一种基于关联规则的离群数据挖掘算法及其应用

张璐璐[1,2]，贾瑞玉[1]，李学俊[1]

（1. 安徽大学 计算机科学与技术学院，安徽 合肥 230039；
2. 解放军炮兵学院 基础部计算中心，安徽 合肥 230031）

摘 要：文中基于对传统 Apriori 算法的改进，提出了一种基于规则的离群数据挖掘算法。该算法在数据结构中增加标识符链表后，计算了 1－离群条件集的幂集，使得仅需对原数据库进行一次扫描，从而降低了该算法的时间复杂度。同时由于兴趣度的引入使得挖掘的结果也更有针对性和目的性。该算法被应用于某求职系统的离群数据分析中，实验表明该算法是可行有效的。

关键词：离群挖掘；关联规则；兴趣度；Apriori 算法

中图分类号：TP311.13；TP301.6　　**文献标识码**：A　　**文章编号**：1673－629X(2007)08－0110－03

An Association Rule－Based Algorithm for Outlier Mining and Its Application

ZHANG Lu-lu[1,2]，JIA Rui-yu[1]，LI Xue-jun[1]

（1. Department of Computer Science and Technology，Anhui Univ.，Hefei 230039，China；
2. Computation Centre，Basic Science Department，PLA Artillery Academy，Hefei 230031，China）

Abstract：An improved rule－based outlier mining algorithm is proposed in this paper based on the conventional Apriori algorithm. In this algorithm，after the identifier linked lists being added to the data structure，the power set of 1－outlier－rules is formed. Only one scanning of the data source is needed using this method，which reduces the time complexity of this algorithm distinctly. The introduction of the degree of interest also makes the results more pertinent and reasonable. This algorithm has been used in the analysis of a real－world job－hunting system. It exhibits satisfiable performance.

Key words：outlier mining；association rule；interest；Apriori algorithm

0 引 言

就业问题是目前一个普遍而广泛的话题，各种各样的就业网站数据库中都包含了大量的求职者信息。但由于信息意识和技术的缺乏，管理人员只能够通过简单的统计或查询等功能获得表面的信息，隐藏在这些大量数据中的信息一直没有得到很好的应用。

离群数据挖掘[1]是指从大量数据中挖掘出明显偏离、不满足一般行为模式的数据，简称离群挖掘。目前，离群挖掘已在多个领域得到广泛研究[2]，如防范信用卡欺诈、天气预测、电力和用户分类等方面。但对其在职业领域中的应用与研究尚不多见。利用离群挖掘技术对这些求职数据作理性的分析将在就业引导中起着重要的指导作用。

文中尝试利用离群挖掘方法对求职者信息数据进行分析，发现那些值得注意的例外对象，从而为管理者提供决策支持。

当前离群数据挖掘的算法[3]众多，对于不同的应用范围其实现效果各异：如基于统计的方法，需要用户建立数据点的概率分布模型，应用时需事先知道数据集的分布和分布参数等信息，对数据要求比较严格，因此适用范围不大；基于距离的离群数据挖掘方法，这种方法中的距离难以确定；基于偏离的离群数据发现方法，需要确定相异函数进行离群数据挖掘，但若其相异函数的选取不合适，则得不到满意的结果等。

文中在已有的离群挖掘算法[4]的基础上，针对求职数据库的基本特征，提出了一种基于规则的离群数据挖掘算法。实验结果表明它在求职数据的分析中取得良好的效果，并能够为管理决策提供有用信息。

收稿日期：2006－10－16

基金项目：安徽省教育厅自然科学基金资助项目(2005kj056)

作者简介：张璐璐(1983－)，女，安徽阜阳人，硕士研究生，研究方向为数据挖掘、机器学习；贾瑞玉，副教授，硕导，研究方向为机器学习、计算机图形学。

1　基于规则的离群挖掘算法

1.1　算法思想

文中提出的基于规则的离群数据挖掘算法针对 Apriori 算法不足之处[5] 作如下改进:引入兴趣度以消除不需被重视的规则,从而解决了规则没有价值,甚至错误的问题;在数据结构中包含信息的标识符链表 Tid_ list,并对 1 - 离群条件集 O_1 作幂集运算,以达到扫描原数据库仅需一次的目的从而极大地提高了挖掘的效率。

算法具体步骤如下:

(1)首先逐条扫描数据库产生 1 - 候选集 C_1,在扫描每一条信息时,除了对每一项进行计数以外,还要记录包含该条的信息的标识符 Tid。这样扫描完一遍数据库之后,得到的候选集 C_1 中,每个项都包含了一个相应的标识符链表 Tid_ list。C_1 的每一个元素的结构如下:(项 item,支持度 sup(item),兴趣度值 intr(item),标识符链表 Tid_ list(item));

然后,从 C_1 中删除大于最大离群支持度的条件项,则此时 C_1 即为 1 - 离群条件集 O_1。

(2)求出项集 O_1 的幂集,此时幂集中的每一个元素结构如下:

(项集 item_ set,支持度 sizeof(Tid_ list(item_ set)),兴趣度值 intr(item_ set),标识符链表 Tid_ list(item_ set));

然后,从幂集中删除:大于最大离群支持度的条件项、支持度为 0 的条件项、兴趣度值小于最小离群兴趣度的条件项。此时幂集中的项集即为所有的离群条件项集 O_{set}。

(3)对于每一个包含 i 个条件项的离群条件项集 O_{si},先计算左部只有一个项的,产生所有满足最小置信度 min_ conf 的关联规则的集合 R_i,然后再计算左部有两个项的规则,依次类推,直至左部有 $i-1$ 项为止。同时每条规则所对应的离群数据亦可由 Tid_ list(item_ set) 记录。

1.2　算法具体描述

Input:求职数据库 JDB,最大离群支持度 max_ sup,最小离群兴趣度 min_ intr,最小置信度 min_ conf

Out:Rule=离群条件集;O=离群数据集

Begin

Rule=∅

C_1 = create_ candidate_ 1(JDB);//创建 1 - 候选集

$O_1 = \{c \in C_1 \mid 0 < \sup(c) \leq \max\_\sup$;//生成 1 - 离群条件集

V = power(O_1);//求出 1 - 离群条件集的幂集,且按照数据结构,计算每一元素的属性值

If (V could be inserted in EMS memory whole)

Then { Scan(V);}

Else{divide V into V_i; | $V_i \in V$, V_i could be inserted in EMS memory whole}

For($V_i \in V$)

Scan(V_i);

Endif;

Result = {item_ set | item_ set $\in V$, $0 <$ sup(Item_ set) $\leq$ max_ sup and intr(item_ set) $\geq$ min_ intr};//得到离群条件集

For(item_ set $\in$ Result)//对每一个离群条件

i = Count(item_ set);//记录条件项的个数

For($j = 1; j < i; j++$)

$\text{Rule}_i = (\text{item}_1 \wedge \text{item}_2 \wedge \cdots \wedge \text{item}_j \Rightarrow \text{item}_{j+1} \wedge \text{item}_{j+2} \wedge \cdots \wedge \text{item}_i \mid \text{comf} \geq \text{min\_ conf})$

Rule = $\bigcup \text{Rule}_i$;//得到离群规则集

O = Tid_ list(r_1) $\cup$ Tid_ list(r_2) $\cup \cdots \cup$ Tid_ list(r_k) | $r_1, r_2, \cdots, r_k$ 为 Result 中的所有元素;//得到离群数据集

End

2　实验及结果分析

2.1　求职数据库预处理

对于求职数据库的预处理主要包含以下各步工作:

1)数据清理[2]:去除源数据集中的噪声数据和无关数据,处理遗漏数据和清洗脏数据,考虑时间顺序和数据变化等。

2)数据变换:找到数据的特征表示,用维变换或转换方法减少有效变量的数目或找到数据的不变式,将数据转换成适合于挖掘的形式。

3)属性构造:求职数据库涉及到 15 项。这些信息都是以职位类型、期望行业、期望职位、目前状态、工作经验 、最高学历、毕业院校、专业、教育经历、自我评价等形式描述,无法对就业信息进行量化分析。因此需要采用科学方法对有关信息进行评估。文中采用 FAP(Fill in with Average Poll result)法[6]对求职信息进行属性构造,以便评估分析。

4)布尔转换:文中对求职数据库进行关联规则发现,须将数据库中有关属性进行布尔转换。

下文使用一个具体实例来说明预处理过程,预处理前(见表 1):

预处理后：

0(姓名，此时亦为信息的 Tid)：0(出生)，0(性别)，0(婚姻)，1(全职)，2(行业)，4(职位)，1(工资)，0(现状)，1(工作经验)，2(学历)，1(专业)，0(毕业时间)，0(外语水平)，0(计算机水平)。

表 1 求职数据库某条原始记录

姓名	出生	性别	婚姻	期望行业	期望职位	工资待遇	现状
某某	1987－12	女	未婚	外贸类	市场采购或仓库收货员	1000－1999	在校学生
学历	工作经验	专业	毕业院校		计算机水平	外语水平	
大专	嘉诚外贸	英语	义乌工商学院		一般	英语，一般	

2.2 实验结果分析

为对算法的有效性和性能作测试，使用预处理过的 www.91job.com 求职数据库作为测试数据集进行实现。实验的分析的环境为 P4，CPU：3.0G，内存：512M 的 PC 机 ；以 Windows2000 Advance Server 和 SQL Server 为软件平台。

数据如下：

0:0,0,0,1,2,4,1,0,1,2,1,0,0,0

1:0,0,0,1,2,3,1,1,1,1,3,1,0,1

2:1,0,0,0,4,1,0,1,1,2,6,1,0,1

3:0,1,0,1,2,1,0,1,1,1,6,1,1,1

……

兴趣度设定：

出生:2，性别:2，婚姻:1，全职:3，行业:1，职位:2，工资:3，现状:1，工作经验:4，学历:3，专业:4，毕业时间:2，外语水平:4，计算机水平:3

实验直接在支持度为 0.1 的基础上，对兴趣度的变化所产生的对时间的影响进行考察。

图 1(来源于 http://www.hfrc.cn/addon/2006/paoyuan.htm)表示了算法随兴趣度的变化在算法耗费时间上的变化情况。

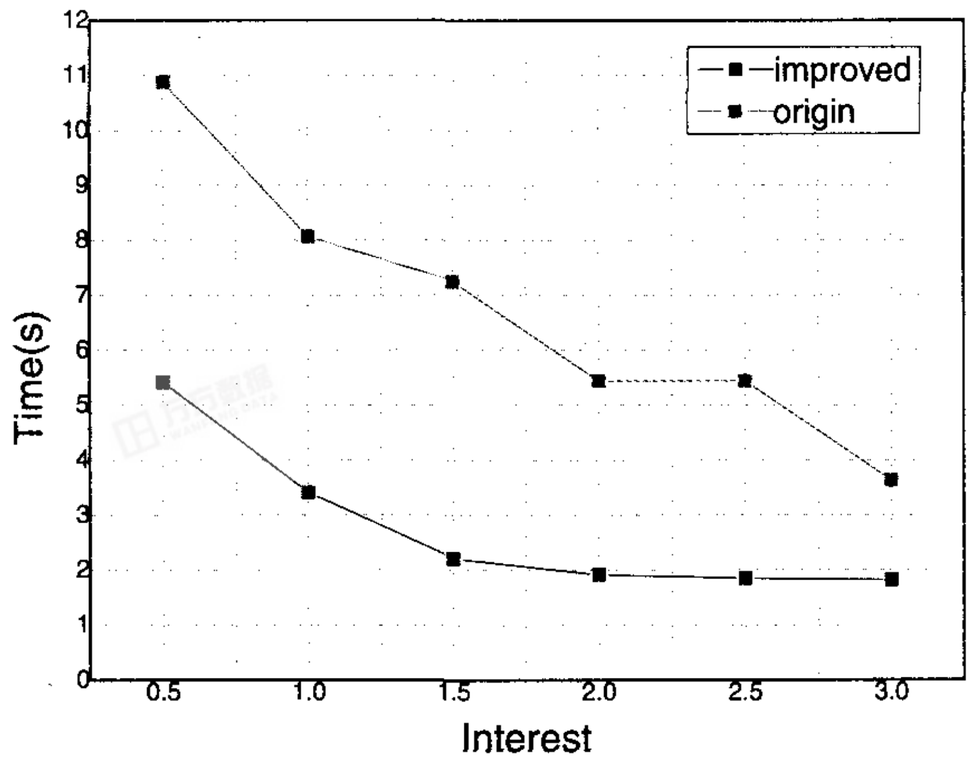

图 1 原算法及改进算法的运行时间曲线

由图 1 知，算法耗费的时间随着兴趣度阈值的增加而递减，因为随着兴趣度的增大，越来越多的条件项集被淘汰，从而约减离群规则。这就说明使用兴趣度作为离群规则度量的有效性。图 1 对比了原算法及改进算法在相同条件的时间耗费。可以看出改进算法在一定程度上降低了算法的时间复杂性，减少了算法耗时，提高了数据挖掘的效率。

实验中，设定支持度为 0.1，置信度为 0.6，对原算法和改进算法在规则的获取上进行比较。由图 2 知，改进算法比原算法过滤掉更多的规则。改进算法在置信度的基础上引入了兴趣度，这样就相当于在原算法的基础上进行二次关联规则过滤，使挖掘的规则更加客观、合理。

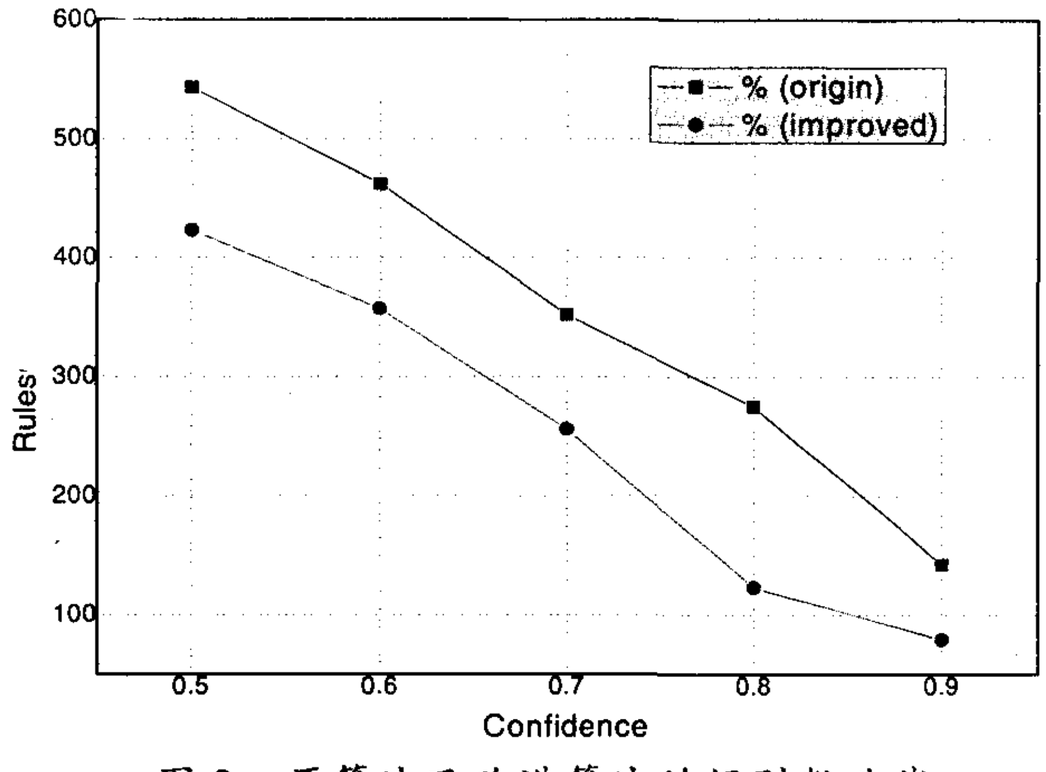

图 2 原算法及改进算法的规则数曲线

实验表明这种方法是正确的，算法中加入了属性的兴趣度能够更加突现有价值的离群条件项集和使决策者感兴趣的离群数据，从而能大量地约减冗余的关联规则，可以更好地起到帮助决策的作用。

3 结束语

文中将关联规则与离群数据挖掘相结合，提出了一种基于规则的离群挖掘算法。该算法能够改进 Apriori 算法的不足：在数据结构中增加标识符链表后，计算了 1－离群条件集的幂集，使得仅需对原数据库进行一次扫描，从而降低了该算法的时间复杂度。同时由于兴趣度的引入使得挖掘的结果也更有针对性和目的性。利用此算法对求职数据进行离群分析，不仅证明了算法的可行有效，而且所得结论对就业指导部门的工作提供了有益的参考。

由于将离群挖掘技术应用于求职数据中，还属于一个比较崭新的课题。如何在关联规则的增量式更新问题上对算法进行改进还有待进一步的研究。

参考文献：

[1] Knorr E，Ng R. Algorithms for mining distance－based outliers in large datasets[C]//In：Proc of the 24 VLDB Conf. New York，USA：[s.n.]，1998：392－403.

(下转第 116 页)

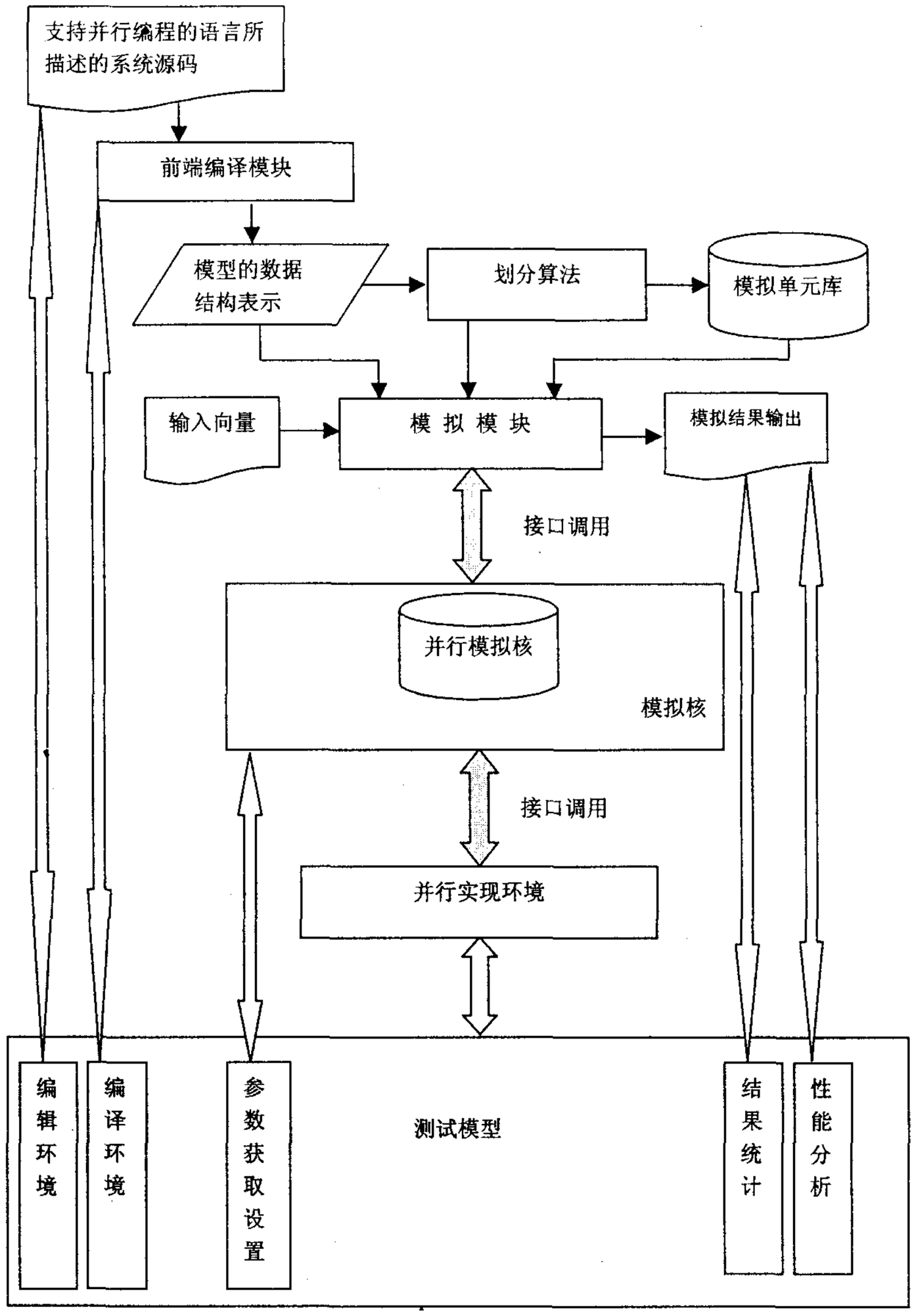

图 1 测试模型结构图

重制约了相关算法的研究工作进展。文中在测试模型的建立方面进行研究,期望提供有效的测试环境,达到简化测试过程,提高测试效率和测试效果的目的。

参考文献:

[1] Subramanian S, Rao D M, Wilsey P A. Study of a Multilevel Approach to Partitioning for Parallel Logic Simulation[C]//14th International Parallel and Distributed Processing Symposium (IPDPS'00). Cancun, Mexico: [s. n.], 2000.

[2] Chamberlain R D. Parallel Logic Simulation of VLSI Systems[C]//in Proc. of 32nd Design Automation Conf. San Francisco: ACM Press, 1995:139 - 143.

[3] Bagrodia R L, Liao W. Maisie: A language for the design of efficient discrete - event simulations[J]. IEEE Transactions on Software Engineering, 1994, 20(4):225 - 238.

[4] Karger D, Stein C. A new approach to the minimum cut problem[J]. Journal of the ACM, 1996, 43(4):601 - 640.

[5] IEEE Computer Socitety. IEEE Standard for VHDL Register Transfer Level(RTL) Synthesis[S]. IEEE Std 1076.6 - 1999. [s. l.]: IEEE, 2000.

[6] Ferxcha A. Parallel and distributed simulation of discrete event system[C]//Parallel and distributed computing handbook. American: Mc Graw - Hill, 1995:85 - 89.

[7] 李 暾,李思昆. 一种启发式并行逻辑模拟划分算法[J]. 系统工程与电子技术, 1999, 21(9):68 - 70.

[8] 李俊红,杨洪斌,吴 悦. 时间偏差算法中通讯接口的研究及实现[J]. 计算机工程与设计, 2005, 26(1):66 - 68.

[9] Wu yue, Li junhong, Yang hongbin. The study of cancellation strategy in rollback mechanism[J]. Journal of Shanghai University, 2005, 9(6): 501 - 505.

[10] Subramanian S, Rao D M, Wilsey P A. Applying Multilevel Partitioning to Parallel Logic Simulation[EB/OL]. 2000. http://www.ececs.uc.edu/paw/lab/pubs.html.

(上接第 112 页)

[2] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明, 孟小峰, 等译. 北京:机械工业出版社, 2001:223 - 261.

[3] Arning A, Agrawal A, Raghavan P. A linear method for deviation detection in large database[C]//Int Conf on knowledge Discovery in Databases and Data Mining. Portland: [s. n.], 1996:169 - 184.

[4] Agrawal R, Inielinski T, Swami A. Mining association rules between sets of items in large database[C]//In: Bunemuu P, Jajodia S eds. Proceedings of the 1993 ACM SCIMOD Conference on Management of Data. New York, NY: ACM Press, 1993:207 - 216.

[5] Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules[C]//In: Proceedings of the 21st International Conference on Very Large Database. Zurich, Switzerland: [s. n.], 1995:432 - 444.

[6] 吕建军. 数据挖掘技术的应用研究[D]. 北京:中国农业大学, 2002.