C-Watcher: A Framework for Early Detection of High-Risk Neighborhoods Ahead of Covid-19 Outbreak

Congxi Xiao[1,2*], Jingbo Zhou[2*], Jizhou Huang[2*], An Zhuo[2], Ji Liu[2], Haoyi Xiong[2], Dejing Dou[2]

[*]Equal contribution, [1]University of Science and Technology of China, [2]Baidu Inc., China

Brief Introduction

- Through extensive data analytics, we explore a set of empirical features related to long-term/regular human mobility patterns (before the COVID-19), which well characterize the socioeconomic and demographic status, as well as the spatial interactions among neighborhoods, to distinguish high-risk neighborhoods from the urban area and predict the potential risks.
- We propose C-Watcher, a COVID-19 risks early detecting method, which generalizes the know-ledge witted in epicenter to target cities without outbreak by adopting an adversarial encoder-decoder framework to learn the "city-invariant" representations from the explored features.
- We collect and construct real-world datasets from the web and conduct extensive experiments for the evaluation of high-risk neighborhoods detection . The results demonstrate the advantages of C-Watcher to early detect high-risk neighborhoods across cities.

Features for Neighborhood Detection

We construct 3 groups of features from human mobility data to characterize a residential neighborhood from the perspectives of demographic/socioeconomic status and spatial interactions for early detection.

POI Radius Features

We compute a group of Point of Interest (POI) radius for each residential neighborhood based on data from Baidu Maps. The POI radius is defined as the shortest distance between the given neighborhood and one certain type of POIs, such as hospital radius and train station radius. We also define an binary feature to directly represent the perfect degree of living facilities. It will be assign as "perfect" if a set of basic living facilities (e.g. hospital, bus stop and so on) are all with in 1km of the given neighborhood, or it is assigned as "poor". Figure 1(a) indicates that more high-risk neighborhoods have poor living facilities.

Demographic Features

Given that the COVID-19 is easy to transmit in a person-to-person way, we also take into account the demographic features of each neighborhood, such as population density. As Figure 1(b) illustrates, on average high-risk neighborhoods do have a higher population density than low-risk one in Wuhan city. Moreover, different groups of residents may face different risks levels in a neighborhood. Therefore, we construct 11 features based on the distribution of residents according to different human attributes, and present each of them as a vector of histogram statistics of residents distribution.

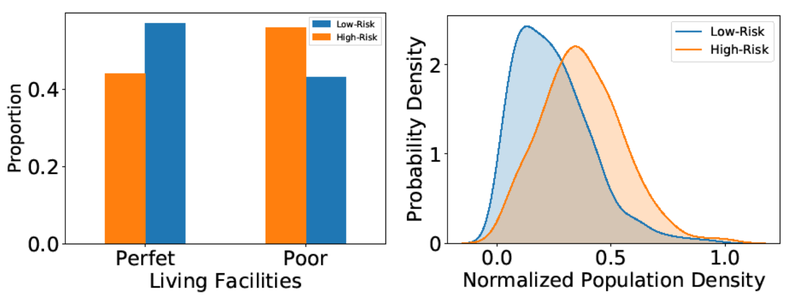

Figure 1: Features of living facilities and population density analysis.

Transportation-Related Features

We also extract features of transportation-related behaviors from human mobility data. The transportation-related behaviors are typically recognized as origin-transportation-destination (OTD) information. Thus, we construct features from the perspective of **T** (transportation means), **OD** (origin & destination venues features including types of visit venues and distance between origin and destination) and **OTD** (the top-20 hottest origin-transportation-destination travel patterns). All the features are extracted from Baidu Maps in a certain time period.

Cross-City Transfer Learning

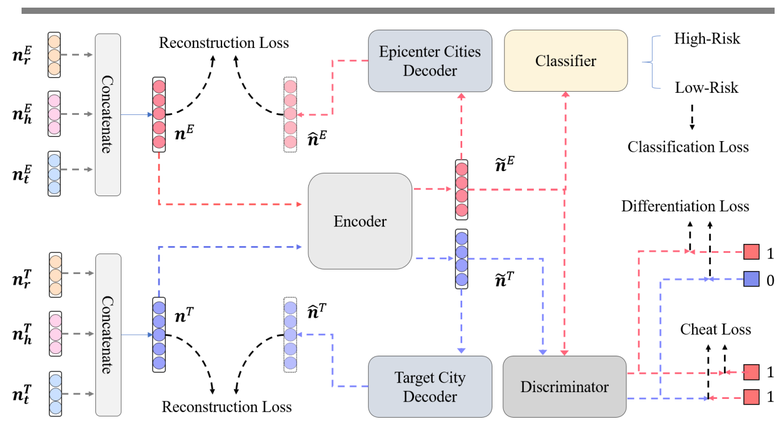

Figure 2: Illustration of cross-city transfer learning model.

The objective of C-Watcher is to make early detection of high-risk neighborhoods without epidemic outbreaks. The cross-city transfer learning is the core component of C-Watcher to improve the performance of early detection of high-risk neighborhoods in the target city via generalizing the knowledge about the COVID-19 infections witted in the epicenter to target cities.

City-Invariant Representation Learning

Discrepancies always exist between different cities and learning unique characteristics of neighborhoods in the epicenter brings little benefit for early detection of high-risk neighborhoods in target cities. Thus, we adopt adversarial learning to learn city-invariant knowledge applicable to both epicenter and target cities. As shown in Figure 2, the adversarial learning framework consists of an encoder and a discriminator. Given the inputs $\boldsymbol{n}^E$ and $\boldsymbol{n}^T$, which denote the features vector of a neighborhood from the epicenter and the target city, the encoder transforms them to encoded representations:

$$p(\tilde{\boldsymbol{n}}^E, \Phi_e) = \int_{\boldsymbol{n}^E} e(\tilde{\boldsymbol{n}}^E|\boldsymbol{n}^E, \Phi_e)\, p(\boldsymbol{n}^E)\, d\boldsymbol{n}^E$$

$$p(\tilde{\boldsymbol{n}}^T, \Phi_e) = \int_{\boldsymbol{n}^T} e(\tilde{\boldsymbol{n}}^T|\boldsymbol{n}^T, \Phi_e)\, p(\boldsymbol{n}^T)\, d\boldsymbol{n}^T$$

where $p(\cdot)$ denotes the data distribution and $e(\cdot, \Phi_e)$ denotes the encoder function with parameters Φ_e. $p(\tilde{\boldsymbol{n}}^E, \Phi_e)$ and $p(\tilde{\boldsymbol{n}}^T, \Phi_e)$ represent the distributions of encoded representations, which are not similar in general. The discriminator $D(\cdot)$ takes the encoded representations as inputs and distinguish whether the neighborhood comes from the epicenter or target city by classifying $\tilde{\boldsymbol{n}}^E$ as true but $\tilde{\boldsymbol{n}}^T$ as false. On the contrary, the encoder tries to confuse the discriminator to classify both of them as true by extracting common features from the 2 neighborhoods from different cities. Thus, the loss function consists of two parts:

$$\mathcal{L}_{diff} = -[\log(D(\tilde{\boldsymbol{n}}^E)) + \log(1 - D(\tilde{\boldsymbol{n}}^T))]$$

$$\mathcal{L}_{ch} = -[\log(D(\tilde{\boldsymbol{n}}^E)) + \log(D(\tilde{\boldsymbol{n}}^T))]$$

Embedding Space Constraints

To improve the high-risk neighborhoods detecting ability, we impose some constraints on embedding space of the encoder by exerting an auto encoder-decoder features reconstruction and a COVID-19 risks prediction components. The reconstruction component, which are designed to retain the information of a neighborhood, consists of two decoders $d(\cdot)$ for the epicenter and target city respectively. The decoders take $\tilde{\boldsymbol{n}}^E$ and $\tilde{\boldsymbol{n}}^T$ as inputs and the reconstruction loss function $\mathcal{L}_{rec}$ approximates the outputs to the original input vectors $\boldsymbol{n}^E$ and $\boldsymbol{n}^T$:

$$p(\hat{\boldsymbol{n}}^E, \Phi_d^E) = \int_{\tilde{\boldsymbol{n}}^E} d(\hat{\boldsymbol{n}}^E, |\tilde{\boldsymbol{n}}^E, \Phi_d^E)\, p(\tilde{\boldsymbol{n}}^E)\, d\tilde{\boldsymbol{n}}^E$$

$$p(\hat{\boldsymbol{n}}^T, \Phi_d^T) = \int_{\tilde{\boldsymbol{n}}^T} d(\hat{\boldsymbol{n}}^T, |\tilde{\boldsymbol{n}}^T, \Phi_d^T)\, p(\tilde{\boldsymbol{n}}^T)\, d\tilde{\boldsymbol{n}}^T$$

$$\mathcal{L}_{rec} = ||\hat{\boldsymbol{n}}^E, \boldsymbol{n}^E||_2 + ||\hat{\boldsymbol{n}}^T, \boldsymbol{n}^T||_2$$

The classifier $C(\cdot)$ takes $\tilde{\boldsymbol{n}}^E$ as input and classify whether the neighborhood is high-risk or low-risk, based on the binary label $y^E \in \{0,1\}$ that we collect on the web. The loss function of risks classification ensures that the encoded representations are instructive to high-risk neighborhoods identification.

$$\mathcal{L}_{cl} = -y^E \log(C(\tilde{\boldsymbol{n}}^E, \Phi_C)) - (1 - y^E)\log(1 - C(\tilde{\boldsymbol{n}}^E, \Phi_C))$$

Reference City Validation Mechanism

As shown in Figure 3, to ensure that our trained model to detect latent high-risk neighborhoods in a target city with best hyper-parameters, without any prior information related to COVID-19 confirmed cases and spreading trend, we train C-Watcher on epicenter cities, and use ground truth data in a reference city as validation data, then evaluate the performance in the target city geographically close to the reference city.

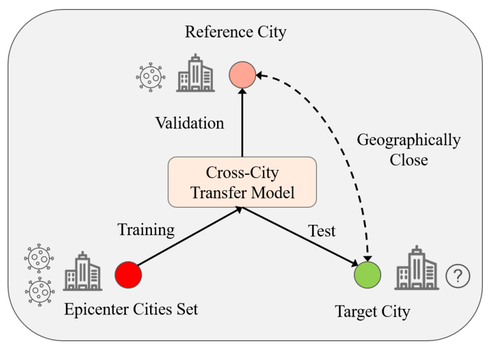

Figure 3: Diagram of reference city validation mechanism.

Experiments

Performance Evaluation of Early Detection

We compare the early detection performance of C-Watcher and baselines on test datasets of 10 target cities. The overall and part of the results are shown in Table 1. We can see that C-Watcher can improve the average AUC over the best baselines (SVM and MLP). The p-values demonstrate that C-Watcher can achieve significantly better performance than other baselines.

	Overall		Huizhou	Shaoyang	Lianyungang	Xuchang	Chongqing
	AUC	P-value	AUC				
SVM	0.5999	0.0005	0.7049	0.5615	0.6728	**0.7330**	0.5693
XGB	0.5810	0.0018	0.6266	0.5190	0.6182	0.7067	0.4901
Lasso-R	0.5853	0.0006	0.6364	0.5410	0.6515	0.7195	0.5263
MLP	0.5963	0.0005	0.6995	0.5594	0.6850	0.7278	0.5438
C-Watcher	**0.6490**	–	**0.7352**	**0.6433**	**0.7218**	0.7312	**0.6142**

Table 1: Early detection performance comparison.

Feature Importance

We conduct a feature importance analysis using Lasso Logistic Regression on epicenter Wuhan dataset and select the top-20 important features according to absolute values of coefficients, to discuss the possible characteristics of a neighborhood leading to the high risk. As shown in Figure 4, for POI radius features, except the effect of perfect and poor living facility (which are denoted by "P:PFLF" and "P:PRLF"), the coefficient of "P:RTS" indicates that long distance to train station can reduce the risks of neighborhoods. For the demographic features, except the high population density (denoted by "D:PD"), the long average commute distance (denoted by "D:ACD") also increases the risks. For transportation-related features, we find that the percentage of travelling on walk (denoted by "T:TW") can reduce the risk of the neighborhood by a large margin.

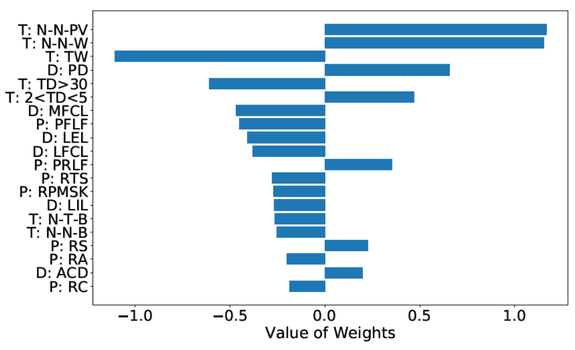

Figure 4: The top-20 most important features for risks prediction.

Effectiveness of Feature Groups

We verify the effectiveness of 3 groups of hand-crafted features by separately evaluating the performance of each group in detecting high/low-risk neighborhoods by MLP on the epicenter Wuhan dataset, and comparing them with taking all 3 groups together. The AUC results (POI radius features (0.8033), demographic features (0.7579), transportation-related features (0.7414) and all 3 groups (**0.8458**)) prove the positive effects of all the 3 feature groups and the significant complementary among them.