Machine Learning for Zoonotic Emerging Disease Detection

Xiaojin Zhu[†], **Jun-Ming Xu**[†], **Christine M. Marsh**[§], **Megan K. Hines**[§], **F. Joshua Dein**[*]
[†]Department of Computer Sciences
[§]Wildlife Disease Information Node, Nelson Institute for Environmental Studies
[*]Wildlife Disease Information Node, National Wildlife Health Center, US Geological Survey
University of Wisconsin-Madison. Madison, WI 53706 USA
jerryzhu@cs.wisc.edu

Abstract

We might have had an earlier identification of West Nile virus ten years ago had people reported that they were seeing dead crows in their backyards. This position paper suggests a wildlife monitoring system for far upstream detection of zoonotic disease outbreaks. Our system integrates wildlife surveillance from experts, news organizations, citizen scientists, and incidental observers. We outline the machine learning opportunities and challenges for such a system.

1. Background

Emerging diseases in the environment could have significant impacts on human and animal health, food production, and biodiversity. There has been heightened global awareness of the emergence of new infectious diseases, and the re-emergence of known diseases in new hosts and geographies. It is known that 75% of emerging pathogens are zoonotic (those exchanged between humans and other animals), and at least 70% of these have a wildlife component, either as a host, reservoir, or vector; recent examples include Avian Influenza, SARS, and West Nile Virus. There is evidence that each of these diseases was present and affecting wildlife before transmission to humans occurred; each of these previously undetected diseases has now caused significant human morbidity and mortality (Woolhouse & Gowtage-Sequeria, 2006; Sims et al., 2005).

Unfortunately, even in developed countries, comprehensive systems for wildlife disease surveillance are not available. Although there have been improvements in capacity over the past few years, they fall far short of the surveillance systems currently in place for humans and agricultural species. The two conditions that often create these gaps are the lack of clear responsibilities for wildlife disease reporting in governmental agencies, as well as the somewhat random nature of observations of mortality events. Detecting diseases in humans, pets and livestock is much easier, because the individual, parent or owner is motivated to seek treatment once illness is seen (Mörner et al., 2002; Rabinowitz & Conti, 2010).

2. Experts, Newspapers, Citizen Scientists, and You

Formal wildlife disease monitoring in the US has traditionally been carried out by experts at the National Wildlife Health Center, which receives reports of wildlife disease events primarily from US federal and state agencies. The Center often accepts diagnostic specimens from these cases to investigate the cause of the event. Once analyses are complete, the information is available in the form of the USGS NWHC Epizootic Database, accessible at http://www.nwhc.usgs.gov/mortality_events/ongoing.jsp. However, this traditional monitoring approach is insufficient. To have the greatest chance of recognizing uncommon and spatially disparate events, we will need to greatly expand the observational corps.

We propose to unite experts, news organizations, citizen scientists, and incidental observers in creating a wildlife monitoring and outbreak detection system, with machine learning serving as the "glue." These groups have different characteristics:

- The experts at the Wildlife Disease Information Node (WDIN; University of Wisconsin and USGS) can provide definite diagnosis, i.e., labels,

Appearing in *ICML 2011 Workshop on Machine Learning for Global Challenges*, Bellevue, WA, USA, 2011.

to an event. Nonetheless, the number of experts and the cases they can process is limited. Furthermore, the diagnosis may have a relatively large latency after the onset of the event.

- Significant wildlife incidents are often recognized by local news organizations. News reports are an effective tool for event awareness, and serve as a substitute for a more structured surveillance program (Keller et al., 2009). WDIN searches and sorts through these reports and produces a Wildlife Disease News Digest (`http:newsdigest.wdin.org`).

- Citizen scientists are non-professionals who agree to actively observe wildlife and report to professionals via specific channels. One such channel is WDIN's Wildlife Health Event Reporter (WHER), a website that enables anyone to report sightings of sick or dead wildlife (`http://www.whmn.org/wher/`). An accompanying mobile phone app "Outbreaks Near Me" accepts wildlife and human illness reports (`http://www.healthmap.org/outbreaksnearme/`). With awareness and interest, there can be many more citizen scientists than experts – the key is to promote participation. The citizen scientists can provide fairly accurate descriptions of an event, but usually cannot diagnose the cause of the event. Their reports can be near real-time.

- Incidental observers are people who produce machine readable information regarding wildlife, but are otherwise not affiliated with nor necessarily interested in wildlife monitoring. We may view them as *passive* citizen scientists, or more abstractly as human sensors with peculiar signal detection characteristics. A good example is Twitter, which has been used in scientific studies including earthquake monitoring (Earle et al., 2009). The following tweet is an example of wildlife encounter: *Dead armadillo on the side of the road with a buzzard picking at it; what a lovely sight on my trip to work. :P* There is a huge number of potential incidental observers, densely distributed in time (around the clock) and space (within populated areas). We speculate that incidental wildlife reports tend to happen at the geographical boundaries where human meets nature – deep inside forests there are few people while deep inside populated areas there are few wild animals. This is particularly relevant for zoonotic disease monitoring where the transmission from animals to human is likely to occur. Such reports are also suitable for studying the impact of human development on ecology. On the other hand, incidental reports are highly noisy. For example, the vast majority of tweets that mention animal names do not contain true wildlife encounters, as exemplified by this tweet: *saw your article on Mario 3DS, how do you feel about them reusing the raccoon tail?* Being able to sift and winnow is therefore crucial. Incidental reports also tend to be incomplete, missing important information such as time, location, species. These reports tend to be real-time.

We believe combining information from all these groups is essential to enhance our ability to detect changes in the patterns of wildlife disease occurrence that may signal the very earliest stages of disease emergence in human / animal systems. We discuss machine learning's role next.

3. Opportunities and Challenges for Machine Learning

We discuss some machine learning research questions that may be important to zoonotic emerging disease detection. When appropriate, we suggest potential approaches to address each question. Our discussion is aimed at improving awareness in both the machine learning and wildlife monitoring communities, rather than offering definite solutions.

1. The information from different groups differ dramatically: we have a small amount of precise diagnosis from experts, larger amount of relatively accurate event descriptions from news and citizen scientists, and a huge amount of very noisy reports from incidental observers. Reports generated by citizen scientists and incidental observers are unlabeled (in terms of disease diagnosis). Furthermore, many features might be either missing or specified at a very coarse level. For example, only 1% of tweets come with GPS coordinates – even that may not correspond to the location of the wildlife event. Machine learning models for learning from **unsupervised, semi-supervised, and weakly supervised** labels might offer a principled way to integrate such disparate information.

2. The experts' effort is limited – they should judiciously investigate the most suspicious wildlife events. Sifting through the large number of reports gathered automatically from news, citizen scientists, and incidental observers, a machine learning algorithm should help the experts prioritize events to investigate in order to maximize the

chance of successfully detecting emerging diseases. It should also take into consideration the availability and cost of acquiring samples. Some relevant machine learning techniques to start with might be **cost-sensitive active learning** and **ranking**.

3. To identify news stories on wildlife events, traditional machine learning approaches in **topic detection and tracking** can be used. Meanwhile, active learning can similarly be used to close the loop, supplying local news organizations with tips automatically aggregated from the other groups.

4. A report from a citizen scientist is more valuable than a report from an incidental observer. We would like to raise awareness of channels like WHER, "upgrading" interested incidental observers into contributing citizen scientists, and maintaining participation from existing citizen scientists. One possibility is to tap into the research in **social networks**, identifying important nodes (i.e., individuals or organizations) who are most likely to not only participate as citizen scientists, but also influence their neighboring nodes in a social network. For instance, local Audubon clubs may be good nodes to contact initially. As another example, one can imagine a cyber robot who automatically tweets and even contacts prolific incidental observers.

5. The noisy nature of incidental reports such as tweets requires significant natural language processing. An important task is to classify whether a report is indeed a wildlife encounter or not (the latter includes animal names used in non-animal senses, historical or imaginary mention of animals, forwarded news about wildlife, etc.). If yes, one also needs to extract the type of event, the time, the location, the animals involved, etc. Standard **information extraction** algorithms need to be adapted to the specific type of incidental reports to be effective. The task is further complicated by the fact that wildlife exists outside English-speaking countries, too, necessitating multilingual natural language processing.

6. Reports generated by citizen scientists and incidental observers are subject to several forms of **selective bias**. One is spatial bias, where there are fewer reporters in areas with abundant wildlife. Another one is temporal bias, where human reporters are less active when nocturnal animals are most active. Yet another one is psychological bias, where one is more likely to tweet the sighting of a dead cow than that of a dead crow. These biases need to be calibrated and removed with statistical methods and clever experimental design.

7. The view that incidental observers are sensors, and the fact that an outbreak has limited spatial and temporal footprint and is thus a sparse signal, suggests that recent theory and algorithms for **sparse signal recovery** might be brought to bear on the task of zoonotic emerging disease detection.

8. Reports from citizen scientists and incidental observers often include images and videos in addition to text description. For example, the WHER system and twitpic.com both allow users to upload them. Images and videos provide vital confirmation to experts, because most people are not good at identifying animals, and because they contain contextual information of the surroundings for the wildlife event. Standard **computer vision** techniques can be employed to help recognize the animal species and the scene. Its output can complement, correct, or even replace textual descriptions. Computer vision is particularly attractive as a cross-cultural approach, with less dependencies on solving multilingual natural language processing problems.

9. Expanding this project to developing countries poses special challenges. Traditionally, disease monitoring in developing countries relied heavily on governmental data. In reality, health event data for human, livestock, and wildlife is difficult to acquire in many developing countries. However, the rapid penetration of technology can change the situation in the near future. For example, local events can be transmitted to a village leader, who increasingly may have a phone or web connection. This trend is evident from the success of mobile-phone based projects such as Ushahidi.org and the Mobile Phonebased Infectious Disease Surveillance System (Robertson et al., 2010).

4. Roadkill on Twitter: The Power of Incidental Reports

As a demonstration of the scientific value in incidental reports, we present a preliminary study of roadkill statistics collected from Twitter. Although most contain non-zoonotic animal deaths, these reports are still useful to establish baseline spatial and temporal distribution statistics. Furthermore, we need to identify roadkill in order to exclude them from zoonotic injured or dead animal reports.

Traditionally, the department of transportation and a few roadkill observation websites collect reports from citizen scientists. For example, the Roadkill project recruited students and teachers to monitor roadkill (http://roadkill.edutel.com/). At the project's peak, during an 8-week period in 1997 it received 3962 roadkill reports from its citizen scientists, or about 70 reports per day. However, such effort is not sustainable year-round nor globally. Most people are not aware of these systems or not willing to participate due to the extra trouble they have to go through.

On the other hand, people are much more willing to share their incidental observations through social media such as Twitter. With our primitive algorithm below, we were already able to collect roadkill at a rate of 120 reports per day. The rate is expected to increase significantly as we improve our algorithm. Our algorithm also has a much better spatial and temporal coverage.

Our algorithm collects tweets through the Twitter stream API using a keyword list containing 370 wild animal names. During the 11-day period from April 23, 2011 to May 3, 2011, we collected 10,834,563 tweets containing at least one keyword in our list. It then uses a pipeline to identify the tweets describing roadkill and extract event details. First, it retains tweets containing `(ran AND over)` or `(dead AND road)`. Then, it removes the tweets written in non-English language by thresholding the fraction of non-English dictionary words. Next, it parses the tweets using the Stanford CoreNLP Tools (De Marneffe et al., 2006). Finally, it extracts the victim animal species using manually specified grammatical dependency rules. Our algorithm also extracts the time stamp and approximate location of each tweet. Among the tweets we collected, 1,294 roadkill tweets are identified, with a precision better than 75%. Here are some examples:

- *"Much love to the dead skunk in the middle of the road @aanchyyy"*
- *"oh god me and hannah just saw a dead squirrel by the road and screamed... awful"*
- *"she ran over a kangaroo yesterday at the new side of epping?! i never knew kangaroos are anywhere near the city! just heard crazy incidents"*
- *"I'm 78% sure I ran over a dolphin with a jet ski today. Shut your mouth hippies, Earth day is over. Next time I'm aiming for the manatees."*

There are some interesting observations from the events we extracted. Table 1 shows the top 20 most frequent roadkill species. Squirrel are the most frequent victims with greater frequency count than all other animals.

Table 1. The top 20 roadkill species. The numbers are average roadkill tweets per day.

Animal	Frequency	Animal	Frequency
squirrel	22	armadillo	3
rabbit	11	fox	3
bird	11	rat	2
skunk	10	turkey	2
snake	9	goose	2
turtle	7	opossum	2
duck	6	mouse	2
frog	5	beaver	1
deer	5	chipmunk	1
raccoon	4	bear	1

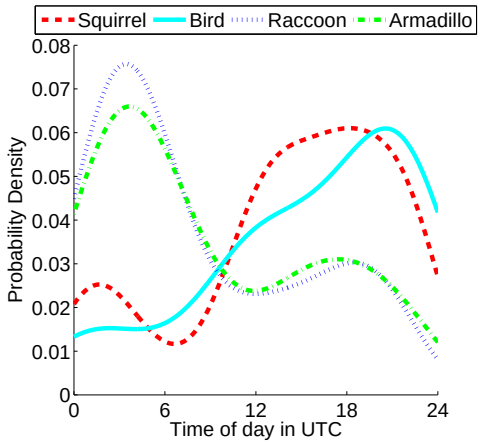
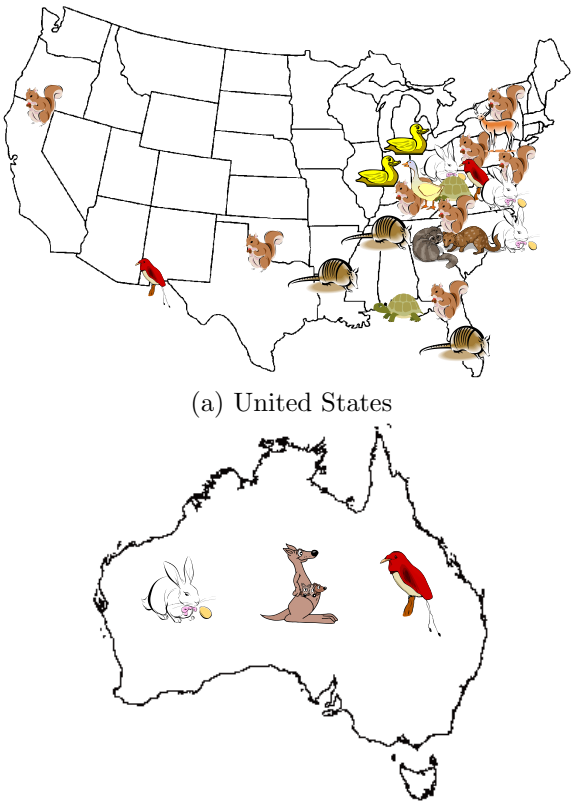


Figure 1. Temporal distribution of roadkill for four species.

Figure 1 shows the temporal distribution of roadkill tweets for squirrel, bird, armadillo and raccoon. Each curve is the probability density function of a species. The x-axis is the time of day in UTC when the tweets were generated. Given that US (especially east coast) dominates the tweets, the most likely local time is EDT and can be obtained by UTC-4. Clearly, squirrels and most birds have more probability density to the right (i.e., approximately daytime), while armadillos and raccoons have more density to the left (night time). This seems to correlate nicely with the fact that squirrels and most birds are diurnal, while armadillos and raccoons are nocturnal.

Figure 2 shows the spatial distribution in selected geographical regions of the most frequent species involved in roadkill tweets. We filtered the self-reported



(a) United States

(b) Australia

Figure 2. Spatial distribution of species in roadkill tweets

"user location" field for recognizable country and state names. 130 out of 1294 roadkill tweets contain this field. In Figure 2, a species is plotted only if it is involved in two or more tweets (among the 130) from that geographical region. Armadillos were more frequently mentioned in southern United States than the north part. More roadkill events were reported from the east coast than elsewhere in the US. Kangaroos have been encounter in Australia but not in the United States, as expected.

5. Conclusion

We believe zoonotic emerging disease detection is a rich domain, with the potential to further advance and complement machine learning research in natural sciences, including ecology and sustainability (Dietterich, 2009).

Acknowledgment

Research supported in part by Great Lakes-Northern Forests CESU Agreement 07HQAG0150. We thank the anonymous reviewers for helpful suggestions that improved the manuscript.

References

De Marneffe, M.C., MacCartney, B., and Manning, C.D. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, volume 6, pp. 449–454, 2006.

Dietterich, Thomas G. Machine learning in ecosystem informatics and sustainability. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Pasadena, CA, 2009.

Earle, P. S., Guy, M., Ostrum, C., Horvath, S., and Buckmaster, R. A. OMG earthquake! Can Twitter improve earthquake response? *Eos Transactions, American Geophysical Union, Fall Meeting Supplement*, 90(52), 2009.

Keller, Mikaela, Blench, Michael, Tolentino, Herman, Freifeld, Clark C., Mandl, Kenneth D., Mawudeku, Abla, Eysenbach, Gunther, , and Brownstein, John S. Use of unstructured event-based reports for global infectious disease surveillance. *Emerging Infectious Diseases*, 15(5):689–695, 2009.

Mörner, T, Obendorf, DL, Artois, M, and Woodford, MH. Surveillance and monitoring of wildlife diseases. *Rev Sci Tech*, 21(1):67–76, 2002.

Rabinowitz, PM and Conti, LA. Sentinel disease signs and symptoms. In Rabinowitz, PM and Conti, LA (eds.), *Human-Animal Medicine*, pp. 18–23. Saunders, New York, 2010.

Robertson, Colin, Sawford, Kate, Daniel, Samson L.A., Nelson, Trisalyn A., and Stephen, Craig. Mobile phonebased infectious disease surveillance system, Sri Lanka. *Emerging Infectious Diseases*, 16(10):1524–1531, 2010.

Sims, LD, Domenech, J, Benigno, DVM, Kahn, S, Kamata, DVM, Lubroth, J, Martin, V, and Roeder, P. Origin and evolution of highly pathogenic H5N1 avian influenza in Asia. *Veterinary Record*, 157:159, 2005.

Woolhouse, MEJ and Gowtage-Sequeria, S. Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, 11(12):1842–47, 2006.