Aggregation of Location Attributes for Prediction of Infection Risk

Slobodan Vucetic and Hao Sun

Center for Information Science and Technology, 303 Wachman Hall,
Temple University, 1805 N. Broad St., Philadelphia, PA 19122

Abstract. In this study we propose an algorithm for the prediction of an infection risk that is based on the aggregation of locations and their use as prediction attributes. Our algorithm is tested on a specific instance of EpiSims simulated data for Portland, OR. The results indicate that location aggregation is very promising approach that can result in high prediction accuracy and could be helpful in modeling of epidemics.

1. Introduction

Despite numerous advances in medicine, the risks associated with occurrences of well-known, mutated, or new pandemic diseases, such as H5N1 avian influenza in Southeast Asia [1], are among the largest threats facing world health. Traditionally, key pandemic response elements have included: (i) surveillance, investigation, and protective health measures, (ii) viral and anti-viral drugs, (iii) health care and emergency response [1]. Several of these response actions directly motivate research and developments in data mining.

An open question is how modern data collection techniques and data mining could be integrated to better understand the spread of a new infection. The recently developed EpiSims [2] simulation tool provides an excellent environment for the development and testing of various data mining techniques for pandemic response. Recently, the EpiSims [3] team has published simulated data corresponding to a particular instance of infection outbreak in Portland, OR. The data consists of 5 data tables with detailed information for about 1.6 million people and 240 thousand locations in Portland, including people's movement, activities, social contacts, and the infection spread.

While highly detailed simulated data are useful for understanding the properties of various infections, we also need to consider the type and availability of information real life situations would provide and how such information might be used in response to an outbreak. This problem is the primary motivation for our study. Our assumption was that, using the current technology while considering privacy issues, it would be possible to collect very valuable information for disease response. For example, by using cell phone records, it would be possible to track human movement quite accurately. In addition, by performing appropriate surveys, information about the type of activities occurring at every location could be obtained. Using this information, our hypothesis is that data mining can be very useful when predicting those people most at risk, shortly upon the outbreak of an infection. Additionally, by analyzing the developed predictor, it could be possible to gain better understanding of the epidemics and use the results to improve quality of epidemics modeling [4].

The goal of our study is to explore the ways in which the risks of infection could be predicted using human movement and location-specific activity data. To constrain the scope of our study, we concentrate on diseases that are transmitted only through human contact. In such cases, close proximity among people is a necessary condition for infections to spread. Our approach is based on an assumption that that the type of location is an important determinant of infection risk. For example, people who are congregated in a large public area are at a lesser risk of infection than those who are in closer physical contact with an infected person. The proposed approach is based on the aggregation of locations into specific types and their uses as prediction attributes. In this paper, we propose to aggregate locations based on the nature of the ongoing activities.

2. Data Sets and Data Preprocessing

2.1. Data Sets

The original data [3] consists of five data tables that are the result of one simulation by the EpiSims model:

People (PortlandProtoPopulation). This table consists of basic information about 1.6 million inhabitants of Portland.
Locations (PortlandProtoLocations). Contains spatial coordinates of about 240 thousand locations in Portland.
Activities (PortlandActivities). Provides information about activities of each person including location of activity, type and time of the activity.
Contacts (PortlandContactGraph). Provides a detailed

graph of contacts between people, their duration, and their type of contact.

Infections (PortlandDendrogram). Provides detailed information about infection spread through the first 20 generations. For each person infected in generation i, we are given information about where infection occurred and which person from generation (i–1) transmitted it.

2.2. Data Preprocessing

The original data provides all pertinent information about the specific instance of an infection outbreak as simulated by EpiSims model. Our goal was to transform the original data into a form that could be realistically collected with the use of current technology and without overstepping boundaries of privacy. We used the following assumptions:

1. It is possible to **track movements** of people through the use of cellular phone tracking data.
2. It is difficult to measure **activity type, actual contacts and contact type** with other people. This information could not be easily obtained by the current technology, and would represent a serious privacy breach.
3. Types of **activities at each location** could be obtained. For example, it could be measured that 100 people are present at location L_i between hours 7–12 and that 70% of them are there for work, while 30% are there for recreational purposes. This type of information can be collected anonymously and without any intrusion of privacy.
4. It is difficult to know **where an infection occurred and who transmitted it**. Although this information could be estimated by interviewing infected people, we assume it would be difficult to collect it in a timely and comprehensive manner.
5. It is possible to **estimate the generation** of each infected person. For example, a reasonable generation estimate could be obtained by recording the time at which symptoms become visible.

Given the assumptions, we produced several data sets:

LocationActivity. For each location L_i, i =1…L (L = 243,423), we recorded 33 activity types, A_j, j = 1…33. These activity types were derived from the **Activities** table. Elements of the resulting table LA were $LA(i,1)$ – total person-hour occupancy at location L_i during the day by people having "Home" activity; $LA(i,2)$ to $LA(i,5)$ – the occupancy during hours 1–6, 7–12, 13–18, 19–24 by people having "Work" activity. The remaining 7 quadruples of columns were filled in similarly to $LA(i,2)$ – $LA(i,5)$, by recording the occupancy of people engaging in "Shop," "Visit," "Social/ Recreational," "Other," "Pick up or drop off a passenger," "School," and "College" activities.

PersonLocationTime. For every person P_i, i = 1…M (M = 1.6 million), we constructed a binary matrix PLT_i with L (=243,423) columns and 24 rows. Element $PLT_i(l,h)$ has a value 1 if person P_i was at location L_l during hour h, and 0 otherwise.

Generation. For every person P_i, i = 1…M (M = 1.6 million), we recorded its infection generation. Elements of the resulting vector $G(i)$ were set to 0 if the person was not infected during the time of simulation, and to a number between 1 and 20 to indicate the infection generation.

GenerationLocationTime. For every triple (g, L_l, h), we recorded the total number of infected people from generation g (=1…20) present at location L_l at hour h, and saved it as element $GLT_g(l,h)$ of matrix GLT_g.

GenerationPersonLocation. For every triple (g, P_i, L_l) we recorded the total number of contact hours person P_i spent with infected people from generation g at location L_i, and saved it as element $GPL_g(i,l)$ of matrix GPL_g. This value can be obtained by using the dot product between PLT_i and GLT_g matrices.

Out of these five sets, we used **LocationActivity**, **Generation**, and **GenerationPersonLocation** data in the further study.

3. Methodology

3.1. Problem Definition

Our objective was to explore whether or not we could predict if healthy person P_i will become infected in generation g. This task can be defined as a classification problem by representing person P_i as an L-dimensional vector $x_i^g = (x_{i1}…x_{iL})$, where $x_{il} = GPL_{g-1}(i,l)$, and labeling it as $y_i^g = 1$ if $G(i) = g$, and $y_i^g = -1$ otherwise. We denote the resulting data set as $D^g = \{(x_i^g, y_i^g), i = 1…M\}$.

Given these definitions, the classification problem can be stated as learning from historical data $D^1…D^{g-1}$ to build a prediction model $f(x)$ that predicts whether a person P represented with an attribute vector x will become infected in generation g. Instead of pure classification, it is more appropriate to use prediction model $f(x)$ to rank all people by their risk of becoming infected. By default, we assume that the risk of infection is negligible for people with zero vectors x (i.e., people not in contact with infected people from generation g–1). This assumption, while valid for the specific data set studied here, might not be appropriate for infections that do not require direct human contact.

3.2. Approach

The basic idea of our approach was that the risk of infection varies with the type of location where infected people and people at risk are gathered.

In the simplest case, which we will call **the baseline predictor**, the total number of contact hours with infected people is used as measure of infection risk – the larger it is, the greater the risk of contracting the disease. Therefore, the baseline predictor can be described as $f(x_i) = \Sigma_l(x_{il})$.

As another extreme, one can attempt to build predictor f directly from the L-dimensional data D^{g-1}. There are two major problems with this approach. One is dealing with highly dimensional attribute space, because every location is represented as an attribute. Another is ability to generalize. For example, if the outbreak was in the eastern part of a city, the trained predictor will not be able to generalize and predict infection risk in the rest of the city.

Our approach is to aggregate all locations into a small number of clusters and use the clusters as attributes in classification. Let us assume that location L_l is assigned to one of the K clusters as $c(L_l) \in \{1 \ldots K\}$. The original L-dimensional attribute vector x_i can be transformed into a K-dimensional new attribute vector $z_i = (z_{i1} \ldots z_{iK})$, whose elements z_{ij} are defined as

$$z_{ij} = \sum_{c(L_l)=j} x_{il} \ .$$

This operation results in reduction of the original L-dimensional into K-dimensional attribute space. In the following section we describe an approach for clustering of locations.

3.3. Clustering of Location Attributes

Our approach is based on clustering of **LocationActivity** data LA (see Section 2.2). The LA data provides useful information about the type of activities occurring at any given location. Our hypothesis was that those locations with similar types of activities are likely to pose similar infection risks.

We used the k-means algorithm to cluster L (= 243,423) locations into K = 2, 5, 10, 20, 50 clusters. Before clustering, we transformed the LA data in two different ways:

LogLA. In this case, prior to clustering, the LA matrix was transformed to $LogLA$ as $LogLA(i,j) = log(1+ LA(i,j))$. This step is justified by the nearly lognormal distribution of each of the 33 LA attributes.

LogNormLA. In this case, $LogLA$ was further scaled to $LogNormLA$, such that every of the column of the new matrix has mean 0 and standard deviation 1. This step is justified by the need to decrease the influence of the most common activities (e.g. "Home" and "Work") on clustering.

3.4. Classification Models

In this study, we considered Linear Regression (LR) and Support Vector Machines (SVMs) with linear kernels. Both algorithms are directly applicable to infection risk prediction because their outputs are correlated with the probability of infection. Let us represent each person P_i with pair (z_i, y_i), where $z_i = [z_{i1} \ldots z_{iK}]$ is attribute vector and y_i is class label.

LR is optimized to find coefficients α_0, $\alpha_1 \ldots \alpha_K$ that minimize the mean squared prediction error $E[(y - f(z))^2]$, where

$$f(z) = \alpha_0 + \sum_{i=1}^{K} \alpha_i z_i \ .$$

SVMs (Vapnik, 1995) are optimized to find the decision hyperplane with the maximum separation margin between positive and negative data points. The output of SVMs for attribute vector $z = [z_1 \ldots z_K]$ is calculated as

$$f(z) = b + \sum_{j=1}^{N_S} \alpha_i y_i K(z_j, z) \ ,$$

where N_S is number of support vectors selected from training data, α_i, $i = 1 \ldots N_S$, and b are model parameters obtained by optimization, and K is an appropriate kernel function. For SVM training, we used SPIDER version 1.6 package with SVMLight optimizer. Linear kernel was used and slack variable was set to C=100. Because the number of positive and negative examples is unbalanced, we used the balanced ridge method.

3.5. Accuracy Measure

The objective of infection risk prediction is to achieve a high ranking of people that are most at risk from infection. To evaluate prediction quality we used AUC accuracy obtained as the area under the ROC curve. An ROC curve measures the trade-off between true positive (TP; fraction of positives predicted as positives) and false positive (FP; fraction of negatives predicted as positives) prediction rates for different prediction cutoffs. Given the prediction cutoff θ, all data points with prediction above θ are considered positive and all below negative. If θ is very small, no positives are predicted, and TP = 0 and FP = 0, while if θ is very large TP = 1 and FP = 1. Predictors that achieve high TP over a range of FP are considered accurate – AUC measures exactly this aspect of prediction quality. Perfect predictors achieve AUC = 1, random predictors have AUC = 0.5.

4. Results

We first explored the usefulness of LogLA and LogNormLA transformations prior to location clustering. In Table 1, we show the AUC accuracy of LR predictors trained on generation 5 (G5) data and

tested on generation 6 to 15 (G6 – G15) data. The second column of Table 1 is the AUC accuracy of the baseline predictor that ranks people by the total number of contact hours with infected people from the previous generation (G4). Columns 3 – 7 represent AUC accuracies obtained on data obtained by applying k-means clustering with K = 2, 5, 10, 20, 50 on LogLA transformed data. Columns 8 – 12 represent the corresponding AUC accuracies obtained when using LogNormLA transformation before clustering.

The results indicate that K = 10 and K = 20 clusters is the best choice with LogLA transformation, while K = 20 clusters is the best choice with LogNormLA transformation. Moreover, LogNormLA transformation resulted in an increase of AUC accuracy by up to 0.02, as compared to LogLA. Additionally, as compared with the baseline predictor, LogNormLA predictors have up to 0.05 higher AUC accuracy. The accuracy of LR predictors decreases with generation number. Additionally, the difference between LR predictors and the baseline predictor seems to slowly decrease with the generation number.

In the next set of experiments we first explored if successful predictors could be obtained quickly after the onset of an infection outbreak. Columns 4 – 7 in Table 2 represent accuracies of LR predictors learned on generation 1, 3, 5, 10 data and tested on all generations (G1–G15). The results show gradual improvement of AUC accuracies with the generation number. While predictor trained on G1 data was slightly inferior to the baseline predictor, the G3 predictor was more accurate than the baseline predictor. LR predictors trained on G5 and G10 data were the most accurate, and had a slight difference between them. Based on the LR results, it seems that successful predictors could be developed very early after the outbreak.

Columns 8 – 11 in Table 2 show the accuracy of SVM predictors with linear kernel trained on G1, G3, G5,

Table 1. AUC*100 accuracy of LR model trained on G5 data and tested on G6–G15 data.

TEST	BASELINE	LogLA Clustering, K =					LogNormLA Clustering, K =				
		2	5	10	20	50	2	5	10	20	50
G6	74.0	74.7	74.6	75.9	76.9	77.4	74.2	75.9	77.0	**78.2**	78.0
G7	74.0	74.5	74.3	76.7	77.8	77.5	74.4	75.3	77.3	**79.2**	78.2
G8	73.7	74.2	73.8	75.0	75.6	75.8	73.8	75.5	76.4	**77.8**	76.5
G9	71.0	71.6	72.0	73.5	73.9	73.7	71.2	72.9	73.6	**74.3**	73.6
G10	68.0	68.8	68.9	70.6	71.2	71.5	68.0	69.9	70.7	**72.2**	71.9
G11	68.2	68.6	69.2	70.5	70.8	70.9	68.0	69.7	70.6	**72.1**	71.0
G12	67.3	67.7	68.3	69.1	69.5	69.4	66.8	68.4	69.0	**70.4**	69.4
G13	67.0	67.3	67.8	68.6	68.4	68.6	66.0	67.7	68.1	**69.8**	68.1
G14	65.8	66.2	66.3	67.0	67.1	67.0	64.6	66.0	66.5	**68.1**	66.7
G15	65.5	66.2	65.7	66.4	66.5	66.5	63.8	65.0	65.7	**67.2**	65.9

Table 2. AUC*100 accuracy of LR and SVM models trained on G1, G3, G5, G10 data and tested on G1 – G15 data. LogNorm clustering was used with K = 20. Second column is the number of infected people in generation Gi.

TEST	|Gi|	BASELINE	LR				SVM			
			G1	G3	G5	G10	G1	G3	G5	G10
G1	84	73.4	80.8	79.3	81.1	75.9	82.1	72.5	81.5	78.3
G2	149	78.9	79.1	79.8	81.2	82.2	78.4	81.3	80.7	81.8
G3	238	76.6	76.7	78.9	79.7	78.1	70.9	79.3	79.5	79.4
G4	420	76.4	73.7	**77.4**	79.1	79.6	68.8	**76.8**	78.8	78.3
G5	648	75.5	73.8	**75.8**	79.2	77.3	71.4	**74.0**	78.6	75.5
G6	1101	74.0	72.7	76.3	**78.2**	78.4	68.0	76.2	**78.0**	77.7
G7	1881	74.0	74.5	76.9	**79.2**	78.7	70.1	76.6	**78.7**	77.8
G8	2997	73.7	72.7	75.5	**77.8**	77.3	67.5	75.6	**77.6**	77.0
G9	4718	71.0	69.4	72.3	**74.4**	74.9	58.1	71.9	**74.3**	74.6
G10	7570	68.0	67.6	69.6	**72.2**	73.1	64.6	68.8	**72.0**	72.2
G11	12765	68.2	67.3	69.5	**72.1**	**72.2**	62.7	69.1	**72.0**	71.5
G12	21167	67.3	65.9	68.2	**70.4**	**70.5**	66.0	67.8	**70.4**	69.9
G13	34063	67.0	65.4	67.1	**69.8**	**70.1**	63.1	68.2	**69.9**	68.9
G14	51711	65.8	64.5	65.9	**68.1**	**68.3**	61.3	65.5	**67.8**	67.0
G15	70174	65.5	64.0	65.4	**67.2**	**67.3**	59.7	64.8	66.5	**67.1**

and G10 data and tested on G1 – G15 data. Accuracies of SVM and LR models are comparable. We also evaluated SVM with nonlinear kernels, such as a polynomial kernel of degree 2, and a radial basis function kernel. However, we did not observe significantly improved accuracy. (Results are not shown.) Considering significant efforts needed for selection of parameters of SVMs and high computational complexity of SVM training, linear regression seems to be the more attractive alternative

Both LR and SVM predictors can be used to analyze the influence of each location type on infection spread. In Figure 1, we illustrate how to visualize the results and potentially gain an insight into properties of infection. In the upper panel of Figure 1, we illustrate each of the $K = 20$ location types obtained by k-means clustering of LogNormLA transformed data. Each of the 33 rows represents the number of people pursuing one of the 33 activities represented by the LocationActivity data (Section 2.2), while each of the 20 columns is a representative of one of the 20 clusters. Dark boxes mean that many people pursue the given activity at the given location, while light boxes mean that only a few people are involved. For example, column 17 corresponds to locations that most likely represent colleges since the most prevalent activity is of type "College" and occurs between hours 7-24 (boxes in rows 31-33 are the darkest). Bars in the lower panel of Figure 1 represent coefficients $\alpha_1 \ldots \alpha_{20}$ of the LR model. Analysis of the coefficients provides an insight about risk factors associated with each location type. For example, location types 8 and 11, with the highest α values, seem to correspond to predominantly residential locations (with the darkest box is in row 1 and relatively light remaining boxes), suggesting that being collocated with an infected person at home carries the largest risk of infection.

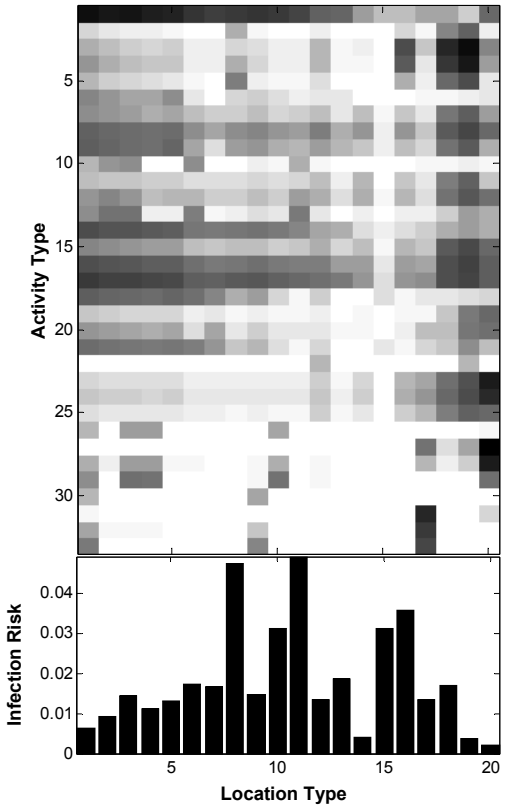

Figure 1. Illustration of location clusters and the associated risks.

4. Conclusions

Our results indicate that the proposed location aggregation approach is very promising for the prediction of infection risk. As compared to the baseline approach that uses number of people-hours in contact with the infected people as predictor of infection risk, AUC accuracy of our method is up to 0.05 higher. Additionally, it seems that accurate predictors could be developed early after the outbreak, by using only the first few generations of infected people. Analysis of the prediction model provided an insight into the risk factors associated with various location types. This information can be useful in disease modeling and developing successful control strategies and vaccination deployment.

Further improvements in location aggregation are possible, possibly by integrating k-means clustering with expert knowledge about various location types, or by including demographic and health related information of the exposed people.

Acknowledgement

This study was supported in part by NSF grant IIS-0546155 to S. Vucetic.

References

1. Ferguson NM; et al.,. Strategies for Containing an Emerging Influenza Pandemic in Southeast Asia, Nature, 437, 209-214, 2005.
2. Synthetic Data Products for Societal Infrastructures and Proto-Populations: Data Set 1.0, NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, VA, ndssl.vbi.vt.edu/Publications/ndssl-tr-06-006.pdf
3. HHS Pandemic Influenza Plan. US Department of Health and Human Services, November 20.
4. Colizza V, Barrat A, Barthelemy M, Vespignani A., The role of the airline transportation network in the prediction and predictability of global epidemics, Proc Natl Acad Sci USA 14;103(7), 2015-20, 2006.