文章编号：2095-2163(2021)07-0124-05 中图分类号：TP399 文献标志码：A

Apriori 算法在手足口病与气象因素关系分析中的应用

降 惠[1]，尹振保[2]，武丽娟[1]，崔永梅[3]，魏 晋[1]
（1 长治医学院 计算机教学部，山西 长治 046000；2 长治市气象局，山西 长治 046000；
3 长治市疾病预防控制中心，山西 长治 046000）

摘 要：探讨长治市手足口病(HFMD)与气象因素的关联关系。利用 Apriori 关联规则算法挖掘不同气象因素组合引发手足口病的频繁程度。长治市手足口病的周高发、中等发病率与前一周的平均水汽压、气温具有显著相关性。而周低发病率与前一周的气温(最低、平均)、降水量存在显著相关性。长治市手足口病周发病率与气象因素存在一定的关联性，应根据气象条件进行流行风险评估，及时采取相应的防控措施。

关键词：手足口病；气象因素；关联规则分析

Application of Apriori algorithm in the analysis of the relationship between hand-foot-mouth disease and meteorological factors

JIANG Hui[1], YIN Zhenbao[2], WU Lijuan[1], CUI Yongmei[3], WEI Jin[1]
(1 Department of Computer Teaching, Changzhi Medical College, Changzhi Shanxi 046000, China;
2 Changzhi Meteorological Bureau, Changzhi Shanxi 046000, China;
3 Changzhi City Center for Disease Control and Prevention, Changzhi Shanxi 046000, China)

【Abstract】To investigate the correlation between HFMD and meteorological factors in Changzhi city, the Apriori association rule algorithm is used to mine the frequency of HFMD caused by different meteorological factors. The high and moderate incidence of HFMD is significantly correlated with the mean vapor pressure and temperature in the previous week.But the low weekly incidence is correlated with temperature (lowest,mean) and the rainfall in the previous week. There is a certain correlation between the weekly incidence of HFMD and meteorological factors in Changzhi City. Epidemic risk assessment should be carried out according to meteorological conditions, and corresponding prevention and control measures should be taken in time.

【Key words】hand-foot-mouth disease(HFMD); meteorological factor; association rule analysis

0 引 言

在全球气候变化的背景下，气象因素导致的健康效应备受关注[1]。许多传染病的流行都与气象因素有关[2]。手足口病(hand, foot and mouth disease,HFMD)是由肠道病毒感染引起的一种常见传染病[3]。国内外大量研究表明，气象因素会影响手足口病流行[4-5]。近年来，许多流行病学、统计学专家学者致力于研究手足口病与气象因素的关系，但传统统计分析只能揭示手足口病对气象因素的依赖程度，而关联规则分析作为数据挖掘中的一项重要技术，可以通过检验各种气象因素组合引发手足口病的频繁程度[6]，得到定量表达手足口病随气象因素变化的情况，有效简化数据处理过程。因此，本文尝试采用关联规则分析法探索气象因素对手足口病的流行影响，为预防手足口病提供借鉴和参考。

1 数据来源与预处理

1.1 研究区域概况

长治市位于山西省东南部，辖 4 区 8 县(包括潞州区、屯留区、潞城区、上党区、长子县、壶关县、平顺县、黎城县、沁县、武乡县、襄垣县、沁源县)，人口 347.8 万人，属暖温带半湿润大陆性季风气候区。

1.2 数据来源

本研究以 12 个县区的周数据作为研究单元，时间跨度为 2009～2018 年。研究中涉及手足口病数据、气象数据和人口数据三类数据。手足口病数据来自于“国家疾病监测信息管理系统”。因 2018 年长治市行政区划调整，将 2009～2017 城区与郊区手足口病周发病数合并为潞州区发病数。气象数据来源于长治市气象台(11 个国家级地面气象观测站)，共采集到十年来 11 个县区 9 种气象因素(定时风

基金项目：山西省高等学校科技创新项目(2019L0682)。

作者简介：降 惠(1983-)，女，硕士，副教授，主要研究方向：医学数据挖掘。

收稿日期：2021-04-18

速、相对湿度、降水量、最高气温、平均气温、最低气温、日照时数、平均气压与平均水汽压）的周数据。潞州区因无国家级气象观测站，气象数据根据屯留区、潞城区、上党区数据取均值进行统计分析。人口数据来自于 2010～2019 年山西统计年鉴。

1.3　数据预处理

研究中，考虑到手足口病潜伏期为 2～10 天，因此选取周发病率与前一周的 9 项气象因素建立二维关系表。其中，含有的缺失值和异常值采用行删除法或替换法处理[7]。对于有较大缺失值的观测样本采用减少样本量，即行删除法处理。因研究中涉及的数据均为数值型，所以对于样本中存在的个别缺失值和异常值，使用前后一周数据的均值进行替换。经过清洗，最终确定用于研究的数据为 63 300 个。

2　关联规则分析

2.1　关联规则与 Apriori 算法

关联规则反映一个事物（或属性）的出现对其他事物（或属性）的出现有多大的影响。关联规则分析是从大型关系数据库或事务数据库的海量数据中发现并提取频繁出现的或人们感兴趣的知识，是一种无监督学习的数据挖掘方法[8]。

在关联规则分析中，一条样本记录称为一个事务。样本的属性称为项，多个属性组成的集合称为项集，k 个属性组成的集合称为 k －项集。对于事务数据库中的一条记录，如果同时具有互不相交的 2 个子项集 A 和 B，则项集 A 和 B 是关联的，即 A －> B。A 称为前项，B 称为后项。关联规则分析可以从大量数据项集中发现频繁出现的模式和关联性。但得出的关联规则并不能直接使用，还需要根据置信度、支持度和提升度指标进行评估，从而得出具有一定参考价值的关联规则[9]。支持度是指项集 A、B 同时出现的频率，主要体现关联规则的重要性，置信度是项集 A 发生前提下 B 发生的频率，主要体现关联规则的准确性[10]。提升度是项集 A 发生前提下 B 发生的概率与 B 总体发生的概率之比。在关联规则分析中，最小支持度表示挖掘出的关联规则必须满足数据项频度的最小支持阈值，其取值影响着生成频繁项集的数量[11]。最小置信度体现关联规则的最低可靠性，其取值影响着生成强关联规则的数量[12]。

目前，常用的关联规则算法有 Apriori、FP-Tree、Eclat 和灰色关联算法。其中，Apriori 是最经典、也是最常用的挖掘频繁项集的算法。Apriori 算法采用逐次迭代的方法，通过反复扫描事务数据库，连接产生所有的频繁项集，然后根据预先设定的支持度、置信度和提升度参数，利用剪枝的方法得到感兴趣的强关联规则。本研究拟采用 Rstudio 软件，借助 arules 和 arulesViz 程序包中的相关函数实现 Apriori 关联规则分析。

2.2　数据离散化

在构建关联规则模型时，为缩小数据的覆盖范围，使数据更适应模型，匹配 Apriori 关联规则建模的格式要求，分析中首先对各数据项进行离散化分组。为保证每组中样本量的一致性，本研究利用 arules 包中的 *discretize*( ) 函数，将每个属性值分组数预设为 7[13]，按照等深分组的方法，识别出相应的阈值区间，各数据项具体分组情况见表 1。数据离散化后，将其导入到 Rstudio 中，并将其转换为“transcations”格式，建立事务数据库。

2.3　不同程度手足口病周发病率与气象因素的关联规则分析

在事务数据库中，每个样本记录包含 10 个属性，即：手足口病发病率与 9 种气象因素值。为了分析不同程度手足口病周发病率与气象因素的关联关系，分析中将前一周 9 种气象因素值作为 9 －项集 A，手足口病周发病率作为项集 B。对于任意一条记录，如果同时具有项集 A 和 B，则项集 A 和 B 是关联的，即 A －> B。

2.3.1　手足口病高发病率与气象因素的关联规则分析

本研究中将最小支持度和置信度分别设定为 0.011、0.55，共生成关联规则7 385条。为了求出频繁项集中手足口病高发病率与气象因素之间的关联关系，研究中将气象因素设置为前件，将手足口病高发病率 HFMD5 设置为后件。高发病率与气象因素的强关联规则见表 2。当提升度（*lift*）> = 3.5 时，共得到 3 条强关联规则。

表 2 结果显示，手足口病的高发病率主要有 2 种气象特征：

（1）前一周平均水汽压为 VapPres5，最低气温为 LTemp6，特别是平均气温为 MTemp6 时。

（2）前一周平均水汽压为 VapPres5，最高气温为 HTemp7。

高发病率与气象因素的强关联规则如图 1 所示。由表 2 与图 1 可以看出，手足口病的高发与平均水汽压、气温具有显著的相关性，结果与国内相关报道一致[4]。

表 1 各数据项分组表

Tab. 1 The grouped table for each data item

属性	标识组	阈值区间	属性	标识组	阈值区间
风速/(m·s^{-1})	WindSp1	[0,1.2)	相对湿度/%	RelHumid1	[16.86,41)
	WindSp2	[1.2,1.46)		RelHumid2	[41,48.8)
	WindSp3	[1.46,1.64)		RelHumid3	[48.8,56.8)
	WindSp4	[1.64,1.84)		RelHumid4	[56.8,64.1)
	WindSp5	[1.84,2.06)		RelHumid5	[64.1,71.3)
	WindSp6	[2.06,2.46)		RelHumid6	[71.3,77.4)
	WindSp7	[2.46,5.44]		RelHumid7	[77.4,97.9]
降水量/mm	Rainf1	[0,0.16)	平均气温/℃	MTemp1	[-11.2,-2.94)
	Rainf2	[0.16,0.54)		MTemp2	[-2.94,2.29)
	Rainf3	[0.54,1.39)		MTemp3	[2.29,8.91)
	Rainf4	[1.39,3.23)		MTemp4	[8.91,14.27)
	Rainf5	[3.23,37.94]		MTemp5	[14.3,18.65)
	Rainf6	-		MTemp6	[18.65,21.87)
	Rainf7	-		MTemp7	[21.87,27.1]
日照时数/h	SunDura1	[0,3.59)	最低气温/℃	LTemp1	[-18.8,-8.14)
	SunDura2	[3.59,4.91)		LTemp2	[-8.14,-2.96)
	SunDura3	[4.91,5.83)		LTemp3	[-2.96,2.89)
	SunDura4	[5.83,6.64)		LTemp4	[2.89,8.42)
	SunDura5	[6.64,7.46)		LTemp5	[8.42,13.08)
	SunDura6	[7.46,8.41)		LTemp6	[13.08,16.73)
	SunDura7	[8.41,12.53]		LTemp7	[16.73,22.81]
平均气压/hpa	AirPres1	[879.03,896.05)	最高气温/℃	HTemp1	[-3.73,4.31)
	AirPres2	[896.05,900.53)		HTemp2	[4.31,9.39)
	AirPres3	[900.53,904.19)		HTemp3	[9.39,16.38)
	AirPres4	[904.19,908.07)		HTemp4	[16.38,21.59)
	AirPres5	[908.07,911.16)		HTemp5	[21.59,25.26)
	AirPres6	[911.16,916.4)		HTemp6	[25.26,28.15)
	AirPres7	[916.4,941]		HTemp7	[28.15,35.7]
平均水汽压/hpa	VapPres1	[0.86,2.4)	HFMD 发病率/(/10 万)	HFMD1(低)	[0,0.34)
	VapPres2	[2.4,3.63)		HFMD2(中低)	[0.34,0.78)
	VapPres3	[3.63,5.7)		HFMD3(中)	[0.78,1.85)
	VapPres4	[5.7,9.08)		HFMD4(中高)	[1.85,4)
	VapPres5	[9.08,12.65)		HFMD5(高)	[4,70.6]
	VapPres6	[12.65,17.81)		HFMD6	-
	VapPres7	[17.81,28.51]		HFMD7	-

表 2 高发病率与气象因素的强关联规则

Tab. 2 Strong association rules between high incidence and meteorological factors

编号	规则	支持度/%	置信度/%	提升度
1	$\{E_3 = \text{MTemp6}, E_5 = \text{VapPres5}, E_8 = \text{LTemp6}\} => E_{10} = \text{HFMD5}\}$	1.2	63.1	4.4
2	$\{E_5 = \text{VapPres5}, E_8 = \text{LTemp6}\} => \{E_{10} = \text{HFMD5}\}$	1.6	58.8	4.1
3	$\{E_5 = \text{VapPres5}, E_9 = \text{HTemp7}\} => \{E_{10} = \text{HFMD5}\}$	1.1	57.3	4.0

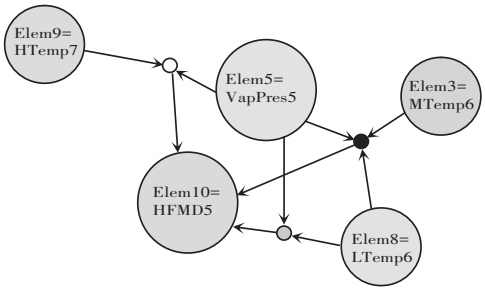

图 1　高发病率与气象因素的强关联规则图

Fig. 1　Strong association rules between high incidence and meteorological factors

2.3.2　手足口病中等发病率与气象因素的关联规则分析

为了探讨手足口病中等发病率与气象因素的关联关系，将中高发病率 HFMD4、中发病率 HFMD3、中低发病率 HFMD2 作为后件，将气象因素作为前件，将最小支持度、置信度分别设置为 0.02 和 0.2，共生成关联规则 3 404 条。中等发病率与气象因素的强关联规则见表 3。当提升度（*lift*）> = 2 时，生成 5 条强关联规则。在生成的强关联规则中，后件均为 HFMD4，说明中高发病率与气象因素的关联性更强。

表 3　中等发病率与气象因素的强关联规则

Tab. 3　Strong association rules between moderate incidence and meteorological factors

编号	规则	支持度/%	置信度/%	提升度
1	$\{E_5 = \text{VapPres7}, E_9 = \text{HTemp7}\} => \{E_{10} = \text{HFMD4}\}$	2.4	29.9	2.1
2	$\{E_5 = \text{VapPres7}, E_8 = \text{LTemp7}\} => \{E_{10} = \text{HFMD4}\}$	3.6	29.5	2.1
3	$\{E_5 = \text{VapPres7}, E_3 = \text{MTemp7}, E_9 = \text{HTemp7}\} => \{E_{10} = \text{HFMD4}\}$	2.3	29.4	2.1
4	$\{E_5 = \text{VapPres7}\} => \{E_{10} = \text{HFMD4}\}$	4.2	29.9	2.0
5	$\{E_5 = \text{VapPres7}, E_3 = \text{MTemp7}\} => \{E_{10} = \text{HFMD4}\}$	2.9	28.8	2.0

中等发病率与气象因素的强关联规则如图 2 所示。由表 3 和图 2 可以看出，HFMD 中等程度的发病率与前一周平均水汽压、气温（最高、最低、平均）均具有显著的相关性，当前一周平均水汽压、气温（最高、最低、平均）位于最高区间时，会造成手足口病中等程度的流行。

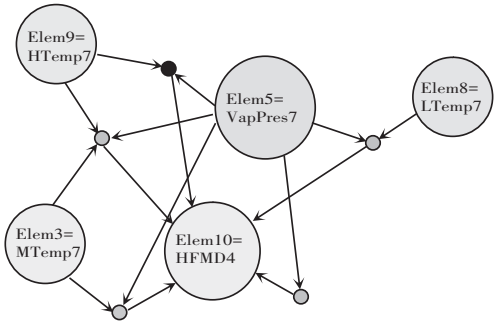

图 2　中等发病率与气象因素的强关联规则

Fig. 2　Strong association rules between moderate incidence and meteorological factors

2.3.3　手足口病低发病率与气象因素的关联规则分析

为了探讨手足口病低发病率时的气象特征，研究中将气象因素设置为前件，将 HFMD1 设置为后件，最小支持度和置信度分别设定为 0.1、0.8，共生成关联规则 25 条。低发病率与气象因素的强关联规则见表 4。当提升度（*lift*）> = 1.9 时，得到 3 条强关联规则。

低发病率与气象因素的强关联规则如图 3 所示。由表 4 与图 3 可以看出，手足口病的低发与最低气温、平均气温、降水量存在显著的相关关系，当最低气温、平均气温、降水量位于最低区间时，手足口病的发病率较低。

表 4　低发病率与气象因素的强关联规则

Tab. 4　Strong association rules between low incidence and meteorological factors

编号	规则	支持度/%	置信度/%	提升度
1	$\{E_2 = \text{Rainf1}, E_8 = \text{LTemp1}\} => \{E_{10} = \text{HFMD1}\}$	10.3	82.8	1.94
2	$\{E_8 = \text{LTemp1}\} => \{E_{10} = \text{HFMD1}\}$	11.8	82.7	1.93
3	$\{E_3 = \text{MTemp1}, E_8 = \text{LTemp1}\} => \{E_{10} = \text{HFMD1}\}$	10.0	81.9	1.92

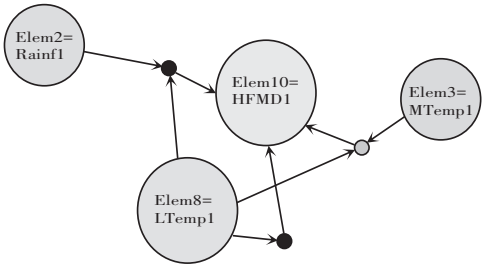

图 3 低发病率与气象因素的强关联规则图

Fig. 3 Strong association rules between low incidence and meteorological factors

3 结束语

目前,关联规则分析在医学领域的应用主要集中于中医用药规律分析、慢性病患病因素分析、上呼吸道疾病与气象因素相关性分析等。本研究采用 Apriori 关联算法分析了长治市 2009~2018 年各县(区)手足口病与气象因素的关联性。

研究结果显示,不同程度的手足口病发病率与各气象因素的关系存在一定的差异。手足口病的高发、中等发病率与前一周平均水汽压、气温(最高、最低、平均)存在显著的相关性。高发病率有 2 种气象特征:

(1)平均水汽压为中等([9.08 hpa,12.65 hpa)),最低、平均气温为次高([13.08 ℃-16.73 ℃)、[18.65 ℃-21.87 ℃))。

(2)平均水汽压为中等([9.08hpa,12.65hpa)),最高气温为最高(高于 28.15 ℃)。

这可能有 2 方面的原因:一是湿热的气象环境,适合肠道病毒的繁殖与快速传播,二是适宜的气象环境下,易感人群室外活动频率增加,感染几率增大。当平均水汽压、气温满足这 2 个条件时,HFMD 下周暴发的可能性最大,在这个时期应加大防控知识宣传力度;提醒家长少带孩子到拥挤的公共场所,不喝生水,不吃不卫生食品;加强食品和卫生监测;增加幼儿园、学校、青少年活动中心、文体中心等聚集场所的卫生清洁与消毒频次。

手足口病的低发与气温(最低、平均)、降水量存在相关性。当降水量最少、平均气温最低时,环境干燥寒冷,大部分病毒干冷而死,发病率低。

综上所述,本研究利用 Apriori 关联规则算法,通过反复扫描 2009~2018 年长治市手足口病周发病率与前一周9种气象因素建立的事务数据库,得出了频繁出现的项集,最后根据提前设置的最小置信度等参数得出强关联规则。研究结果与国内外文献报道一致[4,14-15]。但研究中以周作为时间尺度,可能不能精准地反映气象因素对手足口病的流行效应。今后,有待选择日作为研究单元,分析气象因素对不同滞后天数手足口病的流行影响,研究结果可能会更准确。此外,手足口病的发病可能受人口密度、经济条件等多种因素的影响,下一步应综合考虑这些因素,为手足口病的预防控制提供更为准确的参考依据。

参考文献

[1] 吴衍嘉,孙杨青,陆芳芳,等. 日光照射时间对 2015-2018 年深圳宝安区儿童手足口病的影响[J]. 现代预防医学,2021,48(6):1029-1033,1049.

[2] 阚海东,姜宜萱,陈仁杰. 气象因素与人群健康研究的前沿进展[J]. 山东大学学报(医学版),2018,56(8):7-13.

[3] 国家卫生健康委员会. 手足口病诊疗指南(2018 年版)[J].中国病毒病杂志,2018,8(5):347-352.

[4] DUAN Chunxiao, ZHANG Xuefeng, JIN Hui, et al.Meteorological factors and its association with hand, foot and mouth disease in Southeast and East Asia area: a meta-analysis[J]. Epidemiology and Infection,2018,147(50):1-18.

[5] NGUYEN H X, CHU G, NGUYEN H L T, et al.Temporal and spatial analysis of hand, foot, and mouth disease in relation to climate factors:A study in the Mekong Delta region, Vietnam[J]. Science of the Total Environment,2017,581/582:766-772.

[6] 王哲,李琳,王凯,等. 基于关联规则分析的慢阻肺就诊人数与气象空气条件关系研究[J]. 中国数字医学,2018,13(4):2-4,47.

[7] 张良均,云伟标,王路,等. R 语言数据分析与挖掘实战[M]. 北京:机械工业出版社,2021.

[8] 张良均,谢佳标,杨坦,等. R 语言与数据挖掘[M]. 北京:机械工业出版社,2017.

[9] 郭慧敏. 基于关联分析的中老年体检数据的挖掘[J]. 软件工程,2021,24(5):7-9.

[10]陈梦蝶. 数据驱动的慢性疾病风险因素关联分析及再入院预测研究[D]. 成都:电子科技大学,2020.

[11]李宇斐. 基于关联规则的电子病历数据挖掘应用研究-以糖尿病及其并发症为例[D]. 武汉:华中科技大学,2017.

[12]李毛琳. 空气质量与慢病关联模型研究[D]. 荆州:长江大学,2018.

[13]翟广宇,王式功,董继元,等. 兰州市上呼吸道疾病与气象条件和空气质量的关联规则分析[J]. 兰州大学学报(自然科学版),2014,50(1):66-70.

[14]杨雅斯,卢雅陵,方莅媛,等. 气象因素对四川省手足口病发病率的影响及预测模型构建[J]. 四川大学学报(医学版),2021,51(5):685-690.

[15]张翠平,张勇,刘辉,等. 安阳地区 2008-2019 年手足口病发病与气象因素的相关性分析[J]. 医学理论与实践,2021,34(8):1415-1417.