基于改进 Apriori 算法的纺纱生产质量预测研究

邢鹏程[1,2]，曾献辉[1,2]

（1.东华大学 信息科学与技术学院，上海 201620；
2.数字化纺织服装技术教育部工程研究中心，上海 201620）

摘　要：随着工业大数据时代的到来，纺织企业正加速向智能制造进行产业转型升级。以提高纺织品质量预测准确度为研究目标，在基于关联规则 Apriori 算法及引入兴趣度的 I_Apriori 算法的纺纱生产质量预测模型基础上，针对 Apriori 算法效率低、时间复杂度大、不精确的缺点，提出了一种基于遗传算法的全局优化策略，对 Apriori 算法进行了改进和优化。通过对纺纱厂现场数据的试验和分析，对 Apriori 算法、I_Apriori 算法和优化算法效果进行了对比，结果显示优化算法的处理效率更高、规则挖掘更准确，对预测效果有显著提升。

关键词：纺纱生产；质量预测；Apriori 算法；遗传算法

中图分类号：TS104.1　　**文献标识码**：A　　**文章编号**：1673－0356(2017)12－0019－04

0　前言

质量预测作为一种质量控制的高级手段，是纺织生产中重要的环节之一，准确的纺织质量预测可以大幅度地降低成本、提高生产效率。传统的质量预测方式主要是凭借技术人员的经验来判断配棉方案的可行性，缺乏对产品质量和原棉性能指标之间关系的细致研究，因此经常会出现产品指标要求不符合、产品质量波动、成本增高等问题，浪费了很多宝贵的时间和资源[1]。

近年来，我国有关部门和纺织企业加大了对纺织品质量预测的重视和研究，国家经贸委和科技部都设立了相关项目资助该方面的工作[2]。目前智能化质量预测系统多采用人工神经网络技术来实现分类、回归与预测预报等，例如提出了神经网络预测模型，预测纱线质量指标[3]。

随着智能质量预测控制研究的不断深入，许多优秀的智能算法被提出来应用于纺织领域，Apriori 智能算法对企业的质量预测和生产决策产生了良好的效果[4]，但在实际应用中，尤其在处理大数据的情况下，Apriori 算法仍有很多缺陷：(1)扫描频繁项集时会产生大量的候选项集，并且在剪枝过程中需要计算出每个候选项集的所有子集并判断它们是否是频繁的，因此算法的时间复杂度过大，同时候选项子集的重复组合增加了计算时间；(2)在计算候选项集的支持度时，需要多次重新遍历数据库。基于纺织企业的海量数据规模，Apriori 算法的处理效率会大大降低，系统的 I/O 负载也会增大；(3)仅通过设置支持度和置信度来寻找关联规则并不能保证数据的完整挖掘[5]，最终得到的一些强关联规则会与实际情况不符，无法满足纺织企业智能制造的技术要求。

针对 Apriori 算法的不足，本文提出了一种基于遗传算法的全局优化算法，在优化了修剪频繁策略的基础上，引入遗传算法来避免穷举搜索和搜索过程中的局部最优解，对全局的搜索过程进行了优化。将传统的 Apriori 算法、引入兴趣度的 I_Apriori 算法以及全局优化算法应用到纺织大数据中并进行了对比，试验显示了改进算法在大数据处理上的优越性，提高了效率的同时有效地提取了最有价值的关联规则。

1　Apriori 算法和 I_Apriori 算法

1.1　Apriori 算法

Apriori 算法的原理是利用频繁项集性质的先验知识，通过逐层搜索的迭代方式，来基于 k 项集搜索 $k+1$ 项集直至穷尽数据集中的所有频繁项集[6]，再根据置信度阈值从频繁项集中产生关联规则。Apriori 算法计算频繁项集可分为两步：连接和剪枝。

1.1.1　连接

通过频繁 k- 项集 L_k 与自身进行连接来产生候选 $(k+1)$- 项集 C_{k+1}。连接规则如下：

频繁 k- 项集 L_k 的任意两个子集 l_a、l_b 可以连接的条件是：若它们的前 $k-1$ 项相同，则可连接。即

$$(l_a[1]=l_b[1]) \wedge (l_a[2]=l_b[2]) \wedge \cdots \wedge (l_a[k-1]=l_b[k-1])$$

则 l_a、l_b 连接产生的结果是：

收稿日期：2017-09-19

作者简介：邢鹏程(1993-)，男，硕士研究生，主要研究方向为数据库应用技术、大数据分析，E-mail：491472180@qq.com。

$l_a[1]l_a[2]\cdots l_a[k-1]l_a[k]l_b[k]$

1.1.2 剪枝

剪枝主要分为两个部分：

(1)依据 Apriori“任一频繁项集的所有非空子集也必须是频繁的”的性质，对候选 k- 项集 C_k 的所有项集进行扫描求出它们的($k-1$)- 项子集，并判断这些子集是否是频繁的；

(2)扫描数据库，求出候选 k- 项集 C_k 的每个候选项集的支持度，并与支持度阈值进行比较，删除小于支持度阈值的项集，得到最终的频繁 k- 项集 L_k 。

1.2 I_Apriori 算法

I_Apriori 算法是针对 Apriori 算法容易忽视规则负相关性的缺点而产生的，例如某条强关联规则 $A\Rightarrow B$ 满足可信度阈值，但其负相关规则置信度同样也很大，导致此问题的原因可能是项集间存在负相关的抑制作用，或者项集之间相互独立，因此此规则是相互矛盾的，是错误的。基于上述问题，提出一个兴趣度模型：

$$\begin{aligned}\text{interest}(A\Rightarrow B)&=Conf(A\Rightarrow B)-Conf(\overline{A}-B)\\&=\frac{P(AB)}{P(A)}-\frac{P(\overline{A}B)}{P(\overline{A})}\\&=\frac{P(AB)-P(AB)P(A)-P(\overline{A}B)P(A)}{P(A)[1-P(A)]}\\&=\frac{P(AB)-P(A)P(B)}{P(A)[1-P(A)]}\end{aligned}\tag{1}$$

此兴趣度的范围是[-1,1]，即当 $\text{interest}(A\Rightarrow B)>0$ 时，A 对 B 是促进作用，且当兴趣度越接近 1，则 A 和 B 的关联性越强；当 $\text{interest}(A\Rightarrow B)<0$ 时，A 对 B 是抑制作用，且当兴趣度越接近 -1，则 $\overline{A}$ 和 B 的关联性越强，可以看出 $A\Rightarrow B$ 规则的负关联规则 $\overline{A}\Rightarrow B$ 并没有被忽视，所以此规则是相互矛盾的，可删除此规则；当 $\text{interest}(A\Rightarrow B)=0$ 时，A 与 B 独立不相关。

I_Apriori 算法有效地去除错误的强关联规则，但和 Apriori 算法一样，在寻找频繁项集时仍要对数据库进行规模较大的遍历，同时也无法避免大量的候选项子集的重复组合。本文针对上述缺点提出了基于遗传算法的全局优化算法。

2 基于遗传算法的 Apriori 全局优化算法

此算法主要从两个方面进行优化：一是在执行连接剪枝步骤之前对频繁项集进行修剪，以减少候选项集数目，提高效率；二是引入遗传算法对寻找频繁项集的搜索过程进行全局优化。

2.1 修剪频繁项集策略

在生成频繁 k- 项集后，利用 Apriori 算法的性质 2，来简化执行连接剪枝步骤所要用的项集数量，从而减少了连接产生候选 $(k+1)$- 项集的过程中重复组合的数量级，优化了执行时间，预计采用改进后的 Apriori 算法可以使扫描次数减少一半[7]。

性质 2 为某元素要成为频繁 k- 项集的一元素，该元素在频繁 k- 项集中的出现次数必须不小于 k 次，否则包含此元素的项集不能产生候选($k+1$)- 项集[8]。

2.2 遗传算法全局优化策略

Apriori 算法的核心问题是如何找到频繁项集，利用遗传算法对此全局搜索过程进行全局优化，可以大幅度地提高 Apriori 算法的效率。

根据实际问题，先对纺织生产中的数据进行编码，例如，影响棉纱线单强的因素有唛头、技等、回潮率等物理属性，由于设备采集的此类数据类型属于非布尔型数据，且每个属性的取值是连续的、不固定的，因此需要根据历史数据和实际情况对属性取值划分为区间，并将这些区间定义为“值 1”、“值 2”，…，“值 n ”。如表 1 所示。

利用适应度函数来评价个体的优劣，并决定此个体是否可以进入下一代，因此定义适应度函数是算法的关键。衡量项集是否频繁的依据是此项集的支持度，因此根据支持度来定义适应度函数。

一般来说适应度函数定义为：

$$Fit(X)=\frac{Supp(X)}{MinSupp}\tag{2}$$

其中，$Supp(X)$ 代表当前项集 X 的支持度，$MinSupp$ 代表最小支持度阈值。

但由于算法目的是寻找单强属性最好的关联规则，所以我们更需要单强属性较大的规则；同时，对于单强属性很低的关联规则同样有价值，因为我们可以通过此规则知道造成单强属性低的因素，从而避免它。因此，在公式(2)的基础上，通过设定单强属性的权值来重新定义适应度函数：

$$Fit(X)=\frac{Supp(X_1,X_1,\cdots,X_{n-1})+W_{1\sim6}}{MinSupp+W_{1\sim6}}\tag{3}$$

其中，$Supp(X_1,X_2,\cdots,X_{n-1})$ 是除单强属性其他所有属性组合的支持度，$W_{1\sim6}$ 是单强属性的权值，其中单强属性数据设置为 6 个区间，单强属性越高或者越低的权值越大，单强属性趋于中间值的权值最低，如表 2 所示。

表 1　棉线物理属性的数据区间编码表

数组元素	A[1]	唛头	A[2]	回潮率/%
项		I_1		I_2
取值情况	值 1	2.00～2.20	值 1	7.0～7.6
	值 2	2.21～2.40	值 2	7.7～8.2
	值 3	2.41～2.60	值 3	8.3～8.8
	值 4	2.61～2.80	值 4	8.9～9.4
	值 5	2.81～3.00	值 5	9.5～10.0
	0	无	0	无

表 2　单强属性区间及权值分配表

	A[n]	单　强	权　值
取值情况	值 1	12.0～12.7	4
	值 2	12.8～13.5	2
	值 3	13.6～14.3	1
	值 4	14.4～15.1	1
	值 5	15.2～15.9	2
	值 6	16.0～16.7	4

根据适应度函数的大小，进行选择、交叉、变异的遗传操作，产生下一代规则；经过反复迭代以后，直到满足终止条件，得到一组规则；最后利用置信度和兴趣度对产生的所有规则进行筛选和提取。流程如图 1 所示。

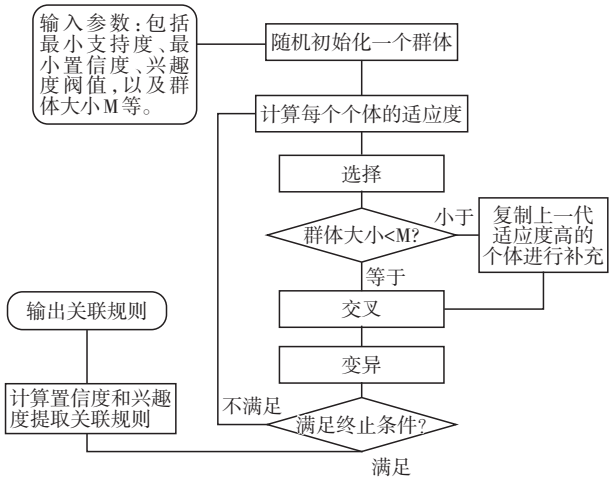

图 1　优化算法的运作流程

3　对比和分析

以纺纱生产中对纺织产品的质量预测为研究目标，根据纺织工艺特点，确立了可通过棉纤维各项性能指标来定量地预测成纱质量[9]。通过 Apriori 算法、I_Apriori 算法以及基于遗传算法的全局优化算法对棉纤维性能数据的试验和仿真，分析 3 种算法的预测性能，从而得出优化算法的优越性。

试验使用的是浙江某棉纺企业提供的现场数据，共 1 500 条数据组成了棉纺纱线单强试验训练样本，该组数据是在转杯纺快速纺纱系统中普梳 18.2 tex 纱线所取得[9]，其中截取的部分数据如图 2 所示。

唛头	技等	回潮率	强力	主体长度	均匀度	品质长度	马克隆值	含杂	成熟系数	短绒	断裂强度	单强
2.32	2.356	8.2	3.45	29.52	1102	32.62	4.61	2.5	1.54	10.66	20.23	12.6
2.01	2.562	8.9	3.78	29.36	1040	32.81	4.71	1.4	1.52	11.25	18.9	12.9
2.51	2.895	9	3.87	29.89	1134	33.01	4.59	1.9	1.62	11.89	19.56	13.6
2.01	2.452	7.1	3.56	30.38	1042	33.25	4.49	2.2	1.49	11.12	19.42	15.9
2.81	2.829	8.2	3.46	30.45	1013	32.45	4.56	2.4	1.6	12.23	19.23	16.5
2.46	2.459	8.6	3.25	29.56	1060	32.96	4.62	2.3	1.62	12.01	19.12	16.3
2.61	2.798	8.4	3.62	29.89	1134	31.95	4.66	2.3	1.52	12.62	18.23	14.7
2.32	2.789	9.6	3.68	29.05	1117	31.5	4.59	2.4	1.59	11.52	18.56	14.9
2.51	2.569	9.1	3.87	28.93	1076	31.45	4.62	2.4	1.57	11.45	18.95	15.4
2.22	2.989	8.1	3.6	30.54	1144	32.85	4.56	2.1	1.49	11.62	19.02	14.9
2.41	2.229	7.6	3.59	30.12	1029	30.12	4.59	2.5	1.62	11.89	20.01	16.1
2.36	2.829	8.4	3.34	29.89	988	30.89	4.69	2.5	1.52	10.89	19.45	13.6
2.58	2.346	9.1	3.56	29.31	1006	31.56	4.71	2.3	1.59	11.45	18.95	14.2
2.56	2.156	7.9	3.5	30.12	1027	32.45	4.65	2.2	1.62	11.62	19.45	15.3
2.61	2.874	8.1	3.72	29.58	1134	33.04	4.72	2.2	1.49	12.05	19.78	14.2
2.74	2.529	8.6	3.3	29.78	1086	31.85	4.62	2.1	1.56	12.42	18.45	15.2
2.52	2.529	7.1	3.34	29.44	1060	33.06	4.56	2.1	1.61	11.56	19.99	14.1
2.55	2.529	9	3.45	29.8	1061	32.68	4.59	2	1.62	11.78	20.21	14.6
2.52	2.365	8.8	3.65	30.38	1013	32.98	4.5	1.9	1.67	11.89	20.01	15.9
2.12	2.985	8.7	3.95	30.06	1032	32.45	4.43	2.4	1.71	11.2	19.25	16.2
2.83	2.456	8.2	3.12	30.22	1048	33.2	4.58	2.2	1.59	11.23	20.23	15.7
2.49	2.452	8.3	3.23	29.99	1077	33.12	4.44	2.3	1.54	10.98	19.45	12.9
2.41	2.874	7.8	3.55	29.41	1106	32.65	4.72	2.1	1.64	10.9	19.87	13.6
2.23	2.145	9.1	3.45	30.15	987	32.45	4.67	1.8	1.71	11.32	19.24	13.9

图 2　棉纺纱线单强实验数据

通过测试样本来检验 3 个算法的预测性能。测试规则如下：试验遵循单一变量原则，在相同支持度和置信度条件下，分别通过 3 种算法得到了单强数据的预测值，并与真实值一同记录下来。其中，每组试验的测试样本是从数据库中随机选取 5 条数据（每条数据由唛头、技等、回潮率等 12 项组成），共做 10 组试验。

将每次的试验结果做成横坐标为真实值、纵坐标为预测值的散点图上，如图 3 所示，其中图中的斜线表示真实值和预测值绝对相等的轨迹。

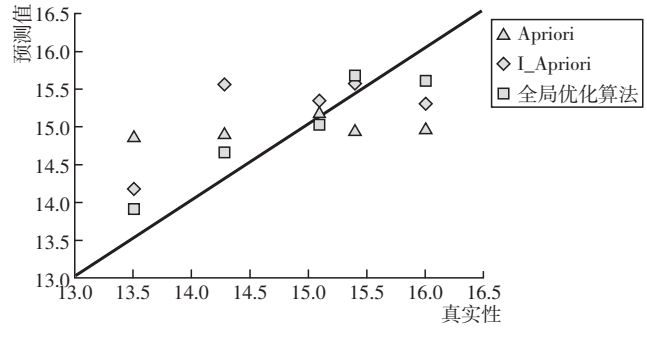

图 3　测试结果

对 10 组试验结果进行单因素方差分析，从数学的角度分析不同的算法是否对预测效果产生了显著影响。利用均方差计算公式：

$$\sigma=\sqrt{\frac{1}{5}\sum_{i=1}^{5}(x_{预}-x_{真})^2} \tag{4}$$

即预测值和真实值之差的算术平均数的平方根，它可以反应某算法得出的预测值距离真实值的离散度，是表示预测精确度的重要指标。

通过 10 组试验的测试分析，最后得到的结果是：Apriori 算法的均方差为 3.131，I_Apriori 算法的均方差为 2.862，基于遗传算法的优化算法均方差为 1.11。因此可以看出，本文提出的全局优化算法在准确度上大大优于传统的 Apriori 算法和 I_Apriori 算法，传统的 Apriori 算法并不能满足实际生产需要，它的预测值离散度太大，不能很好地预测纺纱品的质量；I_Apriori 算法在一定程度上修正了传统算法，其预测值的均方差降低，但预测效果仍然差强人意；对比 Apriori 算法和 I_Apriori 算法，全局优化算法的预测效果得到了显著的提升，预测结果较为理想，因此其性能远优于传统算法。

4 结语

在纺织智能制造中常用的两种关联规则数据挖掘算法 Apriori 算法和 I_Apriori 算法，针对它们的不足提出了一种基于遗传算法的全局优化 Apriori 算法，并通过对棉纺质量数据的试验对比分析，证明此算法有效地弥补了 Apriori 算法的不足。

未来通过对算法的进一步完善可应用到大数据上，由于纺织企业数据库规模较大，因此扫描和比较时间的缩减将会更加明显，大幅度的优化了算法的效率，满足了纺织企业的生产要求。

参考文献：

[1] 吴军辉.纺织加工质量预测技术的研究与应用[D].上海：东华大学，2009.

[2] 王侃枫.基于计算智能的精梳毛纱质量预测[D].上海：东华大学，2009.

[3] 孙海兰.纱线质量分析与预测[D].苏州：苏州大学，2004.

[4] 王达明.基于云计算与医疗大数据的 Apriori 算法的优化研究[D].北京：北京邮电大学，2015.

[5] 徐章艳，刘美玲，张师超，等.Apriori 算法的三种优化方法[J].计算机工程与应用，2004，6(36)：190－193.

[6] 陈　东.基于 Apriori 算法的大数据相关性分析研究[D].北京：中国地质大学(北京)，2016.

[7] 欧阳桃红.一种基于遗传算法的关联规则改进算法[J].杭州电子科技大学学报(自然科学版)，2015，9(5)：79－81.

[8] 肖冬荣，杨　磊.基于遗传算法的关联规则数据挖掘[J].通信技术，2010，1(43)：205－207.

[9] 李利强.支持精益生产的数据挖掘技术的研究与应用[D].上海：东华大学，2010.

Spinning Production Quality Prediction based on Improved Apriori Algorithm

XING Peng-cheng[1,2], ZENG Xian-hui[1,2]

(1. School of Information Science and Technology, Donghua University, Shanghai 201620, China;
2. Education Engineering Center of Digital Textile Technology, Shanghai 201620, China)

Abstract: With the advent of the era of industrial big data, textile enterprises are accelerating transformation and upgrading to intelligent manufacturing industry. With quality prediction as the research object, intelligent textile quality prediction model based on the association rules Aproiori algorithm and I_Apriori algorithm with interest degree was presented. At the same time, aiming at the shortcomings of poor efficiency, large complexity of time and imprecise, a global optimization strategy based on genetic algorithm was proposed, and Apriori algorithm was improved and optimized. Through the experiment and analysis, the Apriori algorithm, I_Apriori algorithm and optimization algorithm were compared. The results showed that the improved algorithm was more high efficiency and precise in dealing with big data, the prediction effect had been improved significantly.

Key words: spinning production; quality prediction; Aproiori algorithm; genetic algorithm