不确定性数据上频繁项集挖掘的预处理方法

李海峰 章 宁 柴艳妹
(中央财经大学信息学院 北京 100081)

摘 要 传统频繁项集挖掘技术无法高效获取不确定性数据中有价值的信息。通过研究频繁模式增长树的算法原理,根据不确定性数据的特点提出了一种有效的不确定性数据预处理方法 PCAFP-Growth。利用主成分分析的方法进行数据的降维,并使用模糊关联分析法将数据概率进行分类,实现数据剪枝。在理论研究基础上,通过实验对数据集进行了验证。结果表明,基于主成分分析法的剪枝策略在稠密数据集上能够有效提高运算速度,减少内存的使用。

关键词 不确定性数据,频繁项集,主成分分析,模糊关联

中图法分类号 TP312 **文献标识码** A

Uncertain Data Preconditioning Method in Frequent Itemset Mining

LI Hai-feng ZHANG Ning CHAI Yan-mei
(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract Traditional studies of frequent itemset mining cannot obtain information from uncertain data efficiently. We studied the frequent pattern tree and proposed an effective uncertain data preconditioning method, the PCAFP-Growth, which can reduce the itemset dimensions with principal component analysis method, and prune data with fuzzy association analysis. Our experimental results over real world datasets show that our method is effective and efficient.

Keywords Uncertain data, Frequent itemset, Principle component analysis, Fuzzy association

1 引言

数据挖掘技术已经历近20年的发展,理论上日趋成熟,但以前的研究都是基于确定性数据上的挖掘算法。最近,人们的视角逐渐转向不确定性数据。在实际生产生活中,对不确定性数据上的知识挖掘变得更有意义。例如,在无线传感器网络中,由于无线信号的不稳定性和传感器能量的有限性,导致其不确定性无处不在,数据有可能无法完整传输,甚至是缺失的。对这样的数据赋予一个存在概率是必要的,这也是目前数据挖掘研究人员面临的难题之一。不确定性数据产生的原因比较复杂:可能是原始数据本身不准确或是采用了粗粒度的数据集合,也可能是为了满足特殊应用目的或是在处理缺失值中产生,同时数据集成过程中也会产生不确定性。本文将针对不确定性数据上的频繁项集挖掘算法进行研究,研究的数据对象是可能世界模型下的不确定性数据。假设数据库集合 D,其中项的集合为 $I=\{I_1,I_2,I_3,\cdots,I_n\}$,事务的集合为 $T=\{T_1,T_2,\cdots,T_m\}$,$T\subseteq I$,事务 T_i 中项 I_k 发生的概率为 p_i(其中 $0<p_i<1$)。给定最小支持度 λ,若一个项集的存在数量超过该阈值,则称该项集为频繁项集。

在进行数据挖掘工作中,需要设计高效的算法来寻找不确定性数据中的频繁项集。目前这方面的研究还比较少,必须在数据挖掘领域中探索一种更加行之有效的算法来适应不确定数据转变为确定性数据产生的海量规模。目前存在的问题主要包括两个方面:一是数据库实例太多,二是目前研究方法误差较大。

1.1 实例太多

当数据库中有 k 项时,若只考虑项的存在级不确定性,可能世界的数目将达到 2^k 个,数据库规模呈指数倍增长。当各项还有属性级不确定性时,可能世界的数目将远超过 2^k 个。在此情况下,罗列出所有可能世界实例是不可行的,更何况需要进一步处理各项复杂的数据挖掘操作,其计算开销惊人,并且会占用大量内存。如表1所列,当每个事务项的个数为2时,可能世界模型中将会出现8个实例,这8个实例发生的概率如表2所列。

表1 不确定性事务的数据集

ID	Transaction
TA	$(I_1,1.0);(I_3,0.1)$
TB	$(I_2,0.5);(I_3,0.6)$

表2 可能实例

World	TransactionDB	Prob.
1	$\{I_1\};\{\}$	0.18
2	$\{I_1,I_3\};\{\}$	0.02
3	$\{I_1\};\{I_2\}$	0.18
4	$\{I_1,I_3\};\{I_2\}$	0.02
5	$\{I_1\};\{I_3\}$	0.27
6	$\{I_1,I_3\};\{I_3\}$	0.03
7	$\{I_1\};\{I_2,I_3\}$	0.27
8	$\{I_1,I_3\};\{I_2,I_3\}$	0.03

另一方面,某些小概率事件对关联规则中查找频繁项集并没有太大的实际意义,如表3中的 $\{I_1,I_3\};\{\}$ 和 $\{I_1,I_3\};$

到稿日期:2011-08-15 返修日期:2011-11-23 本文受国家自然科学基金项目(61100112),中央财经大学科研创新团队支持计划资助。
李海峰(1979-),男,讲师,主要研究方向为数据挖掘、商务智能管理,E-mail:mydlhf@126.com;章 宁 女,教授,主要研究方向为数据挖掘和数据外包;柴艳妹 女,副教授,主要研究方向为图形图像处理和电子商务安全。

$\{I_2\}$的概率仅为0.02；而最大概率值达到0.27，是前者的13.5倍，这种小概率事件可以忽略不计。

1.2 目前研究方法误差较大

目前的研究方法大都将项的概率值简单相加，得到期望支持度，其计算方法如式(1)所示：

$$E(I_n)=\sum_{t_i\in D}P(I_n\subseteq t_i)(n=1,2,3,\cdots,k) \quad (1)$$

此时，引入方差概念，计算方法如式(2)所示：

$$D(I_n)=\sum_{I_n\in T_i}[P_n^i-E(I_n)/|T|]^2 \quad (2)$$

表3中的数据库是一个不确定性数据集合，包含了每个项的概率分布。

表3 $\{D_1\}$不确定性数据集

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
T_1	0.8	0.2	0	0.5	0	1	0
T_2	0	0.1	0.7	1	1	0	0.1
T_3	0.5	0	0	0.2	0	0.5	1
T_4	0	0	0	0.8	0.2	0	0.9
T_5	0	0	1	0.5	0.8	0	1
T_6	1	0.2	0.1	0	0	0	0

将期望及方差公式代入计算后，得到的结果排序如表4所列。可以看出，I_4和I_7的期望支持度相同，但是$I7$的存在可能性明显比I_4的要高，原因是在统计过程中，部分小概率事件多次发生，会对整体结果造成一定的干扰。故本文引入了基于主成分分析和模糊理论的剪枝策略，旨在剪去干扰项，尽可能产生接近实际情况的排序，并在运算过程中对小于支持度的项进行剪枝，以减少运算量。

表4 $\{D_1\}$项的期望支持度和方差

I_n	I_4	I_7	I_1	I_5	I	I_6	I_2	I_4
$E(I_n)$	3	3	2.3	2	1.8	1.5	0.5	3
$D(I_n)$	0.68	1.32	1.008	1.013	0.96	0.875	0.048	0.68

2 相关工作

最近几年，不确定性数据[1]逐渐成为研究的热点。根据建模方法的不同，不确定性数据的频繁项集挖掘算法[2]可以分为3类：可能世界模型下基于支持度和概率的频繁项集挖掘技术；概率模型下基于估计支持度的频繁项集挖掘技术；DS理论模型下基于信任度的频繁项集挖掘技术。

可能世界模型下基于支持度和概率的频繁项集挖掘技术：在文献[3]中定义的可能世界模型是不确定性数据建模的通用模型。可能世界模型下的频繁项集有两种定义方法：第一种是基于期望支持度的频繁项集，其使用支持度的期望值代替传统的支持度定义来计算频繁项集。在这种定义中，频繁项集挖掘的方法多是对确定性数据的数据挖掘算法的扩展。文献[4]提出的U-Apriori算法采用Apriori-Gen方法生成候选频繁项集，然后扫描数据集来验证候选项集的期望支持度。文献[5]在U-Apriori的基础上提出了递减计数器的剪枝策略，其可以迅速降低候选集合的规模。文献[6]提出了FP-Growth算法的扩展UFP-Growth，其利用UFP-tree的分裂节点来保存数据项目的发生概率，并通过离散概率值的方法或直接计算所有可能子项集期望支持度的方法来压缩UFP-tree。文献[7]应用了概率聚类，提出进一步的扩展算法UH-Mine，其采用链表来实现事务集合的遍历，用来代替实际建立的数据映射。第二种是基于概率的频繁项集，其使用了支持度和概率值两个参数作为决定频繁项集的衡量标准。文献[8]提出的PFIM算法采用了动态规划的思想计算项集成为频繁项集的概率，并利用EHH和单调性的剪枝策略来减少计算代价。

概率模型下基于估计支持度的频繁项集挖掘技术：概率模型认为，对于现实世界中真实情况的直接完美观察即为真实数据集。由其他不确定因素得到的观察为不确定性数据集，不确定性数据集都可以看作是由真实确定数据集随机转换得到的。概率模型就是对不确定性数据集根据概率论原理建模，通过不确定性数据集得到估计值作为真实确定性数据集最有可能的取值特征。EST模型[9]是针对频繁项集这一特定问题提出的广义模型，它对不确定性数据集建模，得到项集支持度的无偏估计值，将其作为真实数据集上最可能的取值特征，并使用该估计支持度定义频繁项集。

DS理论模型下基于信任度的频繁项集挖掘技术：DS理论是对贝叶斯理论的一般性扩展，是一种推理理论，也是建模不确定性数据的基本理论。它认为，在实际问题中，人们对命题的相信程度并不能反映出对其否命题的信任程度。DS证据理论在不确定性数据集的基础上定义基本概率函数，为不确定性数据分配信任度。在这种建模方法下，项集的信任度对应着支持度，由此获取不确定性频繁项集。文献[10]在只有一个不确定性数据维的证据数据库上计算频繁项集，将概率维放在数据结构的最底层；然后使用最大频繁项集进行剪枝，并采用深度优先遍历的方法挖掘频繁项集。文献[11]在含多个概率维的证据数据库上建立数据结构，支持度的计算取决于其数据结构建立时属性的顺序。文献[12]使用了RidList数据结构，综合包含项集X的元组来计算X出现的概率，从而得到项集X的支持度。

数据流环境中不确定性数据的频繁项集挖掘方法：尽管针对动态确定性数据和静态不确定性数据的频繁项集挖掘的研究都已经取得了丰硕的成果，但在数据流环境中针对不确定性数据的研究还处于起步阶段，其研究内容还比较分散。

文献[12]提出了在数据流上基于可能世界模型的概率频繁项目挖掘算法PHH。由于不确定性项目在向可能世界转化的过程中需要进行海量计算，因此PHH算法采用抽样技术来降低这种计算代价，同时设置阈值来保证抽样造成的误差被控制在一定范围之内。文献[13]通过对数据流上确定性数据的挖掘方法FP-Stream进行扩展，提出了改进的方法UFP-Stream，来获取近似的概率频繁项集；UFP-Stream采用近似阈值的方法来保证挖掘的实时性。另外，文献[14]提出了有延迟的频繁项集挖掘方法SUF-Stream，其通过将每个项集按照概率不同分裂为多个概率项集，来保存所有流数据，获取精确的概率频繁项集。文献[15]则基于具体约束条件来挖掘不确定性数据的频繁项集。

3 PCAFP-Growth方法

本文利用主成分分析法和基于模糊关联分类法的剪枝策略降低存储代价和计算代价，提出了PCAFP-Growth预处理方法。

3.1 主成分分析法

在不确定性数据集上，支持度计数不能再使用一个唯一的数值，必须用一个离散的概率分布来表示。在此，定义P_i

(I)为 I 的支持度为 i 时项集 I 的支持度概率，则有：

$$P_i(I)=\sum_{T_j\in T,(S(I,T_j)=i)}P(T_j) \tag{3}$$

式中，$S(I,T_j)$表示 I 在 T_j 的支持度。

数据集合中的项集存在性往往存在着一定的关系，这些项与项之间的关系不同于挖掘算法里的关联关系，它们的相关关系是基于概率模型下的相关。如果在进行关联规则挖掘之前通过主成分分析的方法对较小概率事件的项进行剪枝，将会大大提高运算速度，降低运算量。

主成分分析法的基本思想是这样的，主成分就是指多个变量的线型组合，不同主成分之间相互关联极小来实现用最小的主成分代表最多的信息的目的。其中，用项的概率值替代主成分分析法中的各项的贡献度；若存在 m 个事务，每个事务有 n 个项，那么可以利用这 n 个项进行线性组合：

$$F_i=a_{1i}X_1+a_{2i}X_2+\cdots+a_{ni}X_n(i=1,2,\cdots,n)$$

式中，$a_{1i}^2+a_{2i}^2+\cdots+a_{ni}^2=1$。$F_i$ 与 $F_j(i\neq j,i,j=1,2,\cdots,n)$ 不相关，F_1 到 F_n 方差依次递减。

主成分分析的计算步骤如下。首先计算相关系数矩阵：

$$R=\begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1n}\\ r_{21} & r_{22} & \cdots & r_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ r_{n1} & r_{n2} & \cdots & r_{nn}\end{bmatrix} \tag{4}$$

式中，$r_{ij}(i,j=1,2,\cdots,n)$为原变量的 x_i 与 x_j 之间的相关系数，其计算公式为：

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x_i})(x_{kj}-\bar{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x_i})^2\sum_{k=1}^{n}(x_{kj}-\bar{x_j})^2}} \tag{5}$$

通过式(5)计算出的 R 是实对称矩阵，故只需计算其上三角元素即可。其次，计算特征值与特征向量解特征方程 $|\lambda I-R|=0$，得特征值 $\lambda_i(i=1,2,\cdots,n)$，将其按大小顺序排列；根据特征值求出相应的特征向量 $e_i(i=1,2,\cdots,p)$，要求 $\sum_{j=1}^{p}e_{ij}^2=1$，其中 e_{ij} 表示向量 e_i 的第 j 个分量；然后计算主成分贡献率及累计贡献率。其中主成分 z_i 的贡献率为 $\lambda_i/\sum_{k=1}^{n}\lambda_k(i=1,2,\cdots,n)$，累计贡献率为：

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{n}\lambda_k}(i=1,2,\cdots,n) \tag{6}$$

通过式(6)的计算，一般取累计贡献率达 85%以上的特征值所对应的分别为第 1、第 2、…、第 $m(m<n)$个主成分。最后，计算主成分载荷 $l_{ij}=p(z_i,x_j)=\sqrt{\lambda_i}e_{ij}(i,j=1,2,\cdots,n)$。得到各主成分的载荷以后，进一步计算各主成分的得分。

通过主成分分析的方法降低项的维度，将降维后的主成分项重新进行计算，得到了一组新的项集。继而对新项集的概率进行调整，并遵循以下规则：1)将概率值小于零的概率项的值调整为零；2)将概率值大于 1 的项的概率调整为 1。在第一次扫描数据库时，得到相对准确的排序表，再根据此排序表构建树。同时在构建树的过程中直接减去低于最小支持度计数的项。

3.2 基于模糊关联分类法的剪枝策略

不确定性数据的概率分布情况未知，此时可以通过模糊分类法将数据概率进行分类，取大类进行频繁项集的挖掘。模糊关联分类法将根据 I_n 的概率取值范围划分为 K 类，概率 $p_{I_k}^i$ 属于第 j 类的隶属度函数为：

$$\mu_{K,j}^{p_{I_k}^i}(p_{I_k}^i)=\max\{1-|p_{I_k}^i-p_j^K|/b^K,0\} \tag{7}$$

式中，p_j^K 是 I_k 类的中心，表示为 $p_j^K=m_i+(m_a-m_i)(j-1)/(K-1)$；$b^K=(m_a-m_i)/(K-1)$，$m_a$ 是概率 $p_{I_k}^i$ 取值范围里的最大值，而 m_i 是概率 $p_{I_k}^i$ 取值范围里的最小值，b^K 是对应类边界。对于划分中心的选择，可以先结合建模样本对每个概率值按照模糊区间数利用 K-means 聚类方法进行聚类，找到相应的类中心并将其作为属性模糊区间的中心。相应地，取两个类别最靠近中心的点的距离中点为边界。

最后，给出模糊定义下的支持度的定义：

$$Supp_{\text{fuzzy}}=\sum_{T_i}^{D}\mu_{K,j}^{p_{I_k}^i}(p_{I_k}^i)/|D|$$

通过模糊关联分类法将一定概率分布的数据进行分类，直接剪去小于最小支持度阈值的项。在建立树的过程中，直接使用这些大类得到一个排序树，然后进行频繁项集的挖掘。由于事先对类进行了划分，因此在进行计算时得到的可能排序树的种类较原来少，而随着数据量的增加，计算的量级将会明显减小。

3.3 例子

以表 4 中的事务项集合为例，取最小支持度阈值为项集总个数的 20%。首先使用 UFP-Growth 算法得到频繁项集如下：$\{I_4,I_6\}$，$\{I_4,I_3\}$，$\{I_4,I_5\}$，$\{I_7,I_6\}\{I_7,I_3\}$，$\{I_1,I_6\}$，$\{I_3\}$，$\{I_5\}$，$\{I_4\}$，$\{I_7\}$，$\{I_4,I_7\}$，$\{I_1\}$，$\{I_6\}$。运行时间为 0.068s。再用本文中所使用的方法进行计算，得到支持度的概率分布，如图 1 所示。

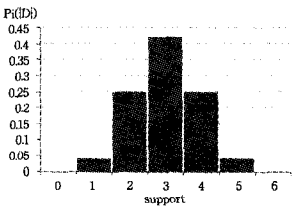

图 1 支持度的概率分布

项的相关系数矩阵计算如表 5 所列。进一步计算，得到总方差，如表 6 所列。从表中可以看出，前 3 个成分的累积贡献率超过了 85%，前 3 个主成分已经足够描述这些变量的水平，因此可以得到 3 个主成分的系数表，如表 7 所列。

表 5 相关系数矩阵

	1	2	3	4	5	6	7
1	1.0000	0.1781	−0.2735	−0.5955	0.3809	0.1278	0.3119
2	0.1781	1.0000	−0.2568	−0.5872	−0.2493	0.0747	−0.6850
3	−0.2735	−0.2568	1.0000	0.8510	−0.5630	0.1939	−0.3647
4	−0.5955	−0.5872	0.8510	1.0000	−0.4253	−0.0297	−0.1429
5	0.3809	−0.2493	−0.5630	−0.4253	1.0000	−0.5448	0.4821
6	0.1278	0.0747	0.1939	−0.0297	−0.5448	1.0000	0.2081
7	0.3119	−0.6850	−0.3647	−0.1429	0.4821	0.2081	1.0000

表 6 总方差

成分	初始特征值			提取平方和载入		
	合计	方差的%	累积%	合计	方差的%	累积%
1	2.84226	40.60374	40.60374	2.84226	40.60374	40.60374
2	1.95796	27.97088	68.57462	1.95796	27.97088	68.57462
3	1.38161	19.7373	88.31192	1.38161	19.7373	88.31192
4	0.68501	9.78586	98.09778			

表 7　成分系数矩阵

	Comp1	Comp2	Comp3
I_1	0.6551	−0.0872	0.3900
I_2	0.1845	−0.9789	−0.0661
I_3	−0.8758	0.1001	0.0981
I_4	−0.8835	0.4409	−0.1180
I_5	0.7613	0.4396	−0.3650
I_6	−0.1981	−0.1679	0.9364
I_7	0.4611	0.7525	0.4375

用 3 个主成分解释原始的 7 个变量。在矩阵中变量 I_1、I_5 对成分 1 的贡献度较大。I_7 对成分 2 的贡献度较大，I_6 对成分 3 的贡献度较大。通过计算由新变量构成的概率情况表，得到表 8。修正后的概率分布表中有些项的概率接近于 1，在此基础上运用 UFP-Growth 算法进行频繁项集的挖掘。运行时间为 0.039s，这个时间要比 UFP-Growth 的时间短。在数据量较小的情况下，主成分分析法并没有特别的优势，进行降维之后其可能会比传统方法消耗更多的时间。而随着数据量的增加，本文方法的优势将会逐渐体现出来。

表 8　修正后的概率分布表

ID	Comp1	Comp2	Comp3
T_1	1.0000	0.5022	1.0000
T_2	0.3294	0.3936	0.5998
T_3	0.0184	0.4910	0.7982
T_4	1.0000	0.3957	0.1950
T_5	0.1845	0.0000	0.9364
T_6	0.8396	0.0200	1.0000

4　实验与性能分析

采用 Matlab 在内存 2G，CPU 为 Xeon 2.0GHz，操作系统为 Windows Server 2003 的机器上实现并运行了本文提出的 PCAFP-Growth 预处理方法，同时将其与目前性能最优的不确定数据的频繁项集挖掘算法 UFP-Growth 进行了比较。采用了 2 种标准的数据集作为实验的测试数据集：网络链接信息 Connect4 数据集和匈牙利在线新闻门户网站的点击流数据集 KOSARAK。表 9 列举了这两种数据集的主要数据特征。其中前者的密度较高，后者的密度较低。

表 9　数据集的主要特征

数据集	事务数量	平均事务	最小事务	最大事务	项目数量
KOSARAK	990 980	7.6	1	380	41270
Connect4	67 557	42	42	42	128

图 2 表示了在密集型数据集 Connect4 下运行时间的比较情况。该时间是预处理时间与挖掘时间的总和。可以看出，随着最小支持度阈值的降低，UFP-Growth 运行速度明显减慢。相对来说，本文所使用基于主成分分析的 PCAFP-Growth 方法与传统 UFP-Growth 的运行时间相比其变化不大，无论最小支持度较大还是较小，运行速度均很快。这是因为经过主成分分析后数据变量的项明显减少，去除了大量冗余的计算。

如图 3 所示，针对在密集型数据集 Connect4 下的内存使用情况来说，尽管二者的内存使用都会随着最小支持度的降低而增大，但 UFP-Growth 的内存使用变化较大，而 PCAFP-Growth 的内存使用变化较小。

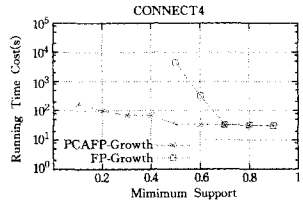

图 2　Connect4 数据集的运行时间比较

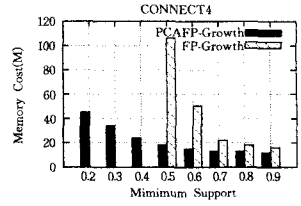

图 3　Connect4 数据集的内存使用比较

图 4 和图 5 表示在稀疏型数据集 Kosarak 下，运行时间和内存使用的比较情况。尽管 UFP-Growth 与 PCAFP-Growth 相比，运行效率和内存使用效率仍然较低，但其已经不会随最小支持度的变化而迅速变化。

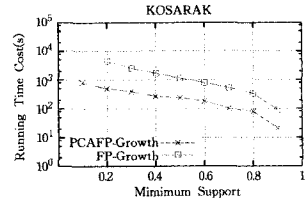

图 4　KOSARAK 数据集的运行时间比较

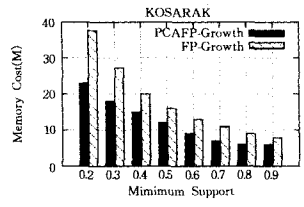

图 5　KOSARAK 数据集的内存使用比较

可以看出，提出的 PCAFP-Growth 算法尤其适合对稠密型数据集进行挖掘。在这种类型的数据上，PCAFP-Growth 算法能够有效进行数据的剪枝，实现运算效率和存储效率的大幅度提高。

结束语　本文在分析传统频繁项集挖掘方法的基础上，发现在构建树的过程中项集的排序直接影响着频繁项集的挖掘速度。在此基础上提出了主成分分析和模糊理论的剪枝策略，其能够利用较少的数据量表示大量的数据信息，从而有效地减小了挖掘的代价。通过数据集的验证，发现经过剪枝之后，在对稠密型数据集进行频繁项集挖掘时，运算速度得到了很大的提高，内存使用也有了明显的下降。

不确定性数据大多以一定的概率分布存在。在未来的研究中，可以利用蒙特卡罗模型对不确定性数据建模，它比 R. Muntz 等人提出的 EST 模型更加精确。

参 考 文 献

[1] 周傲英，金澈清，王国仁，等. 不确定性数据管理技术研究综述[J]. 计算机学报，2009，31(4)

[2] 李建中，于戈，周傲英. 不确定性数据管理的要求与挑战[J]. 中国计算机学会通讯，2009，5(4)

[3] Pei H J, Yin Y. Mining frequent patterns without candidate generation[C]// International Conference of SIGMOD. 2000

[4] Chui C, Kao B, Hung E. Mining frequent itemsets from uncertain data[C]// International Conference of Pacific-Asia Knowledge Discovery and Data Mining. 2007

[5] Chui C, Kao B. A decremental approach for mining frequent itemsets from uncertain data[C]// International Conference of Pacific-Asia Knowledge Discovery and Data Mining. 2008

[6] Leung C, Matco M A F, Brajczuk D A. A tree-based approach for frequent pattern mining from uncertain data[C]// International Conference of Pacific-Asia Knowledge Discovery and Data Mining. 2008

（下转第 199 页）

则第1层至第 d 层全部为真节点，第 $d+1$ 层及第 $d+2$ 层有部分真节点，第 $d+3$ 层及后面层的情况还需进一步判断 θ 与 $AS(W^L_{[n-d-3]})$ 之间的大小关系；

(4)若 $AS(W^L_{[n-d]})<\theta\leqslant AS(W^H_{[n-d-2]})$，则第 d 层、第 $d+1$ 层及第 $d+2$ 层有部分真节点，其他层的情况还需进一步判断；

(5)若 $AS(W^L_{[n-d]})<AS(W^H_{[n-d-2]})<\theta\leqslant AS(W^H_{[n-d-1]})$，则第 d 层及第 $d+1$ 层有部分真节点，$d+2$ 层及其后面无真节点，第 d 层之前的情况还需进一步判断。

对凹超立方体、凸超立方体或 SP 函数的结构进行分析，将其标注到汉明图上，可以发现其属于上述5种情况中的某一部分，但并不清楚是否完全覆盖这5种情况。另外，超立方体可归类为特殊的 SP 函数，而在已知的线性可分函数系中，SP 函数没有找到阈值与权值直接判别法，故需对 SP 函数的判别法提炼出直接判别法，这将会进一步解决线性可分函数系的覆盖问题。

结束语 已知的线性可分结构系中正超立方体、汉明球、线性可分的汉明球突的判别法较为明确简洁，而凹超立方体、凸超立方体及 SP 函数的判别方法较为复杂，有待进一步简化。另外，根据阈值在实轴上的取值范围，通过本文的分析，目前已知的几类线性可分结构系仍不能覆盖所有的二进神经元，即仍存在其他线性可分结构系，有待进一步研究。本文指出了未知的线性可分结构系阈值的取值范围，为提出新的线性可分结构系及判别法提供了研究方向。

参 考 文 献

[1] Gray D L, Michel A N. A training algorithm for binary feed forward neural networks[J]. IEEE Trans. Neural Networks, 1992, 3(2): 176-194

[2] Kim J H, Park S K. The geometrical learning of binary neural networks[J]. IEEE Trans. Neural Networks, 1995, 6(1): 237-247

[3] Muselli M. On sequential construction of binary neural networks [J]. IEEE Trans Neural Networks, 1995, 6(3): 678-690

[4] Yamamoto A, Saito T. A flexible learning algorithm for binary neural networks[J]. IEICE Trans. Fundamentals, 1998, E81-A (9): 1925-1930

[5] Chen Fang-yue, Chen Guan-rong, He Guo-long. Universal Perceptron and DNA-Like Learning Algorithm for Binary Neural Networks: LSBF and PBF Implementations[J]. IEEE Transactions on Neural network, 2009, 20(10): 1645-1658

[6] Chen Fang-yue, Chen Guan-rong, He Qin-bin. Universal Perceptron and DNA-Like Learning Algorithm for Binary Neural Networks: Non-LSBF Implementation[J]. IEEE Transactions on Neural network, 2009, 20(8): 1293-1301

[7] Lu Yang, Yang Juan, Wang Qiang, et al. The upper bound of the minimal number of hidden nodes in parity problem by neural networks[J]. Science China Information Sciences, Oct. 2011

[8] Chua L O. Visions of nonlinear science in the 21st century in CNN: A Paradigm for Complexity[M]. Singapore: World Scientific, 1999

[9] Chen Fang-yue, He Guo-long, Chen Guan-rong. Realization of Boolean Functions via CNN: Mathematical Theory, LSBF and Template Design[J]. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2006, 53(10): 2203-2213

[10] Wegener I. The complexity of Boolean function[M]. New York: Wiley, 1987

[11] Crounse K R, Fung E L, Chua L O. Efficient implementation of neighborhood logic for cellular automata via the cellular neural network universal machine[J]. IEEE Transactions on Circuit System I, Fundam. Theory Application, 1997, 4(3): 355-361

[12] Nemes L, Chua L O, Roska T. Implementation of arbitrary Boolean function on a CNN universal machine[J]. Int. J. Circuit Theory Appl., 1998, 26(6): 593-610

[13] Haykin S. Neural Networks: A comprehensive foundation (2nd ed) [M]. Englewoood Cliffs, NJ: Prentice-Hall, 1999

[14] Negnevitsky M. Artificial Intelligence: A guide to intelligent systems(2nd ed)[M]. New York: Addision-Wesley, 2004

[15] Hassoun M H. Fundamentals of artificial neural networks[M]. New York: MIT Press, 1995

[16] 陆阳，韩江洪，张维勇. 二进神经网络逻辑关系判据及等价性规则提取[J]. 模式识别与人工智能，2001，14(2)：171-176

[17] 陆阳，魏臻，高隽，等. 二进神经网络中汉明球的逻辑意义及一般判别方法[J]. 计算机研究与发展，2002，3(1)：79-86

[18] 杨娟，陆阳，黄振谨，等. 二进神经网络中的汉明球突及其线性可分性[J]. 自动化学报，2011，37(6)：737-745

[19] 陆阳，韩江洪，张维勇. 二进神经网络中笛卡尔球的研究[J]. 模式识别与人工智能，2004，17(3)：368-373

[20] Lu Yang, Han Jiang-hong, Wei Zhen. A general judging and constructing method of SP functions in binary neural networks[J]. Acta Automatica S inica, 2003, 29(2): 234-241

(上接第 164 页)

[7] Aggarwal C C, Li Y, Wang J, et al. Frequent pattern mining with uncertain data[C]// International Conference on Knowledge Discovery and Data Mining. 2009

[8] Bernecker T, Kriegel H, Renz M, et al. Probabilistic frequent itemset mining in uncertain databases[C]// International Conference on Knowledge Discovery and Data Mining. 2009

[9] Muntz R, Mining Y. Frequent Itemsets in Uncertain Datasets [R]. CSD-TR No. 030042

[10] Tobji M A B, Yaghlane B B, Mellouli K. Frequent Itemset Mining from Databases Including One Evidential Attribute[C]// International Conference on Scalable Uncertainty Management. 2008

[11] Hewawasam K K R G K, Premaratne K, Subasingha S P, et al. Rule Mining and Classification in Imperfect Databases[J]. IEEE Transactions on Systems, Man and Cybernetics, 2007

[12] Tobji M A B, Yaghlane B B, Mellouli K. A New Algorithm for Mining Frequent Item-sets from Evidential Databases[C]// International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. 2008

[13] Zhang Q, Li F, Yi K. Finding frequent items in probabilistic data [C]// International Conference of SIGMOD. 2008

[14] Leung C K, Hao B. Mining of frequent itemsets from streams of uncertain data[C]// IEEE International Conference on Data Engineering. 2009

[15] Leung C K, Hao B, Jiang F. Constrained frequent itemset mining from uncertain data streams[C]// IEEE International Conference on Data Engineering. 2010