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 proposed an algorithm for prediction of infection risk that is based on aggregation of locations and their use as prediction attributes. The 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.

Introduction

Despite numerous advances in medicine, the risks associated with occurrences of well-known, modified, or novel pandemic diseases, such as H5N1 avian influenza in Southeast Asia [1], are among the largest threats facing the human race. Traditionally, key pandemic response elements include: (i) surveillance, investigation, and protective health measures, (ii) viral and anti-viral drugs, (iii) health care and emergency response [1]. Several of the 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 spread of a new infection. The recently developed EpiSims [2] simulation tool provides an excellent environment for development and testing of various data mining techniques for pandemic response. Recently, the EpiSims [3] team has published a thorough 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, inclusive of people movement, activities, and social contacts, and infection spread for a particular simulation instance.

While highly detailed simulated data are useful for understanding the properties of various infections, an important question is what type of information we can expect in a real life situations and how can such information be used in response to infection outbreak. This question motivated our study. Our assumption was that, using the current technology while considering privacy issues, it could be possible to collect highly valuable information for disease response. For example, by using cell phone records, it could be possible to track people movements quite accurately. In addition, by performing detailed 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 helpful in predicting people most at risk, shortly upon the outbreak of an infection. To constrain the scope of our study, we concentrate on diseases that are transmitted only by human contact. In this case, collocation is necessary condition for infection to spread.

The goal of our study was to explore if infection risk could be predicted by using people movement and location type data. Our approach is based on an assumption that that the type of location is important determinant of infection risk. For example, being at the same time in a big shopping mall with an infected person bears smaller risk than if the location is a coffee shop or an apartment. However, properties of each location, with the respect to the specific disease, are not known in advance. Additionally, in initial stages of infection, only a small fraction of locations is visited by infected people. Successful prediction algorithm should be able to generalize from this limited number of locations.

The proposed approach is based on aggregation of locations into specific types and their use as prediction attributes. In this paper, we propose an automatic procedure for location aggregation that is based on the nature of activities occurring at any given location.

2. Data Sets and Data Preprocessing

2.1. Data Sets

The original data [3] consists of 5 data tables that are the result of one simulation by 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 detailed graph of contacts between people, their duration and contact type.

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

2.2. Data Preprocessing

The original data is extremely detailed and provides all pertinent information about the specific instance of infection outbreak as simulated by EpiSims model. Our goal was to transform the original data into a form that could be reasonably expected to be collected by current technology and by being sensitive to the privacy issues. We used the following assumptions:

1. It is possible to **track movements** of all people. Given current technology, it is a near possibility (e.g., by access to the cell 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 measured that 100 people were present at location L_i between hours 7–12 and that 70% of them were there for work, while 30% were there for recreation purposes. This type of information can be collected anonymously, without privacy intrusion.
4. It is difficult to know **where infection occurred and who transmitted it**. Although this information could be estimated by interviewing each infected person, we assume it would be difficult to collect it for all infected people.
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 when symptoms became visible.

Given these assumptions, we produced several data tables:

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 $LA(i,1)$ – person-hour occupancy during the day by people having "Home" activity; $LA(i,2)$ to $LA(i,5)$ – 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 similar way to $LA(i,2)$ to $LA(i,5)$, by recording occupancy by people having "Shop", "Visit", "Social/ Recreation", "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 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 hour h, and every location L_l, we recorded the total number of infected people present from generation g (=1…20), and saved it as element $GLT_g(l,h)$ of matrix GLT_g.

GenerationPersonLocation. For every person P_i, every generation g, and every location L_l, we recorded total number of contact hours this person spent with infected people from generation g at the given location, 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 5 sets, we used **LocationActivity**, **Generation**, and **GenerationPersonLocation** data in the further study.

3. Methodology

3.1. Problem Definition

Our objective was to explore if it is possible to predict what people will become infected by generation $(g–1)$ people. If person P_i is infected, he/she becomes a member of generation g. This task can be defined as 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 person P_i will become infected in generation g. Instead of pure classification, it might be more appropriate to use prediction model $f(x)$ to rank all people by their risk of being infected. By default, we assume that the risk of infection is negligible for people with zero vectors x (i.e., people not collocated with infected people). This assumption, while valid for the specific data instance 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 type of location where infected people and people at risk are collocated. In the simplest case, which we will call **the baseline predictor**, the total number of contact hours with infected people is used for prediction of infection risk. In this case, predictor f can be easily constructed as $f(x_i) = \Sigma_l(x_{il})$. As another extreme, one can attempt to build predictor f directly from the L-dimensional data $D^1 \ldots D^{g-1}$. There are two major problems with this approach. One problem is dealing with highly dimensional attribute space, where every location is represented as an attribute. More serious problem is ability to generalize. For example, outbreak of an infection can be in the Eastern part of the city, and only the attributes corresponding to locations in this part of city will have nonzero values. Therefore, 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 a given clustering assigns each location L_l to the corresponding cluster as $c(L_l) \in \{1 \ldots K\}$, where K is the number of clusters. The original attribute vector x_i can be transformed into a new attribute vector $z_i = (z_{i1} \ldots z_{iK})$, whose element z_{ij} is 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. The remaining question is what type of clustering is appropriate for the problem of infection risk prediction.

3.3. Clustering of Location Attributes

We propose the approach 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 locations. Our hypothesis was that that locations with similar type of activities are likely to pose similar risks of infection spread.

In this paper, 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, *LA* matrix is 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* is 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 influence of the most common activities (e.g. "Home" and "Work").

3.4. Classification Models

In this study, we considered Linear Regression (LR) and Support Vector Machines (SVM) with linear kernels. Both algorithms are directly applicable to infection risk prediction because their outputs are correlated with the posterior 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

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

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

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

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. To overcome the problem that 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 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, while predictors that provide random predictions 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 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 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 clusters is the best choice with LogLA transformation, while 20 clusters is the best choice with LogNormLA transformation. Moreover, LogNormLA transformation resulted in increase of AUC accuracy by about 0.01, as compared to LogLA. Additionally, as compared with the baseline predictor, LogNormLA predictors have up to 0.05 higher AUC accuracy. It is interesting to observe that the accuracy of LR predictors decreases with generation number. Additionally, the difference between LR predictors and baseline predictor seems to decrease with generation number.

In the next set of experiments we first explored if successful predictors could be obtained quickly after the 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 generation number. While G1 was often inferior to the baseline predictor, G3 predictor was more accurate and was consistently outperforming the baseline predictor. G5 and G10 LR predictors were the most accurate, and there was quite a small difference between them. Based on the LR

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

TEST	BASELINE	LogLA Clustering, K =					LogNorm Clustering, K =				
		2	5	10	20	50	2	5	10	20	50
G6	**72.1**	72.5	71.6	74.4	**74.7**	73.8	72.1	73.5	74.5	**75.7**	75.5
G7	**70.3**	70.8	70.4	73.0	**73.4**	72.5	70.6	71.5	73.2	**74.8**	74.0
G8	**68.6**	69.0	68.3	**70.7**	70.4	69.8	68.7	70.0	70.6	**71.8**	70.8
G9	**65.7**	66.1	65.5	**68.1**	68.0	66.8	65.8	67.0	67.6	**68.2**	67.6
G10	**60.8**	61.1	60.6	62.7	**62.9**	62.5	60.8	61.9	62.4	**63.3**	63.1
G11	**59.2**	59.4	59.0	**60.7**	60.7	60.1	59.1	60.0	60.4	**61.2**	60.6
G12	**57.1**	57.2	56.8	58.1	**58.1**	57.7	56.9	57.6	57.8	**58.4**	58.0
G13	**55.4**	55.5	55.2	**56.1**	56.0	55.7	55.1	55.7	55.8	**56.3**	55.8
G14	**53.9**	53.9	53.7	**54.3**	54.3	54.0	53.6	53.9	54.1	**54.4**	54.1
G15	**53.1**	53.2	52.8	**53.3**	53.3	53.1	52.7	53.0	53.1	**53.4**	53.2

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 number of infected people in generation Gi.

TEST	|Gi|	BASELINE	LR				SVM			
			G1	G3	G5	G10	G1	G3	G5	G10
G1	84	**74.4**	81.0	79.4	80.7	75.8	82.1	72.5	81.5	78.3
G2	149	**78.3**	78.7	79.8	81.0	**81.6**	78.4	81.3	80.7	**81.8**
G3	238	**76.4**	76.3	78.7	**79.2**	77.8	70.9	79.3	**79.5**	79.4
G4	420	**76.3**	73.4	77.0	78.9	**79.1**	68.8	76.8	**78.8**	78.3
G5	648	**74.2**	72.5	74.4	77.4	75.8	71.4	74.0	78.6	75.5
G6	1101	**72.1**	70.7	74.1	75.7	**75.9**	68.0	76.2	**78.0**	77.7
G7	1881	**70.3**	70.8	72.8	**74.8**	74.4	70.1	76.6	**78.7**	77.8
G8	2997	**68.6**	67.8	70.0	**71.8**	71.4	67.5	75.6	**77.6**	77.0
G9	4718	**65.7**	64.5	66.6	68.2	**68.6**	58.1	71.9	74.3	**74.6**
G10	7570	**60.8**	60.6	61.7	63.3	63.8	64.6	68.8	72.0	72.2
G11	12765	**59.2**	58.7	59.8	**61.2**	61.2	62.7	69.1	**72.0**	71.5
G12	21167	**57.1**	56.5	57.4	**58.4**	58.4	66.0	67.8	**70.4**	69.9
G13	34063	**55.4**	54.9	55.5	56.3	**56.4**	63.1	68.2	**69.9**	68.9
G14	51711	**53.9**	53.6	53.9	54.4	**54.5**	61.3	65.5	**67.8**	67.0
G15	70174	**53.1**	52.8	53.1	**53.4**	53.4	59.7	64.8	66.5	**67.1**

results, it seems that successful predictors could be developed very early after disease outbreak.

Columns 8 – 11 in Table 2 show accuracy of SVM predictors with linear kernel trained on G1, G3, G5, and G10 data and tested on G1 – G15 data. Interestingly, while accuracies of SVM and LR models are comparable on generation G1 – G6 data, the SVM models are significantly more accurate on generations G7 – G15. Additionally, AUC accuracy of SVM predictors is significantly higher than the accuracy of baseline predictors on generations G6 – G15, with the AUC accuracy being up to 0.14 higher.

We also evaluated SVM with nonlinear kernels, such as polynomial kernel of degree 2, and radial basis function kernel. However, we did not observe significant accuracy improvements (results not shown).

Both LR and SVM predictors can be used to analyze the influence of each location cluster on infection spread. In Figure 1, we illustrate how to visualize the results and potentially gain an insight into properties of infection. In upper panel of Figure 1, we visualize each of the 20 clusters obtained by k-means clustering on LogNormLA transformed data. Each of the 33 rows represents number of people pursuing one of the 33 types of activity; each of the 20 columns is a representative of one of the 20 clusters. Dark areas represent highly populated (location, activity). For example, cluster 17 corresponds to colleges since most activity is of type “College”, it occurs between hours 7-24 (attributes 31-33), and there are many people present. 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, attributes 8 and 11, with the highest α values, seem to correspond to locations of predominantly residential type. This agrees with intuition that collocation with infected people within homes carries the largest risk of infection.

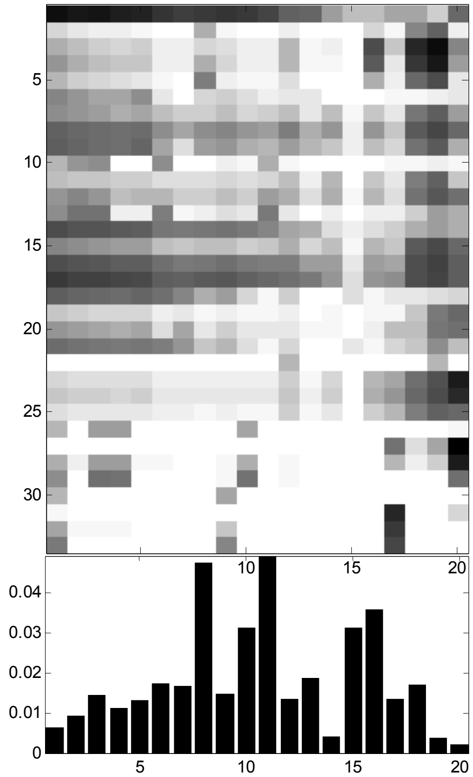

Figure 1. Illustration of location clusters and the associated risks.

4. Conclusions

Presented results indicate that the proposed location aggregation approach is very promising for prediction of infection risk. As compared to the baseline approach, the improvements in AUC accuracy ranged from 0.03 to 0.14. More importantly, it seems that accurate predictors could be developed using only first few generation of infected people.

Further improvements in location clustering are possible, possibly by integrating automatic 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