Apriori算法在通信岗位培训需求下的研究

肖静静，宋哲超

（福建总队综合信息保障中心，福州 350003）

摘　要： 为提高通信岗位培训需求分析的科学性，以某省公司通信岗位人员数据为基础，采用Apriori关联规则算法开展培训需求分析。首先对原始数据进行清理、集成与转换，获得66条有效数据；其次利用Weka平台构建关联规则挖掘模型，并通过支持度和置信度参数设置筛选有效规则。结果表明，在最小支持度为0.1、最小置信度为0.6时，共获得10条关联规则。文章揭示了岗位、学历、技术等级和培训级别之间的关联特征，该结果可用于识别不同岗位人员的培训需求差异，提高培训计划制定的针对性。

关键词： 数据挖掘；关联规则挖掘算法；培训需求

中图分类号： TP311　　**文献标识码：** A

Research on Apriori algorithm for training demand analysis in communication position

XIAO Jingjing，SONG Zhechao

（Integrated Information Assurance Center，Fuzhou 350003，China）

Abstract：To improve the scientific basis of training demand analysis for communication positions，personnel data from communication positions in a provincial company were used to conduct training demand analysis based on the Apriori association rule algorithm. First，the raw data were cleaned，integrated，and transformed，yielding 66 valid records. Next，an association rule mining model was constructed using the Weka platform，and effective rules were screened by setting support and confidence thresholds. The results show that when the minimum support was set to 0.1 and the minimum confidence to 0.6，a total of 10 association rules were obtained. The analysis reveals the association characteristics among position，educational background，technical level，and training level，which can be used to identify differences in training demands among personnel in different positions and improve the pertinence of training plan formulation.

Key words：data mining，association rule mining algorithm，training demand

1　引言

当前，通信行业技术更新速度不断加快，岗位能力要求随之发生变化，传统依赖经验判断的培训需求分析方式已难以充分适应人员能力提升和岗位发展的实际需要。因此，如何结合人员岗位、学历、技能等级及培训情况等信息，准确识别不同岗位人员的培训特征，成为提升培训计划针对性的重要问题。

近年来，数据挖掘技术已广泛应用于风险分析、设备缺陷识别及关联关系挖掘等领域。文献[1]针对网络漏洞挖掘问题，提出了基于Apriori风险数据分析的漏洞挖掘方法，并结合MapReduce架构提高算法处理效率。文献[2]通过分析电力通信设备缺陷影响因素、缺陷特征及其映射关系，提出了基于改进Apriori算法的设备缺陷关联分析方法。文献[3]采用Apriori算法对金融行业公司股东关系进行挖掘，获得了多条强关联规则。上述研究验证了Apriori算法能够从多属性数据中发现潜在关联关系，但现有方法多依赖经验判断或单一指标统计，对岗位、学历、技能等级和培训级别等多属性间关联关系的挖掘仍不充分，难以全面反映不同人员群体的培训特征。

基于此，本文以某省公司通信岗位人员数据为研究对象，在完成数据清理、集成和转换的基础上，利用Weka平台和Apriori算法挖掘岗位、学历、技能等级与培训级别之间的关联规则，并结合规则结果分析不同岗位人员的培训特征，旨在提升培训需求分析和培训计划制定的针对性。

2　数据挖掘算法

2.1　数据挖掘

数据挖掘是利用相关算法从数据库中提取目标信息，并从大量数据中发现有效、新颖且潜在有用模式的过程，也是数据库知识发现（Knowledge Discovery in

Database，KDD）的重要环节。数据挖掘常用方法主要包括分类、聚类、关联规则学习等。其中，关联规则挖掘算法主要包括 Apriori 算法和频繁模式增长算法（Frequent Pattern Growth，FP-Growth）等。

2.2　Apriori 关联规则算法

Apriori 算法主要用于挖掘海量数据集中不同数据项集之间的关联关系[4]。该算法采用逐层搜索的迭代方式，利用先验性质对候选项集进行剪枝，逐步发现满足最小支持度要求的频繁项集，从而缩小搜索范围。Apriori 算法频繁项集生成过程的伪代码如下。

```
k=1
FK={i|i∈I∧σ({i})≥N×minsup)} //发现所有的频繁 1-项集
repeat
k=k+1
Ck=apriori-gen(Fk-1)//产生候选项集
for 每个事务 t∈Tdo
C t=subset(Ck,t)//识别属于 t 的所有候选项集
for 每个候选项集 c∈C t do
σ(c)=σ(c)+1//支持度计数增 1
endfor
endfor
Fk={c|c∈Ck∧σ(c)≥×minsup)}//提取频繁 k-项集
until Fk=∅
Result=∪Fk
```

2.3　算法选择

Apriori 算法与 FP-Growth 算法在数据规模、内存条件及规则复杂性等方面存在一定差异，具体对比如表 1 所列。

表 1　关联规则挖掘算法对比

对比项	Apriori 算法	FP-Growth 算法
数据规模	适用于中小规模静态数据集	更适用于大规模数据集
内存条件	对内存要求相对较低	需要构建 FP-tree，对内存要求相对较高
数据处理方式	多次扫描数据集并生成候选项集	通过 FP-tree 压缩数据，减少数据库扫描次数
规则特征	适用于项集规模较小、规则结构相对简单的场景	更适用于频繁模式较多或长模式挖掘场景

Apriori 算法通过迭代生成候选项集，并扫描数据库验证其频繁性，适用于中小规模数据集；FP-Growth 算法通过构建 FP-tree 减少数据库遍历次数，在大规模数据集中的处理效率更高。考虑到本文数据规模较小、属性结构相对简单，Apriori 算法能够满足关联规则挖掘需求，因此实验选用 Apriori 算法。

3　实验方法及分析

3.1　实验环境

本次实验基于 Weka 3.8.3 数据挖掘开源软件开展。Weka 使用 Java 语言开发，集成回归、分类、聚类、关联规则及属性选择等数据挖掘方法[5]，能够满足本文数据的预处理、关联规则挖掘及结果分析需求。

3.2　挖掘模型

为分析通信岗位人员统计数据中不同属性之间的关联关系，构建数据挖掘模型如图 1 所示。

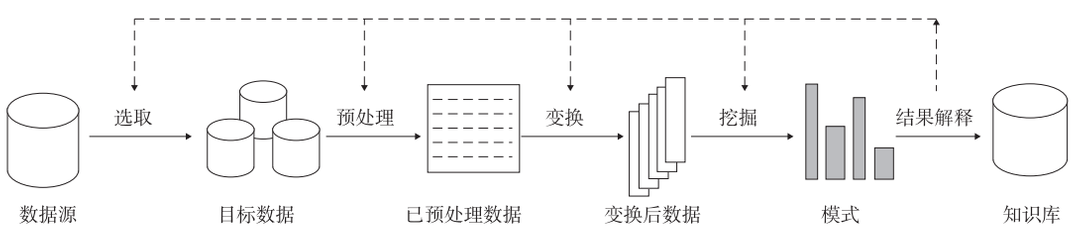

图 1　人员信息数据挖掘模型

该模型主要包括数据选取、数据预处理、数据变换、数据挖掘和结果解释 5 个步骤。（1）从岗位人员数据表中选取与研究目标相关的人员信息；（2）对原始数据进行整理，删除无效元组和无关属性；（3）对数据格式及属性值进行转换，以满足 Weka 的数据处理要求；（4）采用 Apriori 算法开展关联规则挖掘；（5）结合通信岗位实际情况对挖掘结果进行分析，识别岗位、等级、学历及培训级别之间的关联特征。

3.3　数据来源

实验数据来源于某省公司 2023 年 1 月—2025 年 12 月通信岗位人员统计数据，共获得原始样本 106 条，主要包括岗位、等级、学历和培训级别等属性。

3.4　数据预处理

数据预处理主要包括数据清理、数据集成和数据转换等步骤。（1）数据清理：删除信息不完整的无效记录，以及姓名、籍贯、毕业院校、公司名称等与关联规则挖掘目标无关的字段；任职时间、出生年月和入职日期等容易形成明显时间关联的字段不纳入后续分析。经清理后，最终获得 66 条有效记录。（2）数据集成：对整理后的数据进行规范化处理以合并形成统计表，主要保留岗位、等级、学历和培训级别 4 类属性。（3）数据转换：为便于后续数据挖掘，将“岗位”“等级”“学历”“培训级别”分别转换为“job”“level”“certificate”“train”；“等级”属性值转换为“L1”“L2”“L3”“L4”“L5”“L6”；“培训级别”中的“总公司”“市分公司”“县分公司”分别转换为“1”“2”“3”等。

3.5　数据挖掘过程

3.5.1　数据处理过程

将处理后的统计数据保存为 .csv 文件并导入 Weka，

再转换为Weka专用的.arff文件格式。Apriori算法主要处理标称型数据，而“train”属性值“1，2，3”会被识别为numeric类型，无法直接用于关联规则挖掘，因此采用Filter将其转换为nominal类型，具体处理过程如图2所示。

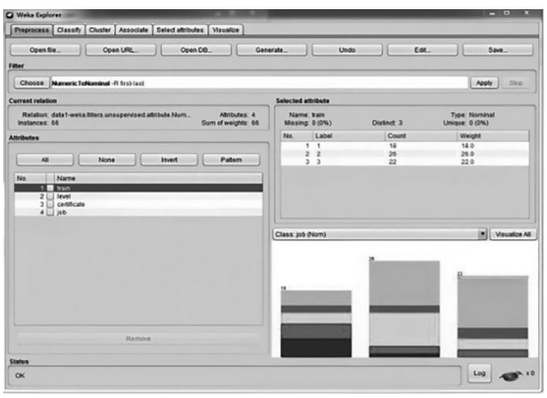

图2 数据处理过程

3.5.2 关联规则参数选择

在Apriori算法中，最小支持度和最小置信度是影响关联规则挖掘结果的关键参数。支持度反映规则在数据集中的出现频率，可用于筛除出现次数较少的规则；置信度反映规则成立的可靠程度，置信度越高，规则的可信程度通常越高。为确定合适的参数组合，分别设置不同的最小支持度和最小置信度进行实验，并统计产生的关联规则数量，结果如表2所列。

表2 置信度与支持度参数选择对比

产生关联规则数量	最小置信度值									
最小支持度值	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.1	56	56	51	38	27	10	5	3	2	0
0.2	8	8	8	8	7	1	1	1	0	0
0.3	0	0	0	0	0	0	0	0	0	0
0.4	0	0	0	0	0	0	0	0	0	0
0.5	0	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0	0	0
0.7	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0
1.0	0	0	0	0	0	0	0	0	0	0

若某关联规则仅适用于数据集中低于10%的实例，则其代表性相对有限[6-7]。本文共有66条有效记录，10%对应6.6条，即规则至少应覆盖约7条记录。由表2可知，当最小支持度为0.1、最小置信度为0.6时，共产生10条关联规则，在规则数量和置信度之间较为均衡，因此采用该参数组合，具体设置如图3所示。

3.5.3 可视化结果

采用Apriori算法进行关联规则挖掘后，Weka按照置信度由高到低输出前10条关联规则（见图4）。

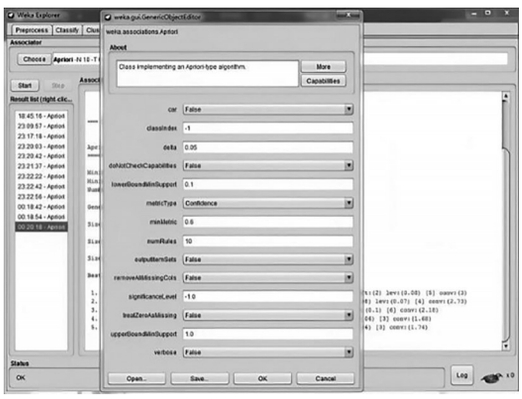

图3 关联规则参数选择

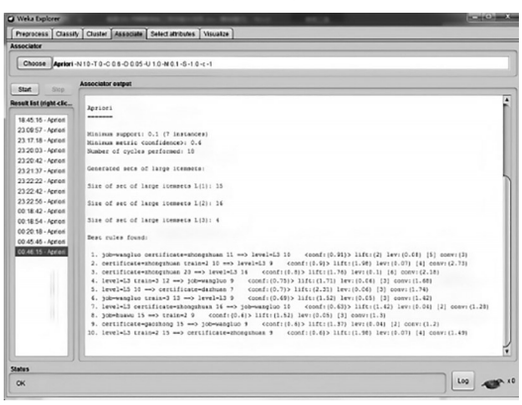

图4 数据关联分析结果

3.5.4 结果分析

实验结果显示，最小支持度为0.1，对应7个实例，最小置信度为0.6，共获得10条关联规则。

（1）关联规则1的置信度为0.91，在11个“学历”为“中专”且“岗位”为“网络维护员”的样本中，有10个样本的技术“等级”为“L3”，说明中专学历、网络维护员岗位与L3技术等级之间具有较强关联。

（2）关联规则2的置信度为0.90，在10个“培训级别”为“市分公司”且“学历”为“中专”的样本中，有9个样本的技术“等级”为“L3”，表明市分公司培训级别、中专学历与L3技术等级之间存在较明显关联。

（3）关联规则3的置信度为0.80，在20个“学历”为“中专”的样本中，有16个样本的技术“等级”为“L3”，说明中专学历与L3技术等级之间存在一定关联。

（4）关联规则4的置信度为0.75，表明县分公司培训级别、网络维护员岗位及相关技术等级之间存在一定关联。

（5）关联规则5的置信度为0.70，表明较高技术等级与大专学历之间存在一定关联。

（6）关联规则6与规则4反映的关联特征较为接近，关联规则7与规则1的结果也较为相似，进一步说明相关属性之间具有一定稳定性。

（7）关联规则8—10的置信度均为0.60，关联强度相

（下转第138页）