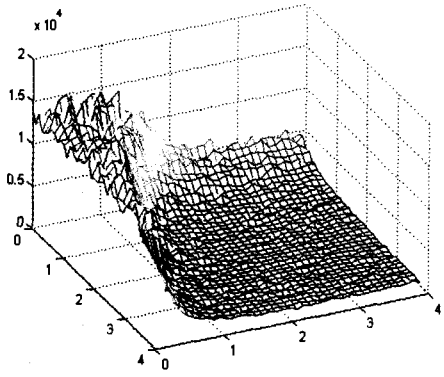

图 4 $w=0.7$,感知系数 c_1、c_2 与 PSO 性能曲面

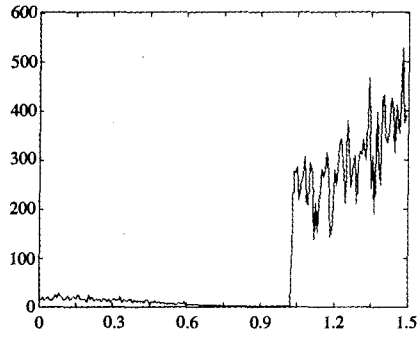

图 5 感知系数 c_1、c_2 一定,w 与 PSO 性能曲线

参 考 文 献

1 Eberhart R C, Kennedy J. A new optimizer using particle swarm theory. In: Proceedings of the sixth International Symposium on Micro Machine and Human Science. Nagoya Japan, 1995. 39～43

2 Kennedy J, Eberhart R C. Particle Swarm Optimization. In: Proc. IEEE International Conference on Neural Networks. IEEE Service Center, Piscataway, NJ, IV, 1995. 1942～1948

3 Clerc M. TRIBES-Aparameter Free Particle Swarm Optimizer. http://clerc.maurice.free.fr/PSO 2002-08-10/2003-10-08

4 Salman A. Discrete Particle Swarm Optimization for Heterogeneous Task Assignment Problem. In: Proceedings of World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001). Orlando, USA, 2001

5 Clerc M. Discrete Particle Swarm Optimization: A Fuzzy Combinatorial Black Box. http://clerc.maurice.free.fr/PSO/Fuzzy_Discrete_PSO/Fuzzy_DPSO.htm. 2000-04-01/2003-10-08

6 Krink T, Vesterstrom J S, Riget J. Particle Swarm Optimization with Spatial Particle Extension. In: Proceeding of the IEEE Congress on Evolutionary Computation, Honolulu, Hawaii, USA, 2002

7 Hirotaka, Yoshida, Kenichi. A particle Swarm Optimization for Reactive Power and Voltage Control Considering Voltage Stability. IEEE International Conference on Intelligent System Applications to Power Systems, Rio de Janeiro, 1999

8 Voss M S, Feng Xin. Arma Model Selection Using Particle Swarm Optimization and Aic Criteria. In: 15th Triennial World Congress, Barcelona Spain, 2002

9 Parsopoulos K E, Vrahatis M N. Particle Swarm Optimization Method in Multiobjective Problems. In: Proc. 2002 ACM Symp. Applied. Computing (SAC 2002), Madrid, Spain, 2002. 603～607

10 Van den Bergh, Engelbrecht A P F. Cooperative Learning in Neural Networks using Particle Swarm Optimizers. South African Computer Journal, 2000, 26: 84～90

11 Carlisle A, Dozier G. Tracking Changing Extrema with Adaptive Particle Swarm Optimizer: [Technical Report CSSE01-08]. Auburn University, 2001b

12 龚全胜,周春光,马铭,刘全. 群核进化粒子群优化方法. 计算机科学,2005,32(8):134～137

13 龚全胜,周春光,马铭. 粒子群优化的两种改进策略. 计算机研究与发展,2005,42(5):897～904

14 Storn R, Price K. Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous space: [Technical report]. International Computer Science Institute, Berkley, 1995

15 Vesterstrom J, Thomsen R. A Comparative Study of Differential Evolution, Particle Swarm Optimization, and Evolutionary Algorithms on Numerical Benchmark Problems. In: Proceedings of Congress on Evolutionary Computation (CEC2004), Poland, 2004

16 Abbass H, Sarker R, Newton C. A Pareto frontier differential evolution approach for multiobjective optimization problems. In: Proceedings of the Congress on Evolutionary Computation, IEEE Service Centre, 2001, 2: 971～978

17 Krink T, Filipic B, Fogel G B. Noisy optimization problems -a particular challenge for differential evolution. In: Proceedings of Congress on Evolutionary Computation(CEC2004), Poland, 2004

18 Landa R, Becerra D, Carlos A. A Cultural Algorithm with differential Evolution to Solve Constrained Optimization Problems. Springer-Verlag, Lecture Notes in Artificial Intelligence Puebla, México, November 2004, 3315: 881～890

(上接第 195 页)

3 Yang Bingru. Knowledge Discovery Based on Inner Mechanism: Construction, Realization and Application. Elliott & Fitzpatrick Inc, USA, 2004

4 Yang Bingru, Shen Jiangtao, Song Wei. KDK Based Double-Basis Fusion Mechanism and Its Structural Model. International Journal of Artificial Intelligence Tools (IJAIT), 2005, 14(3)

5 Yang Bingru, Li Xin, Song Wei. Research on Inner Mechanism and KDTIM Theory of Knowledge Discovery. Engineering Science, 2004, 2(1): 46～51

6 Quest D, Ali H. Ontology specific data mining based on dynamic grammars. In: Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE, Aug. 2004. 16～19

7 Piatetsky-Shapiro G. The data-mining industry coming of age. Intelligent Systems, IEEE [see also IEEE Expert], 1999, 14(6)

8 Ramakrishnan N, Grama A Y. Data mining: from serendipity to science. Computer, 1999, 32(8)

9 Olaru C, Wehenkel L. Data mining. Computer Applications in Power, IEEE, 1999, 12(3)

10 Zeeman E C. Catastrophe theory: Selected papers(1972～1977). Reading, Massachusetts: Addson-Wesley, 1977

11 Goharian N, Grossman D, Raju N. Extending the undergraduate computer science curriculum to include data mining. In: Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on, Volume: 1, 5-7 April

12 周颖,杨炳儒. 基于语言场理论的连续属性离散化方法及实现. 计算机科学,2003,30(5)

13 杨炳儒. 基于内在机理的知识发现理论及其应用. 北京:电子工业出版社,2004

14 周颖. 对规则取舍问题研究. 计算机科学,2003,30(5):126～128

15 杨炳儒,綦艳霞. KDD 中因果关联规则的评价方法. 软件学报,2002,13(6):1142～1147

16 Yang Bingru, Zhou Ying. The Inner Mechanism of Knowledge Discovery System and its influence to KDD mainstream development. In: ICAI'2002, (USA), 2002. 826～832