生化企业生产数据知识挖掘系统①

熊结青[1]，沙宗尧[2]

[1](安徽丰原生物化学股份有限公司，蚌埠 233010)
[2](武汉大学 国际软件学院，武汉 430079)

摘 要：对于生物化工产品的工业生产而言，更要有合适的生产环境条件，然而由于生产过程的复杂性，确定适宜的生产环境较为困难。就生化企业生产的数据特征，提出了生产数据的指标分割预处理及针对稀有数据的关联规则挖掘方法，对数据指标分割的过程进行了详细的阐述，并针对稀有数据挖掘，提出了关联规则挖掘中相对支持度的概念，在此基础上设计并开发生化企业关联规则挖掘数据分析系统，给出了系统的结构和功能，并对系统应用进行了试验和分析，取得了较好的效果。

关键词：关联规则挖掘；知识发现；生化企业；知识挖掘系统

Association Knowledge Discovery System for Bio-Chemical Enterprises

XIONG Jie-Qing[1], SHA Zong-Yao[2]

[1](Anhui Fengyuan Bio-chemical Co. Ltd, Bengbu 233010, China)
[2](International Software School, Wuhan University, Wuhan 430079, China)

Abstract: Data Mining is a process of discovering knowledge or rules from available dataset. It is possible to make optimized production environment based on association rules mined from production data through association rule mining system. During the production process of Bio-chemical enterprises, it is expected that the environment such as temperature, water condition and raw material supply are appropriate to obtain best production output. However, due to the complexity of production, it is not easy to acquire the optimized condition. This paper therefore tries to design and develop a knowledge discovery software program intending to provide such tools. The paper also analyzes index segmentation and association rule mining from rare transactions from large dataset based on relative supporting value, which is illustrated in detail. The developed system was tested and verified in practice, which proves that the system can be applied in Bio-chemical production data analysis.

Key words: association rule mining; knowledge discovery; Bio-chemical enterprises; knowledge discovery system

生物化工企业一般应用微生物、物理、化学的方法对原料进行加工处理，生产出各种产品。在生产过程中，生产环境（如温度、湿度、pH 值等）的差异可导致不同生产效益，因此，工业生产环境一般都有一定的优化范围，来达到最大的产出效应。

知识挖掘是利用某些特定的知识发现算法，从数据中提取出感兴趣的模式，在数据挖掘中，关联规则是最早引入商业应用的领域之一，关联规则反映了数据中不同数据项间的关联性，通过挖掘关联规则，可以分析和理解数据库中不同数据项间的关联关系[1,2]。笔者长期从事生物化工生产中，积累了大量的生产数据，以往大多采用经典统计分析方法，来揭示生产环境与生产目标之间的关系，但随着生产环境及多因子间的交互关系的进一步复杂化，统计方法显得越来越麻烦，效果也越来越差。因此，我们尝试利用关联规则挖掘方法，来进行生物化工生产环境优化。例如对

① 基金项目:国家自然科学基金(41071219);蚌埠市企业节能增效项目(2007-10)
收稿时间:2010-12-24;收到修改稿时间:2011-02-21

于下图 1 所示截取的部分生产数据，对于决策者关注的转化率和总酸两个生产指标，如果能找到较高的转化率和总酸的环境生产因子（如 OD、配方、玉米浆等），对提高工业生产的效益是很重要的。

	B	E	F	N	V	W	X
1	pH	种子罐OD	起始AN	配方	总酸	转化率	玉米浆TN
2	6.85	0.879	0.3342	F	38.70	56.68	4.4486
3	6.87	0.873	0.3207	F	39.97	58.74	4.4486
4	6.79	0.923	0.3100	E	39.83	59.98	3.9906
5	6.84	0.875	0.3171	F	39.86	58.80	3.9906
6	6.78	0.870	0.3257	F	38.25	58.41	3.9906
7	6.92	0.876	0.2779	G	35.99	59.26	3.9906
8	6.92	0.875	0.3214	F	35.50	59.11	3.9906
9	6.84	0.876	0.3010	F	36.64	58.23	3.9906
10	6.94	0.912	0.2872	H	38.41	53.55	3.9549

图 1　部分生产数据示例

对于生化企业，当积累的生产数据量很大、数据指标很多时，挖掘上述可能的所有关联规则需要通过软件系统来实现。本文所提出的“生化企业生产数据知识挖掘系统”，就是针对企业的生产数据，从中发现隐藏规律，为企业生产条件优化提供参考信息。

1　基本理论

关联规则可以用以下数学模型加以描述：令 $I=\{i_1, i_2,\ldots,i_m\}$是字母（指标）集，称为数据项或项目，$D=\{D_1, D_2, D_n, \ldots\}$是全体交易事务的集合，事务 T 是 I 的一个子集，即 T⊂I，每个事务由唯一的标志 TID 标识。对数据项集 X⊂I，称 T 包含 X 当且仅当 X⊆T，这里，因果关联规则具有如下形式：X→Y，这里 X⊆I，Y⊆I，且 X∩Y＝φ，X 称为规则的条件，Y 称为规则的结果。集合 D 中规则 X→Y 的支持度定义为 D 中 s%的事务包含 X∪Y，规则 X→Y 对集合 D 的置信度定义为 D 中 c%的事务包含 X 且包含 Y。支持度表示规则的频度，支持度大于给定的阀值的规则称频繁规则，置信度表示规则的强度，置信度和支持度均大于给定阈值的规则称为强规则。

关联规则的挖掘一般包括以下两个过程：（1）首先找出支持度大于给定值的频繁数据项集。（2）用频繁数据项集挖掘出强关联规则。经典关联规则采掘算法为 Apriori 算法[3]，其基本思想是：首先通过扫描数据库，产生一个大的候选数据项集，并计算每个候选数据项发生的次数，基于预先给定的最小支持度（s%）生成一维数据项集 L_1，然后基于 L_1 和数据库中的数据，产生二维数据项集 L_2；用同样的方法，直到生成 N 维数据项集 L_N，其中已不再可能生成满足最小支持度的 N+1 维数据项集．这样就依次产生数据项集$\{L_1, L_2, \ldots, L_N\}$，最后，通过步骤二，以最小信任度（s%）为过滤条件，从频繁数据项集中产生强规则。

表 1　生产数据集（数据项）

样本号	A	B	C	D
1	1	1	0	1
2	0	0	1	1
3	0	1	1	1
4	0	0	1	1
5	1	1	1	1
6	0	0	0	0
7	0	0	1	1
8	0	0	0	1
9	0	0	0	0

实际生产中的数据都是表的形式存在，如表 1。假设对于上述数据表，A、B、C 为因子项，D 为决策项，关联规则可以表示成：因子项（全部或部分）→ 决策项，符号“→”表示可“可产生”。例如规则 A（=1）+B（=1）→ D（=1），但度量该规则强度时需要支持度和信任度，按上述关联规则定义，该规则的支持度和信任度分别为 0.22 和 1.00。表 1 中，因子项和决策项的取值都是 0 或 1，原始生产数据需要“指标分割”预处理，使之转化成表 1 的格式，例如原始的生产数据中假如指标“温度”，在样本中，温度取值从 32.2~38.8 之间（表 2（a）），可根据生产知识，把指标“温度”分割成低温“温度_0”（32.2~35.0）和高温“温度_1”（35.0~38.8）两个数据项（表 2（b））。可以看出，指标分割前后的样本数是不变的，但数据项一般要比指标项多。

表 2　指标分割示例

样本	温度	…
1	33.7	
2	32.2	
3	35.9	
4	38.8	
5	36.7	
6	37.3	
7	38.2	
8	35.1	
9	34.5	

(a)原始数据

→

样本	温度_0	温度_1	…
1	1	0	
2	1	0	
3	0	1	
4	0	1	
5	0	1	
6	0	1	
7	0	1	
8	0	1	
9	1	0	

(b)分割结果数据

在分析数据时发现，当上述生产数据中存在稀有数据时，采用支持度和信任度往往会遗漏某些重要的规律，例如对于如下数据表（表 1），尽管“A B →D”的支持度仅为 2/9=0.22，但该规则很可能是有效的，因为“A 且 B”是稀有数据，且 AB 同时出现时 D 也出现的可能性大，如果最小支持度设置的太大（如 0.25），这样的规则会误认为不满足最小支持度要求而被过滤掉（生产数据中大量的存在这种情况）。因此，为了挖掘出这样的稀有数据表现的关联性，本文提出通过定义第二支持度及相对支持度来解决稀有数据的挖掘。前述的最小支持度也称为第一最小支持度，当某规则虽然不符合第一最小支持度，但仍然大于第二最小支持度（一般第二最小支持度要远比第一最小支持度小），可以进一步通过相对支持度来挖掘发现。

稀有数据项可以通过相对支持度来衡量，定义为每个单个数据项都成立时，其占各个数据项均成立时的比例最大的值。例如表 1 中，对于规则“A B → D”，ABD 同时成立的样本数为 2 个，A、B、D（单个数据项）分别成立的样本数各为 2、4、7 个，则相对支持度为 2/2、2/4、2/7 中的最大者，即 1.0，设置一个“最小相对支持度”参数值（如 0.8），当相对支持度大于该参数时，则称感兴趣的规则对于数据集是稀有的，需要进一步进行稀有数据的挖掘。因此，尽管规则“A B → D”的支持度为只有 0.22，当在挖掘关联规则时如果最小支持度设置为 0.25，该规则将会被过滤掉，但该规则符合相对支持度，相对支持度的设置为发现稀有数据中隐含的关联规则提供了条件。概括地说，第一最小支持度和最小信任度是进行关联规则挖掘的首选，如果某规则满足这两个条件，则是有效的规则；如果某些规则不满足第一最小支持度的要求，但仍然满足第二最小支持度时，则需要再通过最小相对支持度进行判别。

2 系统实现与应用

2.1 系统结构和功能

系统的总体结构采用客户-服务器（Client/Server，C/S）两层模式，在企业内部，支持在局域网内进行数据访问，在后台服务器端，是一个关系数据库系统，主要是以数据表形式存放的生产业务数据，关系数据库可以是 MS Access、MS SQL Server、Oracle 等，数据库的数据可以通过多种方式进行更新，包括手工数据导入、手工数据填写、以及生产自动化采集设备接口的自动数据存储；在前台客户端，是系统的主体功能和界面，多个客户端通过该软件，可以共享后台的生产业务数据库。前台功能采用 JAVA 程序开发，前台访问数据通过 JDBC 进行，因此后台数据库可以支持大多数流行的数据库系统。

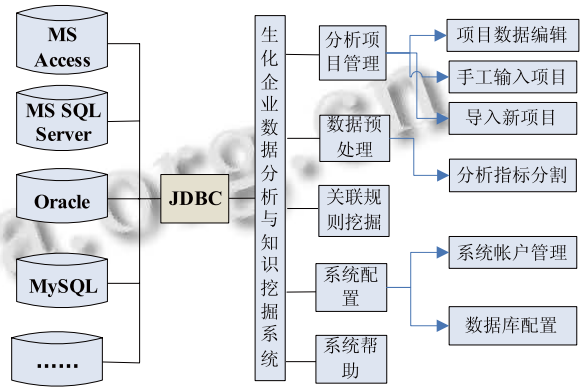

图 2 系统结构和功能模块组成

系统的功能包括分析项目管理、数据预处理、关联规则挖掘、系统配置等四个模块（图 2）及系统帮助，开发的软件所提供的数据分析与知识挖掘功能，是建立在现有生产数据基础上，除了部分自动化程度较高的数据，可通过自动化设备接口实时直接入库外，在分析项目管理模块中，还提供了包括导入新分析项目、单条数据的手工输入（删除、修改）、分析项目数据编辑的功能，为整个系统建立其它非实时数据；数据预处理模块对入库的数据进行预处理，将原始分析项目中包含的多维分析指标进行分割，形成分析数据项，进而可以通过关联规则挖掘模块，对分割后的数据项间的关联性进行挖掘，形成规则集。系统配置模块中，提供了系统用户管理和数据库配置管理两个内容，其中系统用户管理为系统管理员提供了创建新用户、删除用户、修改密码等功能，数据库配置管理可以配置后台数据库的类型和连接，从而使系统不依赖数据库类型。

2.2 系统应用

本系统的研制已初步应用于公司业务生产数据的分析中，数据分析与知识挖掘的实现主要包括以下 3 个步骤：1）分析项目（数据）的建立，2）数据预处理，3）因果关联规则发现。由于分析数据集中经常出现稀有数据，因此，系统提供同时挖掘稀有数据功能。

我们通过第二最小支持度来过滤稀有数据，另外还要设置一个最小相对支持度阈值。

第一步：分析项目的建立。分析项目建立模块中的生产数据导入，提供对企业内基于 EXCEL 数据模板的分析项目数据入库功能，入库过程中对分析项目指标的类型和数值合法性进行全面检测，就分析指标的类型而言，分为分类型和数值型两种，在数据检测中，对数值对应的指标类型（分类型和数值型）的合法性同时检测，例如若某个指标为数值型，则数据记录中，该指标出现了字符型数值时，则会自动检测出数值不合法，提醒用户对该数值进行检查。系统业提供了对感兴趣的分析指标进行导入选择的功能，对于没有意义或不感兴趣的指标，可以舍弃导入。

第二步：分析项目数据预处理。针对所建立的分析项目，将项目中的所有多维分析指标进行分割处理，形成取值为 0 或 1 的分割数据项，在指标分割过程中，依据分析指标的类型是分类型还是数值型（图 3），系统提供了不同的处理策略。对于分类型指标，根据实际的取值类别数，系统自动分割成与类别数相同的数据项；而对于数值型指标，可以直接在分割条件栏里输入以“;”为分隔符的连续区间，也可以通过自动分割栏中的设置按钮，在弹出对话框中进行等值分割和等样本分割，前者等值分割是依据分割的数据项数，按等值区间进行数据项的划分，后者则是以划分区间内的样本数相同为原则。

图 3　分析项目数据预处理（指标分割）

第三步：关联规则挖掘模块中的数据关联分析。采用了向导式的数据挖掘流程，引导用户完成数据中所隐含的关联规则。用户从分析项目集中选择一个分析项目，并选择参与分析的分割数据项（包括决策数据项），例如在柠檬酸生产中，最后产品中高的总酸量（总酸_1）和高的原料转化率（转化率_1）就是两个主要的决策指标，因此应该选中（图 4）。再将所选择的数据项进行分类，例如将前面选择的高的总酸量（总酸_1）和高的原料转化率（转化率_1）作为复合决策项，同时将生产柠檬酸的各类环境因子选中作为条件项。在此步骤中，需进一步设置最小支持度和最小信任度阈值，为了获取稀有数据中可能存在的关联规则，需要将“同时挖掘稀有数据”项选中，并设置第二最小支持度、最小相对支持度阈值。通过设计的算法[4]，进行多线程计算分析，得到所选分析项目的数据项间的关联关系，反映条件项对决策项的决定因素。

图 4　分析项目数据项选择

对企业实际生产数据，运行分析了柠檬酸生产中的数据规律，系统运行的结果如图 5，显示了关联规则挖掘的基本参数设置、各数据项的具体含义以及得到的不同项目集的规则。结果表明多个生产环境（条件）都可以产生较高的决策指标（即决策数据项，高的总酸量和转化率），同时指标间还存在交互关系，将每个数据项通过指标分割的反过程，还原为原始的分析指标，可发现生产环境（条件）和决策目标间的依存关系，例如对于 2-项目集的结果，可以发现在所有玉米浆原料类型中，只有“类型 4”（玉米浆_4）和“类型 7”（玉米浆_7）可产生高的生产率，而配方方法中，只有“方法 5”（配方_5）和“方法 9”（配方_9）可以产生较好的生产率，在 3-项目集中，另外一种原料（由于数据的保密性，用玉米浆_11 表示），类型 4 和 20 分别与 2-项目集中的两个条件项配方_5 和玉米浆_4 一起能够产生较好的生产率，显示出不同指标间的交

互作用。

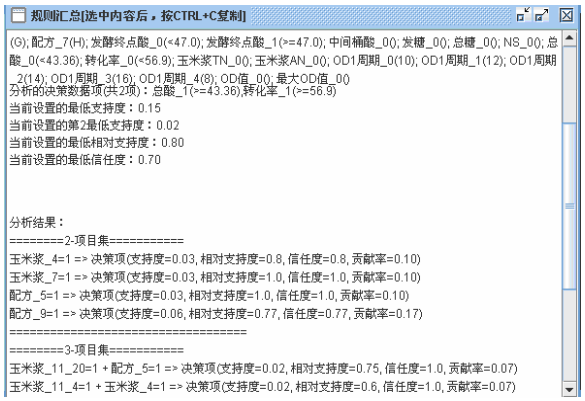

图 5 数据项关联规则挖掘结果

3 结语

对关联规则知识发现的研究一直是数据挖掘研究的重点领域，其市场分析、客户心理分析及环境质量评价中具有较广泛的应用[5-7]，而用关联规则挖掘来分析生物化工生产数据、并从中提取规律的研究现有资料较少[4-8]。本文以生物化工生产中的共性特征出发，研究并完成一个生化企业生产数据关联规则挖掘的应用系统，并对企业生产数据的现有数据集进行了初步试验和测试，取得较满意的效果。

本文设计和开发的生化企业关联规则挖掘分析系统，可以完成分析项目建立、数据预处理和数据挖掘算法实现等过程，特别是在现有数据中存在稀有数据项时，提出了基于相对支持度的解决策略，较好地满足了企业的实际数据情况，所设计开发的软件系统，也可以应用于类似的生产企业数据分析中。

参考文献

1 曹月芹,林枫,陈国浪.基于 Apriori 分类事务库关联规则算法.计算机系统应用,2010,19(8):62–65.
2 陈申燕,曹旻.多层关联规则挖掘算法的研究及应用.计算机工程与设计,2010,4:885–888.
3 殷剑锋,徐建城,李伟强.改进 Apriori 挖掘算法的网格实现.计算机仿真,2010,2:145–148,268.
4 周永生,熊结青,沙宗尧.基于关联规则挖掘的生化企业数据分析及其应用研究.计算机与应用化学,2010,27(9):1252–1256.
5 管乐,王纯.多维关联规则挖掘在彩铃推荐中的应用.计算机系统应用,2009,18(4):155–157.
6 章杰鑫,张烈平.基于时序关联规则的商品需求预测.计算机工程,2009,35(22):65–67.
7 Chen CY, Shyue SW, Chang CJ. Association rule mining for evaluation of regional environments: case study of dapeng bay, taiwan. International Journal of Innovative Computing Information and Control, 2010,6(8):3425–3436.
8 张泉灵,金晓明,荣冈,苏宏业.化工生产过程数据挖掘系统的研究与应用.计算机与应用化学,2008,25(7):769–772.

(上接第 42 页)

的指导下，增加服务台数目（大于 5）时，系统处理能力表现更加优秀。

本文的 SMSDSM 设计方案对大中型企业的短信系统具有一定的参考和借鉴价值；另外，本文实际应用中的 SMSDSM 使用 java 实现，消息队列采用 EJB 组件规范中的 MDB 来支持，可以方便地被集成到不同操作系统的平台中使用。

参考文献

1 沈斌,李兴国,钟金宏,沈丽娜.基于多队列和多线程的短信实时并发控制算法.计算机工程,2008,34(8):62–65.
2 Wang ZT, Guo ZY, Wang ZQ. Analysis of Multi-task Scheduling Based on SMS. 2008 International Conference on Computer Science and Software Engineering. IEEE. 2008. 1087–1089.
3 归敏丹,蒋毅飞,张志敏,吴锡生.多服务员时两种等待队列性能的比较.计算机工程与应用,2008,44(13):44–46.
4 唐应辉,唐小我.排队论基础与分析技术.北京:科学出版社,2006.50–57.
5 曾东海,刘海,金士尧.集群负载调度算法性能评价.计算机工程,2006,32(11):78–79.
6 Meng QY, Qiao JZ. A Dynamic Load Balancing Method Based on Stability Analysis. 2008 International Symposium on Computer Science and Computational Technology. IEEE. 2008.404–408.