信息网络专题研究之应用层

一、文献信息

1. 论文题目：Using News Articles to Model Hepatitis A Outbreaks: A Case Study in California and Kentucky
2. 作者：MarieCharpignon，Maria Mironova，SaeyoungRho，MaimunaS.Majumder, LeoA.Celi
3. 发表途径：NeurIPS 2019
4. 发表时间：2019/12/14

二、问题意义

1. 研究背景及意义

美国最近爆发了一次大规模的疫情，接近实时的发病率估算对于政府预测有效的疫苗接种率从而控制疫情的传播至关重要。这些估计值的传统来源是疾病控制中心（CDC）提供的常规监测数据，通常会出现明显的延迟。因此需要替代方法。在本文中，基于加利福尼亚州和肯塔基州的病例，作者证明在目标疫苗接种率方面，以新闻报道数据为模型的爆发动力学产生的结果与具有 CDC 数据的爆发动力学可比。此外，作者展示了将自然语言处理（NLP）技术应用于与健康相关的新闻内容的方法如何能够提取洞察力，从而为决策者提供信息。

2. 主要研究问题

（1） 两种概念模型发爆发动力学与 CDC 数据的可比性

（2） 自然语言处理技术（NLP）应用于与健康相关的新闻内容的方法

三、思路方法

本研究重点关注加利福尼亚州(CA)和肯塔基州(KY)的疫情动态，这两个州在很大程度上代表了最高风险人群。CA 占美国无家可归人口总数的 24%，而 KY 是与药物过量相关死亡人数最多的 5 个州之一。

CDC 的监测数据集代表传统数据，而从 HealthMap 检索到的与甲型肝炎相关的新闻文章则用作非传统数据源。将发病率衰减和指数调整(IDEA)模型应用于两个数据集，并对其性能进行了比较。此外，作者使用 NLP 技术分析了新闻文章的主体，并测量了语言相似度，以便为卫生决策者提供进一步的见解。

本文数据主要有两个来源。美国疾病控制与预防中心的流行病学研究在线数据(WONDER)是一个搜索引擎，用于从美国疾病控制与预防中心选择数据集。作者提取了 2017

年 3 月 4 日至 2019 年 3 月 31 日关于甲型肝炎发病率的每周报告。这些信息由地方卫生部门自愿提交给国家法定传染病监测系统。HealthMap 是一个包含疾病相关新闻文章的公共数据库。从 2017 年 3 月 24 日到 2019 年 3 月 31 日，从全州各大媒体获取的 568 篇健康地图新闻报道中，都提到了当前甲型肝炎的爆发。仅删除在县级的报道，并收集新闻中报道的病例数，以便进行 IDEA 分析。对于来自 CA 和 KY 的每一篇新闻文章，都对其内容进行了文本分析。

IDEA 模型是一种用于短期流行病学预测的单方程模型。它取决于两个参数:基本繁殖数 R0 和折扣因子 d。R0 表示每个感染者首次进入完全易感人群时成功传播的数量，并描述暴发的初始指数增长。下面的公式显示了(I)随时间(T)的累积发生率，它是 r0 和 d 的函数。T 表示数据收集过程的开始和结束之间的时间点的数量。在此场景中，报告频率以周为单位。

$$I(\mathrm{t})=\sum_{t=0}^{T}\left(\frac{R_0}{(1+t)^t}\right)^t$$

当 R0 相当低(小于 5)时，使用参数化的 IDEA 模型是合适的，这是甲型肝炎病毒传播的情况。为了校准模型参数，作者应用了一个非线性优化过程，并使用 Python scipy.optimization.curve_fit 方法将理论表示与经验数据拟合。

然后进行文本分析。所有新闻文章都聚合在一个词包模型中，以便为文本分析做好准备。预处理的重要步骤包括句子标记化、停止单词移除和单词词形化。再使用 Spearman 进一步测量了 CA 和 KY 两个词袋模型的交集，根据次的相对频率来估计它们的相似度。

四、实验结论

两种概念模型中，美国疾病控制与预防中心的数据作为可靠的串行间隔选择。拟合优度检验表明，模型的性能并不强烈依赖于序列区间的选择。而使用新闻文章处理丢失的数据。与向疾控中心报告的病例数不同，从新闻文章数据中提取的累积发病率会受到丢失数据的影响。进位和线性平滑是两种可以处理它的技术。两种策略都得到了相似的结果。

使用 100%疫苗效力进行疫苗接种评估，估计当前（即截至 2019 年 3 月 31 日数据收集结束时间）的疫苗接种率，并与控制疫情的目标阈值进行比较。由于基于新闻文章的模型对缺失数据具有更高的敏感性，因此使用数据集对疫苗接种率的估计有很大不同。

结果证明了评估目标接种率和当前接种率之间差异的一致性。在流行病仍在蔓延、监测数据尚不可用的情况下，新闻媒体数据有可能被用作另一种数据来源。另一方面，作者注意到新闻文章的质量可能会影响结果。尽管存在这些因素，但该模型在评估估计的疫苗接种率阈值与估计的实际疫苗接种率之间的差距方面仍具有稳健性。同时，与健康相关的新闻报道质量的提高，将提升新闻媒体作为替代数据源的价值。但也存在限制和风险。在未来的工作中，新闻文章可以有效地用于甲型肝炎暴发的建模，以获得遏制其进展所需的增量疫苗接种。在理解流行病时，使用新闻媒体可以带来额外的好处。

五、启发思考

1．机器学习中的文本分析

文中提到了使用文本分析对新闻进行处理。文本分析分为预处理和进一步处理。在本论文中，预处理的重要步骤包括句子标记化、停止单词移除和单词词形化。这些步骤分别在 Python 中使用 nltk.tokenize 中的 word_tokenize 函数，nltk.corpus 提供的英语停用词列表以及 nltk.stem 中 WordNetLemmatizer 模块完成。然后使用 Spearman 进一步测量了 CA 和 KY 两个词袋模型的交集，根据次的相对频率来估计它们的相似度。

使用文本分析可以反映出相关单词与流行病的相关性，在流行病预测中具有重要作用。在其他案例的机器学习与大数据分析中也常采用文本分析，该技术可以提取出文本中的信息，对数据分析具有重要作用。

2．新闻文章质量对结果的影响

在本论文中，新闻文章的质量可能会影响结果。与使用疾病控制中心数据相比，KY 的 IDEA 模型在使用新闻文章时产生了更高的 MAPE，而 CA 的 MAPE 没有显著变化。CA 和 KY 之间的差异可以由当地新闻媒体的不同态度来解释：CA 的第一个新闻报道在宣布疫情的那一周发布，而 KY，观察到延迟了 11 周。此外，当使用新闻文章时，模型输出的疫苗接种率估计值低于使用 CDC-WONDER 数据进行 CA 时的估计值，而 KY 的情况则相反。这可能意味着在 CA news 中过度报道案例数量，而在 KY 中报道不足。因此进行数据的选择时，不同的数据集会对结果产生影响。我们可以通过选择不同数据集，或者对误差进行分析，使数据集对结果的影响降低到允许范围内。