论文

1.导论

1.1 问题背景与选题原因

……

1.2 工作总览

……

1.3 基本假设

【这里参照task1.pdf】

2 数据处理

2.1 数据分析

最好用画图或者画表格方法来展示分析结果，比如：

- 报告的分类情况（Positive, Negative, Unverified, ..）（个人感觉扇形图好）
- 报告的附件类型分布情况（jpg, mov...）（个人感觉扇形图好）
- 报告数随月份的变化情况（用条形统计图比较好）
- And so on

2.2 数据清洗

附的照片不是突变格式的不予以分析

……（具体的处理过程交给处理数据的同学写吧）

（这里要提的一点就是我把那些时间格式不规范的报告给删了（有个好像时间那一栏写的1/21/1200的，具体记不清楚了））

2.3 ……

3 模型建立

- 对于Task 1，我们将注意力主要集中在了那些评判为正确的目击报告上面，建立了一个基于元胞自动机的亚洲大黄蜂传播模型，初始数据使用的是2019年的正确的目击记录。对5年以内的大黄蜂入侵的情况进行了预测。

  【下面参考task 1.pdf】

- 对于Task 2~5 我们分析了提供的目击报告证明，选出了以下四个要素作为我们的评判依据（**这里可以详细说一下是怎么选出来这四个的**），分别是报告的月份、所附的照片、报告的地点（以经纬度形式表示）、以及报告人附的Notes。我们没有直接建立错误分类与各种因素的关系，而是选择了对报告进行可靠度评价的方式。建立了Report可靠度评价系统，对每个Report计算一个可靠度评价指标，作为评估一个Report是否为我们关注的大黄蜂目击记录的依据，它的值越大，可靠程度

也就越大，Report为正面样本的可能性也就越大，反之亦然。并依据我们建立的评价系统，逐个解决了Task 2~5里的问题。

【下面参考Task B.pdf】

4 问题解决（Problem Solving）

- Task 1

  依据我们在上面建立的传播模型，以下给出我们模型的对于不采取任何预防措施情况下五年内亚洲大黄蜂的预测结果，以热力图的形式呈现，其中颜色越红的地方，亚洲大黄蜂的种群的数量也越多。

  【这里贴一下预测结果热力图】

  [精度检验]

  我们采取2020年的数据来对数据进行预测，预测结果如下（0结果不放了，**下面的图建议用Latex画张表格**）：

	经度	纬度	数量
0	−122.9	49.1	2
1	−122.7	48.9	2
2	−123.9	49.1	1

  再来看看实际数据（**这个表格画与不画都行，看最后篇幅够不够**）：

```
956   {1C6D0EAB-F68D-411D-974E-1233618854CC}  2020-05-15 00:00:00  49.060215 -122.641648
834   {AD56E8D0-CC43-45B5-B042-94D1712322B9}  2020-05-27 00:00:00  48.955587 -122.661037
1011  {FC6E894B-F6DF-4FDC-853A-D7372D253988}  2020-06-07 00:00:00  48.777534 -122.418612
3279  {A717D86F-23E9-4C8C-9F12-198A71113E93}  2020-08-17 00:00:00  48.927519 -122.745016
4127  {2138197A-F5CF-4308-93E2-62EA6F84D098}  2020-09-21 00:00:00  48.984269 -122.574809
4338  {AA461F47-1B2B-4EA1-8154-ECF70B55A334}  2020-09-28 00:00:00   48.98422 -122.574726
4208  {DEF5D82B-E326-41A5-9B6C-D46DCD86950C}  2020-09-29 00:00:00  48.984172  -122.57472
4206  {0FAC3767-EAC4-477A-B5F0-24AF8A40BD09}  2020-09-30 00:00:00  48.979497 -122.581335
4207  {BEAC832C-0783-414A-9354-C297F38570AD}  2020-10-01 00:00:00  48.983375 -122.582465
```

  由于所选的经纬度划分粒度，我们视为经纬度差距在0.2度左右的为正确的预测。根据目击报告的数量情况，其中在（-122.9，49.1）附近的应该有1~2个亚洲大黄蜂种群，而（-122.7，48.9）附近，估计有2个以上的种群的数量。因此，大致估计模型的准确度为：

$$precision \approx \frac{num_{right}}{num_{total}} = \frac{3}{5} = 0.6$$

- Task 2

  依据我们建立的Report可信度评价系统，对原本的报告数据使用了我们的指标分析。

  【这里画个我们对原数据进行预测的前20条预测结果与原本的标签的表格，再加一个Postive ID报告的指标值与Negative ID报告的指标值的箱图】

  可以看到，通过我们的评价系统估算每个Report的指标并排序后，大部分的正向样本排在了前面，观察可以得知少部分的正向样本不在其中是因为照片等要素的缺失，处在可解释的范围内。

  回到我们的原来的问题，要衡量一个报告是错误分类的可能性，只需要看它的可靠度指标，如果指标较低，那么它是一个负面Report的概率就越大，反之如果指标较高，那么它是一个负面Report的概率就越小。

- Task 3

对于这个问题，由于人力物力有限，一个很理智的决定便是优先调查那些报告的可靠程度比较高的地区作为调查目标，如果有多个高可靠程度的地区，我们选择较为报告分布更为密集的地区作为我们的调查目标。我们只需要找到那些可靠性比较高的报告集中分布的区域，优先调查即可。我们选择正向样本指标的下四分位数α作为阈值，高于该阈值的报告我们认为其可靠性比较高。对于**无论其Status，对于所有可靠性指标高于阈值的报告**，我们做出其在地图上的分布：

【这里放一下地图上的那些点的分布情况】

具体的地区排序如下：

【这里统计各个方块内的可靠程度比较高的报告点的数量，按数量从大到小排序，排名前十的表格】

所以，我们最好优先调查XX地区。。。

- Task 4

  从上文我们的模型思路来看，我们所建立的模型很大程度上依赖于正面报告数据。但，由于正面报告数据的缺失（只有14个），相较于负面报告的加入，正面报告的数据对我们的模型影响更大。考虑数据越多越准确，模型的更新应该尽量频繁，但同时太频繁的更新会带来难以忍受的计算代价，因此，需要做到更新频度和计算代价的平衡（**这里可以祭出matlab课上的那个舍入误差和截断误差随步长变化的图像（改改坐标轴）**）。我们假定拿到的新的报告是已经经过了实验室的确认过其是不是亚洲大黄蜂的，如果不是，我们使用当前模型进行初步的确认（指标值大于α的认为是正面报告，否则认为是负面报告）。我们采取以下的更新策略：

  - 我们维护一个新报告的队列，开始的时候它是空的。
  - 如果没有新的正面报告出现，那么我们大约一个月更新一次我们的模型。模型的更新包括将新的报告图片加入到图像识别的神经网络训练集里并对神经网络进行进一步的训练，重新计算每个点到最近的k个报告点的距离并更新其RP的值，重新统计正面报告随月份的变化关系来拟合出新的曲线，以及（**重新算那个Note的影响，这里不太清楚，请你们自己改吧**）。不过值得庆幸的是模型的计算复杂度并不是太高，这最多耗费12个小时的时间。
  - 如果有新的正面性较高的报告出现，那么我们立即更新我们的模型。
  - 更新完成后，清空队列，等待下一次更新。

- Task 5

  考虑到亚洲大黄蜂的繁衍传播周期为1年，因此我们选取一年为观察的时间窗口。

  设某一个地区从某一个月开始一年以内的时间里，收到的所有报告的集合为S_i，该月出现正面报告的可能性用$p_i = \max\{Reliability_r | r \in S_i\}$，来衡量，只要$p_i < \alpha$我们认为这个月内不太可能出现正面样本，也就不太可能发现亚洲大黄蜂，便称这个月是**"安全"**的。

  但是这非常依赖于我们所选的月份，有的月份人们发送的报告很少，很可能收到的报告全部符合要么缺失照片要么缺失评论的情况，导致出现指标较低的假象。但通过之前的数据分析可以得知，人们的报告数低迷期不会超过6个月，因此我们只需要持续考察6个月内每个月出现正面报告的可能性，只要这6个月都是**"安全"**的，那么我们便认为该地区发现亚洲大黄蜂的概率极低，可以认为亚洲大黄蜂在该地区被消灭了。

5 敏感性检验

我们的模型的敏感性主要来自于我们对指标中各因素权重的确定，因此，我们对分别改变各个成分的权重大小，来观察其对我们模型分析题目数据的结果的影响。

【分别将某一个因素的权值系数改变（-10%，-5%，+5%，+10%）的相对大小，其他的不变，画出对前20条排序结果的影响的图或者表来说明敏感性关系，共四张】

6 模型评价

6.1 模型长处

- 精确度还是比较高
- ……（**编吧**）

6.2 模型短处

- 具有一定的主观性（权重的分配）
- 数据缺失，对于模型的效果还是有一定的影响
- 训练神经网络需要付出比普通数学计算更高的计算代价
- ……（**未完待编**）

6.3 将来可以改进的地方

【瞎掰吧。。。】

7 Memo

大概要点如下：

1. 问候
2. 告诉农业部门，一种名叫亚洲大黄蜂的害虫已经入侵，它暂时没有天敌，很快将在华盛顿州内迅速传播。我们做了一个仿真模型，仿真结果显示如果不采取任何措施的话，在五年以内，它影响的范围和数量就会达到一个惊人的程度，届时将会对人民的正常生产生活造成很大的困扰，所以我们建议贵部门请尽早注意这种昆虫的动向。
3. 我知道即便是一个州，其人力物力仍然是有限的，没法对这种害虫进行地毯式的排查。好消息是贵部门有人民源源不断的疑似目击报告可以参考，但它们数量比较多，但大部分都属于误判，仅凭手工分析会花费大量的时间和金钱，非常低效。我们不愿意看到亚洲大黄蜂在华盛顿州肆虐，所以我们建立了一个模型来帮贵部门快速确定一个目击报告属于亚洲大黄蜂的目击报告的概率。
4. 我们都知道，物种入侵的初期的调查是一件相当耗费时间的工作。所以，调查的地区将很大程度上影响调查的效率。为了协助你们完成高效的调查目前，我们已经对所有的现有的目击报告完成了评估，并对报告地点的分布做了粗略的分析，根据我们的模型分析结果，亚洲大黄蜂的目击报告更有可能在这些地区出现，请贵部门尽早派人调查。
5. 当然，由于很多原因，我们没能收集到足够多的数据。不过，我们已经考虑到了这一问题。我们的模型支持更新，只需要提供新的报告就可以了，我们的模型会首先自己预测报告是亚洲大黄蜂的概率，再决定下一步要不要立即更新。当然，如果贵部门能够先让实验室准确认定这些报告是不是亚洲大黄蜂的目击报告，再输入我们的模型，这将会更进一步提升我们模型的准确程度，这就更好了。考虑到算力的代价与模型准确度提升的重要程度，我建议贵部门采取以下的更新策略。如果新增的报告里有正面样本，那么就立刻更新模型，否则，每个月月底用新增的报告来更新我们的模型即可。
6. 关于防治措施，我听说鸟类是各类昆虫的天敌，所以可以鼓励人们养鸟。同时对非法捕杀鸟类的人们给予更高的惩罚。当然，大体上可能还是需要依赖人工的力量来防治。
7. 当从某个月开始一年的时间内收到的报告的可靠度预测值都在α以下时，我们可以认为这个月是"安全"的，不过，这也有可能是侥幸。但是如果连续6个月内，每个月都是安全的，恭喜你们成功消灭了入侵的亚洲大黄蜂！
8. 结尾（**注意不要写队员的名字，如果一定要落款，写队伍编号**）

8 参考文献

运用正互反矩阵确定AHP中各成分的权重：

[1]Shiraishi S, Obata T, Daigo M. Properties of a positive reciprocal matrix and their application to AHP[J]. Journal of the Operations Research Society of Japan, 1998, 41(3): 404-414.

纬度对于大黄蜂种群数量的影响的假设来源：

[2]Keeling M J, Franklin D N, Datta S, et al. Predicting the spread of the Asian hornet (Vespa velutina) following its incursion into Great Britain[J]. Scientific reports, 2017, 7(1): 1-7.

传播距离的泊松分布假设：

[3]Kennedy P J, Ford S M, Poidatz J, et al. Searching for nests of the invasive Asian hornet (Vespa velutina) using radio-telemetry[J]. Communications biology, 2018, 1(1): 1-8.

使用元胞自动机仿真的想法：

[4]White S H, Del Rey A M, Sánchez G R. Modeling epidemics using cellular automata[J]. Applied mathematics and computation, 2007, 186(1): 193-202.

【还有一些检验算法来自知乎和百科的什么的也可以标一下，不过我不知道怎么标，另外我在群里发的那本书也可以写成参考文献，这里没写】