Prospective Detection of Foodborne Illness Outbreaks Using Machine Learning Approaches

Aydin Teyhouee[1]✉, Sara McPhee-Knowles[1], Chryl Waldner[2], and Nathaniel Osgood[1]

[1] Department of Computer Science, University of Saskatchewan
{ayt227@mail,sam123@mail,osgood@cs}.usask.ca
[2] Western College of Veterinary Medicine, University of Saskatchewan
chryl.waldner@usask.ca

Abstract. Despite advances in food safety regulations, food-borne illness imposes a heavy health burden, with nearly 50 million estimated incident cases of illness each year. Having a prospective foodborne illness outbreak detection mechanism for more accurate and timely triggering of outbreak control measures would offer notable public health dividends, but is challenging due to the subclinical character of most foodborne illnesses. Within this work, collected synthetic datasets of incident illness cases and vendor contamination records from a previously contributed and empirically grounded model of foodborne illness, are used to study the efficacy of Hidden Markov Models (HMMs) for syndromic surveillance monitoring and disease outbreak detection under two data collection regimes, one involving a sentinel population using smartphone-based app for tracing location of food consumption and subclinical reporting. A support vector machine (SVM) approach was applied to compare the results to the HMM. Findings suggest that while reliance on clinical data offers poor potential for automatic outbreak detection, the use of HMMs offer excellent potential for detecting foodborne illness outbreak when informed by subclinical reporting by even a very small (4% of population) sentinel group. By contrast, SVM offers relatively poor prospects for detection. Furthermore, experiments with an empirically grounded agent-based model suggest that use of an HMM may be advantageous for triggering outbreak investigations among public health inspectors.

Keywords: Foodborne Illness, Outbreak, Prospective Detection, Machine Learning, Hidden Markov Model

1 Introduction

Each year, a large population worldwide suffer from foodborne illness.

While the public health inspection regime of food vendors successfully prevents many potential illnesses, the dynamic nature of restaurantsâĂŹ kitchens, the human resource constraints on carrying out consecutive inspections and the time-consuming character of the inspection process allow violations to remain undetected and limit the completeness of food illness prevention. Moreover, numerous food poisoned people who show mild to moderate symptoms of illness never show up at clinics and health care centers, but are greatly curtailed in their activity. While such subclinical cases impose stiff health, quality of life and economic costs, the absence of such data (of this kind of illness) in public health incidence records makes it almost impossible to figure out a potential outbreak occurrence. On the other hand, outbreak prediction methods mostly rely on telephone interviews of the clinical registered poisoned patients, days or weeks after their illness. This makes the situation even worse in two ways. First, the patient will be subject to forgetfulness about food vendors visited during a specified time, making it hard to prioritize the most probable contaminated restaurants in an investigation. Second, and a consequence, because of inaccuracies in the data collected and the prolonged investigation process, the adverse health and cost impacts of the outbreak will be magnified.

Lately, prospective detection of disease outbreaks in general using machine learning approaches has attracted the attention of researchers.The challenge on this new field is to diagnose the occurrence of an outbreak timely enough, helping the public health agents for taking quick outbreak controlling measurements. However, the applications of such works to foodborne illness outbreaks has been very limited [1]. Machine learning provides a set of tools which can be applied in different problem domains for data analysis. Given that the challenge of detecting foodborne illness outbreaks consists of identifying the evolution of the categorical latent state (outbreak vs. non-outbreak) of a system (municipality) over time in the light of noisy observations (incident cases) strongly influenced by state, Hidden Markov Models (HMMs) offer a particularly attractive analysis lens. Here we seek to distinguish between these outbreak and non-break states based on our observation of the number of reported illnesses. Although this is not the main goal of our presented work, we compare the findings of the HMM with the results of a Support Vector Machine (SVM) model, which fails to take into account the temporal context of data. To achieve this end, we will use synthetic ground truth data from a previously contributed empirically-grounded agent-based model (ABM) of foodborne illness.[2] As Sara M.Knowles shows in her model[2], and reflecting more recent successes in fieldwork by the authors, we further and mainly investigate how use of sentinel reporting of subclinical illnesses via smartphones could improve our inference about the potential outbreaks. To evaluate this, we will simulate two data collection regimes. The first regime focuses complements such traditional data with reports of subclinical illnesses provided by a small sentinel population, constituting just 4% of the total population. While this first data collection regime could be carried out with a number of technologies such as designated social media channels, call-in lines, and web-based mechanisms, we note that such a system has been successfully utilized over many months by the authors in using the Ethica iEpi [3] smartphone-based epidemiological data collection system; this work is currently being prepared for publication. In the second regime, we will use clinical data only, reflecting presentation by victims of possible foodborne illness to healthcare centers.

2 Overview of the Generative Model

Details of the foodborne illness ABM that serves to generate the synthetic time series that is used to train and test the machine learning models is described in a previous contribution [2]. However, we make some general comments about the model here. This model offers a stylistic depiction of a municipality that includes three types of actors: Persons, Restaurants and Inspectors. In the scenarios examined here, the municipality included a population of 5000 persons, 100 restaurants and one inspector. Restaurants can be either in a non-contaminated or contaminated state, with a transition hazard from the former to the latter such that an average of one restaurant per year becomes contaminated. The inspector can be in one of two modes: Routine inspection and outbreak response. In routine inspection mode, the inspector transitions between restaurants in a round-robin fashion In outbreak response mode, the inspector makes prioritized visits to restaurants according to the number of times that they have been identified (via faulty individual memory or via the geostamped records of sentinels) by those with clinical or (for the sentinel scenario) subclinical illness. Visits to restaurants are remembered by an individual. Independent of their source, foodborne illnesses developed in the model are classified clinical with a small probability (0.005), with the remainder remaining subclinical. Following a fixed period of time (2 days), individuals experiencing either of subclinical and clinical symptoms are treated as recovering, and return to a healthy state.

For analysis, each week, the model reports the incident case counts of clinical and subclinical illness and the count of contaminated restaurants.

3 HMM

HMMs are widely used in classification problems. Given a time horizon, they are used to infer the evolution of the system among a set of latent and non-observable categorical states over that horizon, with each of these states being associated with a specific distribution of observables. In this problem, we focus on discrete time characterization, with each time point representing a single week, and transitioning between two states s_t: a state in which the municipality includes a contaminated restaurant (henceforth termed the "outbreak" state) and $s_t = 1$, and one in which no contaminated restaurant is present and $s_t = 0$. Each such state is associated with a distribution for the observables $y_t(t = 1, ..., n)$: Clinical cases and (for the sentinel scenario) subclinical cases, where n is the n'th week. That is, for a given state s_t, $y_t|s_t \sim f_k(y_t; \theta_k)$, where $k \in \{0, 1\}$, f_k is a pre-specified density (e.g., univariate or multivariate Gaussian or Poisson) and θ_k are parameters to be estimated. The unobserved state space, $s_t(t = 1, ..., n)$ is modelled by a two-state homogeneous Markov chain of order 1 with stationary transition probabilities $p_{kl} = P(s_{t+1} = l|s_t = k)$, where $k, l \in \{0, 1\}$ denote the two states of s_t (0: non-outbreak; 1: outbreak). Note that in this Markov-dependent mixture model, y_t is conditionally independent of all the remaining variables, given s_t. As we are working with counted data in this experiment (number of reported illnesses), the above mentioned f_k is a Poisson density. So, to make it short, the expected frequency profile for the events of any dataset is definable in the format of a Poisson where all the mentioned conditions are true. An attraction of HMMs is the fact that it is possible to estimate their parameters using a variety of parameter estimation methods including the iterative Expectation Maximization (EM) algorithm. A preinvestigation over the dataset and the histograms corresponding to each of the two datapoint clusters, one can observe: a) A low level of illness occurrence, where the weekly incident case count can be modeled as a Poisson distribution with parameter λ_1, b) A high level of illness occurrence, where the weekly incident case count can be modeled as a Poisson distribution with parameter λ_2.

The iterating process of converging the Poisson distributions' lambda parameter is performed by assigning the two above mentioned observed Poisson distribution parameters to specify a starting model for the EM algorithm: $\Omega_0 = (\pi_0, P_0, b_0)$, where π_0 is the initial matrix, P_0 is the (2×2) transition matrix and b_0 is the (1×2) emission matrix containing the first guessed lambda parameters for each of the Poisson distributions. In this study, a package named mhsmm [4] in the R statistical computing framework (RDevelopment Core Team 2010) was used for parameter estimation. For training and cross-validation, the number of contaminated restaurants in successive weeks was rendered into a dichotomous variable serving as ground truth, assuming that any contaminated restaurant number greater than 1 corresponded to the state of an outbreak (whether declared or not). Also, the simulated 10,000-day (almost 27 years) dataset captured from the agent-based model was split up into training dataset (75%) and testing dataset (25%). To get an idea how good HMMs are performing in our problem possessing a temporal context, we utilize a Support Vector Machine (SVM) model with a linear kernel over our dataset. The results for both the models are presented in the next section.

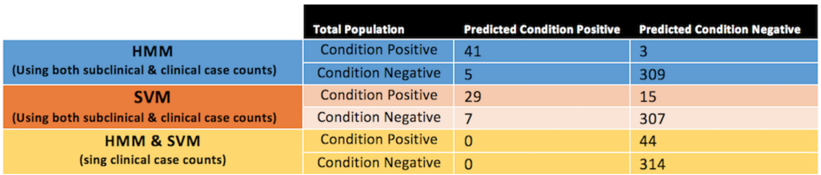
4 Results

- Results of the Hidden Markov Model (Using both subclinical and clinical case counts):
  Our HMM model $\Omega = (\pi_0, P_0, b_0)$ was initialized with $\pi_0 = (0.5\ 0.5)$, $P_0 = \binom{0.5\ 0.5}{0.5\ 0.5}$ and $b_0 = (1\ 4)$. These parameters are used by the EM algorithm to produce a maximum likelihood estimate Hidden Markov Model to describe the data. We evaluated models in terms of confusion matrix, sensitivity and specificity resulting from a cross-validation procedure over the test data. The best model obtained for the scenario 2 (the case where our observation includes both clinical and subclinical instances) has a set of parameters as follows: $\pi = (0\ 1)$, $P = \binom{0.990\ 0.010}{0.055\ 0.945}$ and $b = (7.869088\ 15.860456)$, resulting in a sensitivity of 0.9318182, a specificity of 0.9840764 and a confusion matrix as per Figure 1.
- Results of the Support Vector Machine Model (Using both subclinical and clinical case counts):

The predictive performance of the SVM was measured through a cross-validation process over different cost values with 10-fold sampling method and then a model with lowest misclassification error rate with a linear kernel was chosen.

This model obtained a sensitivity of 0.6590909, a specificity of 0.977707 and a confusion matrix as per Figure 1 over the testing dataset.

- Results of the HMM and SVM (Using clinical case counts):
  In this scenario where only the clinical incidences were considered, both the HMM and SVM approaches failed in labeling the outbreak state. In this case, the number of reported clinical cases were very rare, and all incidences were labeled as non-outbreak state. Figure 1, shows the confusion matrix for this scenario.



	Total Population	Predicted Condition Positive	Predicted Condition Negative
HMM (Using both subclinical & clinical case counts)	Condition Positive	41	3
	Condition Negative	5	309
SVM (Using both subclinical & clinical case counts)	Condition Positive	29	15
	Condition Negative	7	307
HMM & SVM (sing clinical case counts)	Condition Positive	0	44
	Condition Negative	0	314

Fig. 1. Confusion Matrix for Different Scenarios & ML Models

5 HMM-aided Outbreak Triggering System

To investigate whether the HMM could improve syndromic surveillance monitoring and linked disease outbreak detection systems, the ordinary illness triggering method (which is applied once at least 2 clinical cases happen) as mentioned in detail at section 2, was replaced with the resulted HMM in section 4. To carry out this HMM-based outbreak detection mechanism, the ABM uses the HMM parameters calculated in the HMM results of Section 4 to calculate the updated probability of being in an outbreak state in light of the previous value and the reported subclinical and clinical case counts. If the calculated probability at the beginning of any given week is greater than a threshold (which in this experiment was set to 0.6), a message is sent to the inspector to trigger the transition to the outbreak state investigation. The recorded cumulative count of clinical and subclinical illnesses over 10 years for 12 realizations in two different outbreak declaration regime (HMM outbreak triggering method and the ordinary method) is shown in the Fig. 2. Results show a quite significant decrease in the number of illness reports due to the fast detection of contaminated restaurants by applying the HMM outbreak declaration approach. As demonstrated by Fig. 2, this approach reflects a similar decrease in the time period a given contaminated restaurants remains contaminated before being identified and cleared.

6 Conclusion

Performing disease outbreak detection based on reported illness cases is an important function for syndromic surveillance systems. We treated the existence of a foodborne illness outbreak as a latent element of state and developed a

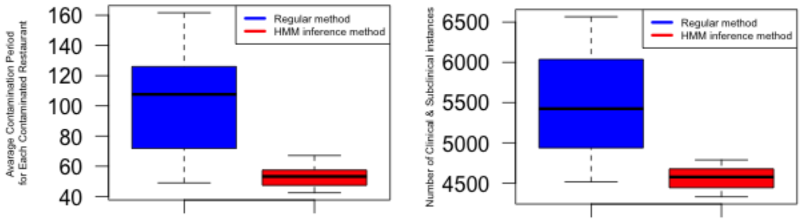


Fig. 2. Regular and HMM-based outbreak declaration comparison over 12 realizations for each: (Left) (Contamination period per contaminated restaurants [day/10-years]) - (Right) (Number of illness incidences [person/10-years]

Hidden Markov model for syndromic surveillance. We evaluated our disease outbreak detection approach using an empirically grounded previously contributed ABM of foodborne illness, comparing the results from HMM to those secured using an SVM approach. Finally, in light of the highly favourable results from the HMM, we further used the foodborne illness ABM to evaluate the public health gains secured through use of a HMM-based outbreak detection trigger, as compared with a traditional one based on case counts. Despite the highly noisy data present, and overlapping distributions of incident case counts between the outbreak and non-outbreak states, the results reported in this paper suggest a promising future for the use of hidden state variables to model the changing dynamics of observed surveillance time series, and for HMMs in general in outbreak signal detection. Moreover, the results from the first and second scenarios (considering both clinical and subclinical reports vs. considering only clinical reports) reveal that use of smartphones that can record locations and offer channels for reporting mild and moderate foodborne illness reports could improve our inference about the potential outbreaks. Finally, evaluation of HMM-based outbreak triggering mechanisms using ABMs suggest that significant public health gains may be secured when combining new technologies for syndromic surveillance with machine-learning based outbreak signal detection mechanisms. This work suggests promising lines of future work, including in extending our outbreak detection approach with multiple data streams obtained from mobile applications, such as restaurant-specific traffic and illness counts counts.

References

1. K. Morrison, K. Charland, A. Okhmatovskaia and D. Buckeridge, "A Framework for Detecting and Classifying Outbreaks of Gastrointestinal Disease," Online Journal of Public Health Informatics, vol. 5(1), 2013.
2. Sara McPhee-Knowles,"The Complex Problem of Food Safety Applying Agent-based Modeling to the Policy Process," digital repository for the College of Graduate and Postdoctoral Studies electronic theses collection, University of Saskatchewan, 2014.
3. "Ethica Data". Ethicadata.ca. N.p., 2016. Web. 29 Apr. 2016.
4. J. O'Connell, S. Hojsgaard."Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R," Journal of Statistical Software, vol. 39(4), pp. 1-22, 2011.