ARTICLE

Inverse problem solver for epidemiological geographic profiling

Yoshiharu Maeno[1] (ORCID)

Abstract

This work studies machine-learning-based inverse problem solvers for a reaction–diffusion process. The study focus is on the performance of a state-of-the-art convolutional neural network in discovering the source of disease spreading. This problem is called epidemiological geographic profiling. The performance is investigated with synthetic datasets for SIR epidemiological compartments on a square grid geospace. The convolutional neural network works effectively in discovering a single source and achieves the largest time average of accuracy for growing infection in a heterogeneous geo-space. The hit score remains near the lower bound over time. Discovering multiple sources is feasible potentially as well by learning the dataset for a single source.

Keywords Convolutional neural network · Geographic profiling · Infectious disease · Inverse problem · Reaction–diffusion process

JEL Classification C13 Estimation: General · C18 Methodological Issues: General · I10 General

1 Introduction

Spreading of an infectious disease is a stochastic reaction–diffusion process in a heterogeneous space. Infection of susceptible individuals is a reaction process in a subpopulation. Removal (or recovery) of infectious individuals is also a reaction process. Movement of infectious individuals is a diffusion process across subpopulation boundaries. A subpopulation is a group of individuals in a small distinct geographic

✉ Yoshiharu Maeno
  maenoy@meiji.ac.jp

1   School of Interdisciplinary Mathematical Sciences, Meiji University, Tokyo 1648525, Japan

🖄 Springer

area. Similar processes underlie many investigation topics in data-driven mathematical sciences, econophysics, and sociophysics. Some examples include dissemination of online fake-news (Vosoughi et al. 2018), spreading of rumors on social networking services (Grinberg et al. 2019), fluctuating prosperity of firms in an economic system (Maeno 2013), structural changes in the international trade network (Ikeda and Iyetomi 2018), and a crash of financial markets (Maskawa et al. 2013). Uncertainty in such processes results in the emergence of systemic risk (Kuyyamudi et al. 2019) and often imposes a large adverse impact on the entire system (Daglis et al. 2022).

Public health authorities are interested in the inception of the spreading. Once they detect a possible outbreak and get a big picture on the spatial distribution of the spreading, it becomes a pressing task for both the first-aid intervention and long-term policy making to discover the index case (Cesar Henrique and da Fontoura 2011). The index case is a so-called patient-zero as a source (individual or sub-population) of spreading. This problem is called epidemiological geographic profiling (Le Comber et al. 2011). It is inspired by criminological geographic profiling which aids crime investigators in searching for a home base of an offender from serial crime sites on a map (Canter et al. 2000). Similarly, it is a significant task for financial regulatory authorities to discover the source of circulating disinformation and transmitting financial distress in the markets. Discovery of a source is an inverse problem for identifying the initial state retrospectively from the current snapshot of variables.

In general, temporal evolution of time-dependent variables is formulated by a Langevin equation (stochastic differential equation). Equivalently, temporal evolution of the probability density function of stochastic variables is described by a Fokker-Planck forward equation (multivariate partial differential equation). The Fokker-Plank formulation is analogous to a Boltzmann transport equation in describing non-equilibrium transient behaviors. The Boltzmann transport equation formalizes collisions and diffusion of molecular species under external forces. Susceptible and infectious individuals in epidemiology are such molecular species in thermodynamics. Note that relaxation to equilibrium, or non-decreasing entropy in the second law of thermodynamics, is an informational process to lose the memory of the initial state irreversibly. Unless the spreading terminates in the equilibrium, the most probable candidate for the initial state can be obtained in principle by analyzing a corresponding Fokker-Planck backward equation. But practical reaction-diffusion processes are too complex to obtain a tractable solution even numerically.

This work studies machine-learning-based inverse problem solvers for epidemiological geographic profiling. The study focus is on the performance of a state-of-the-art convolutional neural network (CNN). The other 3 solvers for performance comparison include a naive Bayes classifier (NBC), a random forest classifier (RFC), and a multinomial logistic regression (MLR). They are one of the most widely used solvers in predicting probability for Bayesian inference, supervised learning, and statistical modeling respectively. For example, some works on molecular transports (Janczura et al. 2020; Kowalek et al. 2019) investigate a random forest classifier in detail. A trajectory of time-dependent variables is computed numerically from a Langevin equation for SIR epidemiological compartments on a square grid geo-space. Snapshots of variables

are recorded from the trajectory and organized as a synthetic dataset. The solvers are applied to discovering either a single source or multiple sources in the dataset. Accuracy and hit score are investigated to quantify the performance of the 4 solvers.

2 Related works

Recent works present that machine learning can play a vital role for data analysis in mathematical sciences. One example is approximating a solution of a complicated diffusion equation for polymer microphase separation (Wei et al. 2018). Another is classifying a diffusion mode in single particle tracking for molecular transports (Janczura et al. 2020; Kowalek et al. 2019). On the other hand, many works study global intercity aviation and human-to-human contacts as a medium of spreading of an infectious disease. The key idea is founded on modeling to discretize the medium as a complex, temporal, or multi-layered network (Li and Saad 2021; Ortega et al. 2022; Torrisi et al. 2021).

Some of the works propose to analyze network topologies and apply meta-heuristics to epidemiological geographic profiling (Menin and Bauch 2018; Nguyen and Vural 2017; Paluch et al. 2021; Shi et al. 2022). But few of them study machine-learning-based inverse problem solvers, which are applicable potentially to a multi-dimensional geo-space and any other models of the medium. This is the motivation of this work.

3 Materials and methods

3.1 Reaction–diffusion process for disease spreading

Time-dependent variables $I_j(t)$ $(j = 1, 2, \cdots, N)$ is the number of infectious individuals in the j-th subpopulation at continuous time $t \geq 0$. N is the number of subpopulations. J_0 is the set of source subpopulations. Temporal evolution of $I_j(t)$ is described by a Langevin equation (system of stochastic differential equations) in Eq. (1) (Maeno 2016).

$$
\begin{aligned}
\frac{dI_j(t)}{dt} =& (\alpha I_j(t) + \xi_j^{[\alpha]}(t)) - (\beta I_j(t) + \xi_j^{[\beta]}(t)) \\
&+ \sum_{k \neq j} (\gamma_{kj} I_k(t) + \xi_{kj}^{[\gamma]}(t)) - \sum_{k \neq j} (\gamma_{jk} I_j(t) + \xi_{jk}^{[\gamma]}(t)) \\
&(j = 1, 2, \ldots, N).
\end{aligned}
\tag{1}
$$

The coefficients α and β are the probability of infection and removal (or recovery) per unit time. These are the state transitions in SIR epidemiological compartments $(S \rightarrow I \rightarrow R)$. In the SIR epidemiological compartments, an individual is assigned to one of 3 compartments with labels S (susceptible), I (infectious), and R (removed (or recovered)). It is assumed in this work that $I_j(t) \ll S_j(t)$ which is the number of

susceptible individuals. The decrease in S_j is negligible. The term for infection becomes $\alpha' S_j I_j(t) \approx \alpha I_j(t)$ in Eq. (1) approximately. Basic reproductive ratio is given by $R = \alpha/\beta$. The Hurst exponent (Vilk 2022) for $I(t) = \sum_{j=1}^{N} I_j(t)$ is $h = 0.5$ (normal diffusion) if $\alpha = \beta$ ($R = 1$). The coefficients $\Gamma = \{\gamma_{ij}\}$ are the probability of unidirectional movements (from the j-th to the k-th subpopulation) per unit time. The probability of staying in the current subpopulation is $\gamma_{jj} = 1 - \sum_{k \neq j} \gamma_{jk}$. The terms $\xi_j^{[\alpha]}(t)$, $\xi_j^{[\beta]}(t)$, and $\xi_{jk}^{[\gamma]}(t)$ are Gaussian white noise. Their functional forms are not known.

Subpopulations are placed on a two-dimensional square grid geo-space. The size of the geo-space is 32×32 ($N = 1024$). A subpopulation is surrounded by 8 neighbors. Heterogeneity parameter r_{ht} is the ratio of missing connections. A missing connection means that the probability of movements is negligible there. The geo-space is sparser and more disjoint for larger r_{ht}. This work studies only non-disjoint connections. It is assumed that $\gamma_{jk} = \gamma$ is a constant for connections. In other words, the probability distribution is $P(\gamma_{jk} = 0) = r_{ht}$ and $P(\gamma_{jk} = \gamma) = 1 - r_{ht}$. The mean and standard deviation of γ_{jk} over j, k in Γ are $\mu = (1 - r_{ht})\gamma$ and $\sigma = \sqrt{(1 - r_{ht})r_{ht}}\gamma$.

3.2 Dataset for epidemiological geographic profiling

The Langevin equation in Eq. (1) is integrated with a Monte-Carlo simulation for a single source ($|J_0| = 1$). The initial state is $I_j(0) = 10^4$ for $j \in J_0$ and $I_j(0) = 0$ otherwise. Time t is discretized into a step Δt. Multinomial random number generators $x \sim \mathcal{M}(\alpha, \beta, \Gamma)$ reproduce stepwise state transitions and fluctuating changes in ΔI_j. Variable snapshot vectors $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)$ are recorded at an observation interval τ. The simulation conditions are $t \leq 200$ and $\tau = 10$ (20 snapshots per a trajectory). The snapshots $\boldsymbol{I}_j$ are the instances of N features with a source identifier $j \in J_0$ as a category label. The pairs of the features and category label are assembled into a training dataset $\mathcal{D}^{[tr]} = \{(\boldsymbol{I}_j, j)\}$. $\mathcal{D}^{[tr]}$ includes $28 \times 28 = 784$ different trajectories (initial states). The source excludes the edges of the geo-space ($1024 - 784 = 240$ subpopulations). The instances ($|\mathcal{D}^{[tr]}| = 20 \times 784 = 15680$) are stored in an arbitrary order. Note that $\mathcal{D}^{[tr]}$ is not informative explicitly on the parameters ($\alpha, \beta, \Gamma, t, \Delta t, \tau$).

A test dataset $\mathcal{D}^{[te]}$ is assembled similarly. The values of the probability parameters (α, β, Γ) are the same. There is either a single source ($|J_0| = 1$) or multiple sources ($|J_0| > 1$) initially. The sources are chosen randomly so that the distance between any pair in J_0 is larger than the mean (0.52 in a unit square) because they are indistinguishable if they are too close. The initial state is $I_j(0) = 10^4/|J_0|$ for $j \in J_0$. The observation interval is $\tau = 25$ (8 snapshots per a trajectory). Figure 1 shows some instances in $\mathcal{D}^{[te]}$ for $r_{ht} = 0.45$ and $R = 1.2$. Each time series (a), (b), (c) represents a trajectory ($t \leq 200$) for $|J_0| = 1$ (single source), 2 (multiple sources), and 3 (multiple sources). The trace of the initial state gets fainter gradually (correlation is lost) as fluctuation accumulates and

Fig. 1 Instances in a test dataset $\mathcal{D}^{[te]}$ for $r_{ht} = 0.45$ and $R = 1.2$ ($\alpha = 0.12, \beta = 0.1, \gamma = 0.1$) The number of sources is (a) $|J_0| = 1$, (b) $|J_0| = 2$, and (c) $|J_0| = 3$
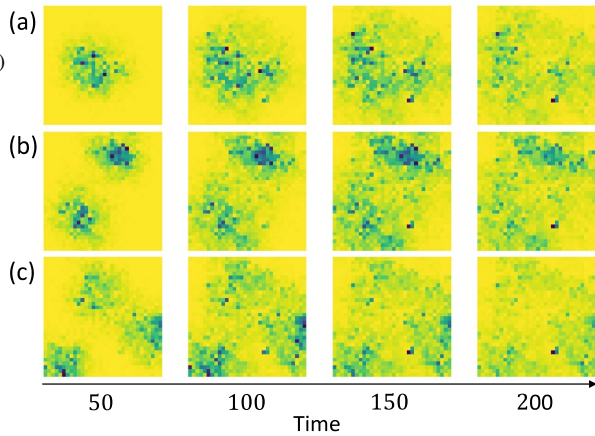
relaxation to the equilibrium progresses. They visualize how the signal level falls while the noise level rises.

3.3 Machine-learning-based inverse problem solver

For an instance I in $\mathcal{D}^{[te]}$, $p_j = P(j \in J_0|I)$ is the probability that the j-th subpopulation is a source. The vector $\boldsymbol{p} = (p_1, p_2, \cdots, p_N)$ is a posterior probability. The role of a machine-learning-based solver is to find the optimal mapping function $\boldsymbol{p} = f(\boldsymbol{I})$ with respect to $\mathcal{D}^{[tr]}$ and predict the source j_0 for $\mathcal{D}^{[te]}$ by eq.(2).

$$j_0 = \underset{1 \leq j \leq N}{\operatorname{argmin}} \, p_j. \tag{2}$$

Prediction succeeds if $j_0 \in J_0$. An accuracy measure A is defined as the number of successful predictions as a fraction of $|\mathcal{D}^{[te]}|$. A hit score measure H is a cumulative search area, which is investigated according to the ranking of p_j in a descending order to discover the source, as a fraction of the total area (Le Comber et al. 2011). The area is measured by the number of grids. Larger accuracy and smaller hit score indicate more effective prediction. The most effective prediction achieves $A = 1$ for $|J_0| = 1$ and $H = |J_0|/N$ for $|J_0| \geq 1$. Mere guesswork results in $H = |J_0|/(|J_0| + 1)$ regardless of N.

A convolutional neural network (CNN) is a state-of-the-art deep learning algorithm (Kowalek et al. 2019; Wei et al. 2018). I and $\boldsymbol{p}$ are processed as if they were the input pixels in a monochrome image (32×32 ($N = 1024$)) and the output likeliness that the image contains a particular image object ($C = 1024$ categories). Figure 2 shows the configuration of an 8-layer convolutional neural network ($L = 8$) in this work. Large kernels ($K = 8 \times 8$) are applied in calculating convolution. Filters decide the number of parallel channels ($F = 32, 64$) in the output. A convolutional neural network with 10 or more hidden layers works as an excellent image classifier empirically. Based on the rule of thumb, the configuration in Fig. 2 is decided by an exhaustive search to improve A and H around $L = 10$, $K = 10$, and $F = 10 \sim 10^2$.

L, K, and F are kept as small as possible. Note that the configuration is not proven the best but verified effective experimentally. The hidden layers 2 through 7 employ leaky rectified linear functions. The activation and output layer employs a soft-max function $f(\boldsymbol{I}) \propto \exp(\sum_{k=1}^{N} a_{jk} I_k + b_j)$ where a_{jk} and b_j are weight parameters. The number of weight parameters is $W_{\mathrm{CNN}} = \mathcal{O}(CLKF)$. $W_{\mathrm{CNN}} \sim 10^6$ if $L, K, F \sim 10$. In total, 7.2×10^5 weight parameters in Fig. 2 are optimized by an adaptive-moment algorithm with cross-entropy $E = -\sum_{j=1}^{N} p_j \log p_j$ as a loss function.

A naive Bayes classifier (NBC) is based on a maximal likelihood estimation. The likelihood $L(j \in J_0) = P(\boldsymbol{I} | j \in J_0)$ (Li and Saad 2021; Torrisi et al. 2021) cannot be computed reliably unless $\mathcal{D}^{[\mathrm{tr}]}$ is dense enough. Instead in this work, it is evaluated approximately by $L(j \in J_0) \propto \max_j \boldsymbol{I} \cdot \boldsymbol{I}_j$ where $\boldsymbol{I}_j \in \mathcal{D}^{[\mathrm{tr}]}$. It does not have any weight parameters. Optimization of weight parameters is omitted, but maximization of $\boldsymbol{I} \cdot \boldsymbol{I}_j$ for prediction consumes more time as $|\mathcal{D}^{[\mathrm{tr}]}|$ increases. A random forest classifier (RFC) is a bootstrap aggregation algorithm for an ensemble of decision tree classifiers. An ensemble of $T = 100$ decision tree classifiers is optimized with respect to cross-entropy. The number of weight parameters is $W_{\mathrm{RFC}} = \mathcal{O}(CT) \sim 10^5$. It is nearly in the same order as W_{CNN} in Fig. 2. A multinomial logistic regression (MLR) is a generalized linear model with a soft-max function. The weight parameters are optimized by a Newton conjugate gradient algorithm. The number of weight parameters is $W_{\mathrm{MLR}} = \mathcal{O}(C^2) \sim 10^6$. It is as large as W_{CNN}.
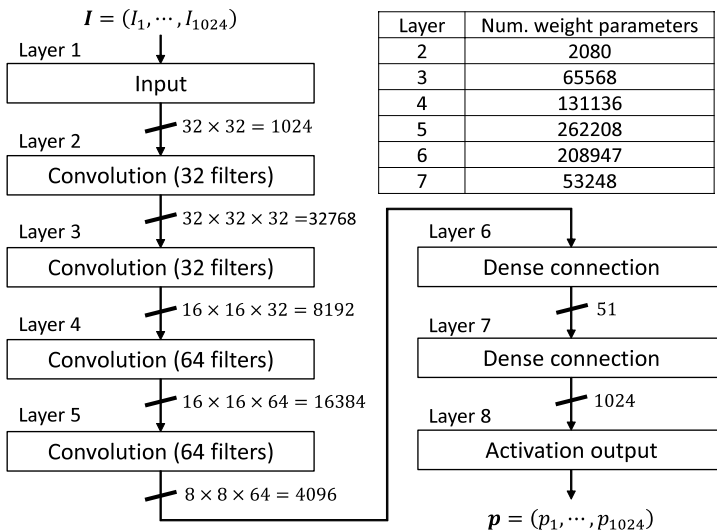

Fig. 2 Configuration of an 8-layer convolutional neural network (CNN). The hidden layers 2 through 7 have a number of weight parameters, which are shown in the inset table

4 Results and discussion

4.1 Discovery of single source

Table 1 shows the time average of accuracy $\bar{A}$ (over the all instances in $\mathcal{D}^{[\text{te}]}$) of the 4 machine-learning-based solvers (CNN, NBC, RFC, MLR) for discovering a single source ($|J_0| = 1$). The geo-space is either homogeneous ($r_{\text{ht}} = 0.05$) or heterogeneous ($r_{\text{ht}} = 0.45$). Infection is either growing ($R = 1.2$), steady ($R = 1$), or declining ($R = 0.9$). The combinations of the values of r_{ht} and R are referred to by the 6 test cases **a**, **b**, $\cdots$, **f**. Note that the solvers do not rely on any prior information on SIR epidemiological compartments, probability parameters, and heterogeneity of a space ($\alpha, \beta, \Gamma, t, \Delta t, \tau, r_{\text{ht}}$) in executing optimization and prediction.

The convolutional neural network achieves $\bar{A} \geq 0.85$ for growing infection (test cases **a** and **d**) and outperforms the others. Growing infection in a heterogeneous geo-space (test case **d**) is of much practical significance. The convolutional neural network works most effectively ($\bar{A} = 0.88$) in this test case. The performance deteriorates to $\bar{A} = 0.51$ for declining infection (test case **f**). The signal gets less conspicuous. But it still achieves the best performance. The multinomial logistic regression fails to work in contrast. The solvers must be capable of learning and expressing a complex mapping function $\boldsymbol{p} = f(\boldsymbol{I})$. The convolutional neural network performs a variety of non-linear computations for any correlated subpopulations in each hidden layer while the others iterate similar computations for the all subpopulations. The naive Bayes classifier and random forest classifier work moderately. They can be an alternative for the test cases **d** and **b**, **c** respectively. But the performance tends to be unstable on a case-by-case basis. It may be difficult to decide confidently when to apply them if r_{ht} and R are not known. The convolutional neural network is not the only choice but the first choice for public health practitioners in investigating a variety of cases.

It is also verified that the convolutional neural network works comprehensively for Γ in other probability distributions ($\neq \Gamma$ in 3.1). For example, the time average of accuracy in the test case **d** is $\bar{A} = 0.91$ for a uniform distribution $\gamma_{jk} \sim \mathcal{U}(0, 2\mu)$, where $\mu = (1 - r_{\text{ht}})\gamma$ and $\sigma = \mu/\sqrt{3}$, and $\bar{A} = 0.93$ for a normal distribution $\gamma_{jk} \sim \mathcal{N}(\mu, \mu/2)$. An interesting finding is that the convolutional

Table 1 Time average of accuracy $\bar{A}$ of the 4 machine-learning-based solvers (CNN, NBC, RFC, MLR)	Test case	Parameter value		Solver			
		r_{ht}	R	CNN	NBC	RFC	MLR
	a	0.05	1.2	0.85	0.76	0.77	0.50
	b	0.05	1	0.49	0.41	0.45	0.22
	c	0.05	0.9	0.38	0.31	0.35	0.17
	d	0.45	1.2	0.88	0.84	0.57	0.56
	e	0.45	1	0.58	0.53	0.27	0.33
	f	0.45	0.9	0.51	0.41	0.19	0.28

neural network works more accurately for heterogeneous geo-spaces. It may ensue because the speed of time evolution is different. The probability of movements is a constant for connections. But the nodal degree is smaller and the hop-count between subpopulations is larger for larger r_{ht}. The effective flux is smaller, the consequent relaxation time is larger, and the trace of the initial state may be kept more recognizable. This may no longer be true after relaxation progresses.

Figure 3 shows the accuracy $A(t)$ as a function of time t ($\tau = 25$). The scatter plots (a), (b), $\cdots$, (f) correspond to the 6 test cases **a**, **b**, $\cdots$, **f** in Table 1. The 4 curves in a scatter plot represent the 4 solvers (CNN, NBC, RFC, MLR). The convolutional neural network almost always works most effectively of the 4 solvers. Note that the naive Bayes classifier works as effectively in the test case **d**, and the random forest classifier works slightly more effectively at $t = 150$ through 200 in the test cases **b**, **c**. These solvers can be alternatives under particular conditions of spreading. In terms of the overall performance in a variety of test cases, it is concluded that the convolutional neural network is still the first choice, and the naive Bayes classifier can be the second choice for helping the practitioners verity
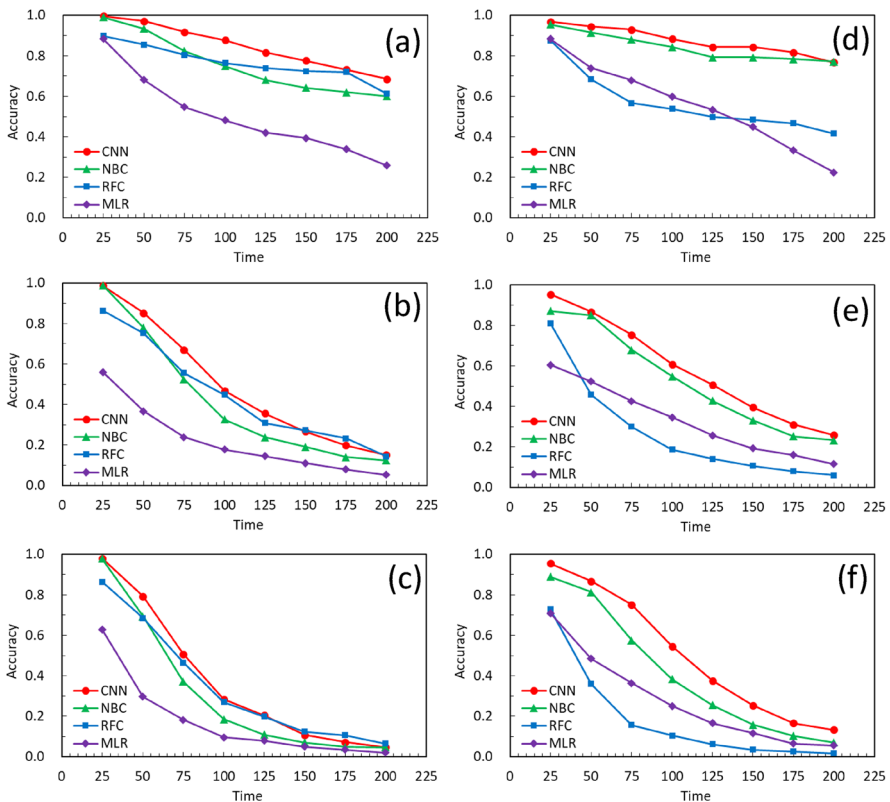

Fig. 3 Accuracy $A(t)$ as a function of time t of the 4 solvers for discovering a single source. The scatter plots (a), (b), $\cdots$, (f) correspond to the 6 test cases **a**, **b**, $\cdots$, **f** in Table 1

the prediction. Performance degradation for declining infection (test cases **c** and **f**) at $t \geq 100$ is evident as the correlation with the initial state $\boldsymbol{I}(0)$ is almost lost (normalized autocorrelation coefficient $\rho(t) \leq 0.05$).

4.2 Discovery of multiple sources

Figure 4 shows the hit score $H(t)$ of a convolutional neural network as a function of time t for discovering a single source ($|J_0| = 1$ for the scatter plot (a)) and multiple sources ($|J_0| = 2, 3$ for the scatter plots (b), (c)). The test cases **a**, **b**, $\cdots$, **f** are the same as in Table 1 and Fig. 3. In the scatter plot (a), the hit scores $H(t) \leq 1.5 \times 10^{-3}$ remains near the lower bound ($1/N = 9.8 \times 10^{-4}$) over time for growing infection in both homogeneous and heterogeneous geo-spaces (test cases **a** and **d**). It means that the second or third candidates from $\boldsymbol{p}$ are promising nearly equally even if the first candidate j_0 in Eq. (2) fails and impairs the accuracy A.

The performance characteristics in the scatter plots (b) and (c) are different from those in the scatter plot (a). The performance worsens as $|J_0|$ increases. But the hit score H is still below 0.01 and 0.1 in the early stage of disease spreading, and converges steadily to the upper bound 0.67 and 0.75 for $|J_0| = 2$ and 3 respectively. The solver always works more effectively for a heterogeneous geo-space (test cases **d**, **e**, and **f**). Note that such instances as in Fig. 1b and c are absent in $\mathcal{D}^{[\text{tr}]}$ and the mapping function $\boldsymbol{p} = f(\boldsymbol{I})$ is optimized with respect to $\mathcal{D}^{[\text{tr}]}$. The number of patterns for the initial state $\mathcal{O}(N^{|J_0|})$ is too large for $|J_0| > 1$. Those findings imply that discovery of small number of sources is potentially
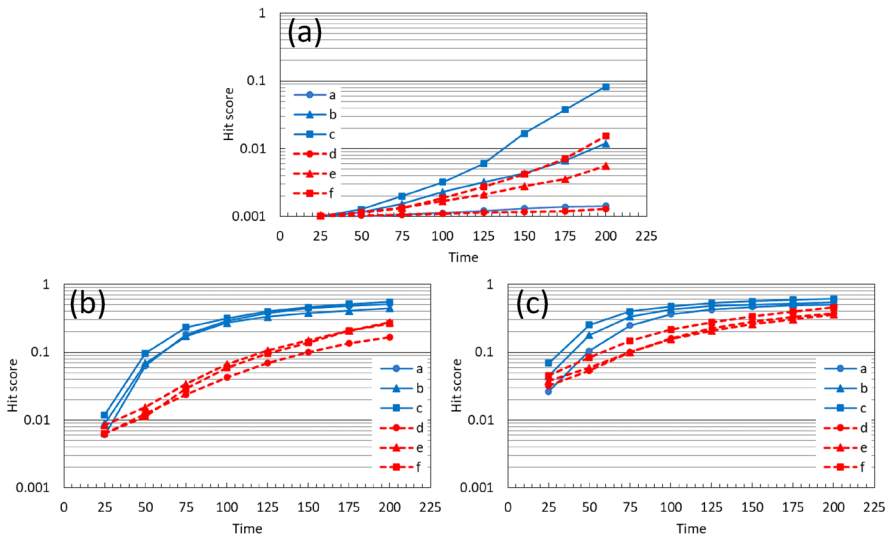

Fig. 4 Hit score $H(t)$ of a convolutional neural network as a function of time t. The scatter plots **a**–**c** represent the hit score for $|J_0| = 1, 2$, and 3. The 6 curves in each scatter plot correspond to the 6 test cases **a**, **b**, $\cdots$, **f** in Table 1 and Fig. 3

achievable, at least in the early stage of spreading, by learning a limited number of instances for a single source.

5 Conclusion

In this work, it is demonstrated that a convolutional neural network works effectively as an inverse problem solver for epidemiological geographic profiling. In terms of the time average of accuracy $\bar{A}$, it outperforms the other 3 machine-learning-based solvers in any test cases of growing, steady, and declining infection in heterogeneous and homogeneous geo-spaces. The largest accuracy ($\bar{A} \geq 0.85$)) is achieved and the hit score ($H \leq 1.5 \times 10^{-3}$) remains near the lower bound over time in discovering a single source ($|J_0| = 1$) for the most significant test case (growing infection in a heterogeneous geo-space). Discovering multiple sources ($|J_0| > 1$) is feasible potentially as well merely by learning of a limited number of instances in the dataset for a single source.

It is anticipated that growing infection in a heterogeneous geo-space generates a unique complex spatio-temporal pattern in the trajectories. The complex pattern may be left as a hidden signal on the source after a simple linear pattern (correlation) is lost in the noise. It is known as a profound design principle that the complexity of a solver is managed to match that of the dataset. Therefore, exploring less complex meta-heuristics as in some previous is a futile effort. Those findings corroborate the conclusion that the complex nature of a convolutional neural network deserves the first-choice solver for detecting such a hidden signal. Note that the number of hidden layers and filters, and the kernel settings are arbitrarily flexible, and the configuration in this work is not proven the best. Performance limits with respect to the nature of inverse problems and the capability of solvers are yet to resolve as future theoretical, experimental, and empirical works.

Recently, large datasets become available by collecting real-time data from sensors in every corner of human activities. Then in turn, machine learning aids investigators in identifying the initial state (and even boundary conditions, external forces, stochastic terms, and equational forms possibly) and in solving miscellaneous inverse problems for a reaction-diffusion process. For example, the inverse problem solvers in this work can be applied to analyzing such an economic system as a financial market, a stock exchange, a supply chain, and a trade network. The price of a stock or a group of stocks in an industry sector is a time-dependent variable (similar to the variables $I_j(t)$ in 3.1). Variable snapshot vectors represent the entire market conditions (similar to the instances I in the datasets $\mathcal{D}^{[\mathrm{tr}]}$ and $\mathcal{D}^{[\mathrm{te}]}$ in 3.2). Retrospective nowcasting is an inverse problem to discover the recent source incident on a stock which causes a big impact on the current market volatility. Such an analysis is beneficial to regulatory authorities, financial institutions, and investors in making a decision to deal with systemic risk, financial crisis, and a crash of markets. An inductive approach as in this work can be an essential investigation tool for data-driven mathematical sciences, econophysics, and sociophysics, and may replace conventional reductionistic modeling approaches.

Declarations

Conflict of interest The author declares that the author has no Conflict of interest.

References

Canter D, Coffey T, Huntley M, Missen C (2000) Predicting serial killers' home base using a decision support system. J Quant Criminol 16:457

Cesar Henrique C, da Fontoura CL (2011) Identifying the starting point of a spreading process in complex networks. Phys Rev E 84:056105

Daglis T, Melissaropoulos IG, Konstantakis KN, Michaelides PG (2022) The impact of COVID-19 on global stock markets: early linear and non-linear evidence for Italy. Evolut Inst Econ Rev 19:485

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. Science 363:374

Ikeda Y, Iyetomi H (2018) Trade network reconstruction and simulation with changes in trade policy. Evolut Inst Econ Rev 15:495

Janczura J, Kowalek P, Loch-Olszewska H, Szwabinski J, Weron A (2020) Classification of particle trajectories in living cells: machine learning versus statistical testing hypothesis for fractional anomalous diffusion. Phys Rev E 102:032402

Kowalek P, Loch-Olszewska H, Szwabinski J (2019) Classification of diffusion modes in single-particle tracking data: feature-based versus deep-learning approach. Phys Rev E 100:032410

Kuyyamudi C, Chakrabarti AS, Sinha S (2019) Emergence of frustration signals systemic risk. Phys Rev E 99:052306

Le Comber SC, Rossmo DK, Hassan AN, Fuller DO, Beier JC (2011) Geographic profiling as a novel spatial tool for targeting infectious disease control. Int J Health Geogr 10:35

Li B, Saad D (2021) Impact of pre symptomatic transmission on epidemic spreading in contact networks: a dynamic message-passing analysis. Phys Rev E 103:052303

Maeno Y (2013) Transient fluctuation of the prosperity of firms in a network economy. Phys A 392:3351

Maeno Y (2016) Detecting a trend change in cross-border epidemic transmission. Phys A 457:73

Maskawa J, Murai J, Kuroda K (2013) Market-wide price co-movement around crashes in the Tokyo Stock Exchange. Evolut Inst Econ Rev 10:81

Menin OH, Bauch CT (2018) Solving the patient zero inverse problem by using generalized simulated annealing. Phys A 490:1513

Nguyen VAT, Vural DC (2017) Morphological inversion of complex diffusion. Phys Rev E 96:032314

Ortega E, Machado D, Lage-Castellanos A (2022) Dynamics of epidemics from cavity master equations: susceptible-infectious-susceptible models. Phys Rev E 105:024308

Paluch R, Gajewski LG, Suchecki K, Holyst JA (2021) Impact of interactions between layers on source localization in multilayer networks. Phys A 582:126238

Shi C, Zhang Q, Chu T (2022) Source estimation in continuous-time diffusion networks via incomplete observation. Phys A 592:126843

Torrisi G, Annibale A, Kühn R (2021) Overcoming the complexity barrier of the dynamic message-passing method in networks with fat-tailed degree distributions. Phys Rev E 104:045313

Vilk O et al (2022) Unraveling the origins of anomalous diffusion: from molecules to migrating storks. Phys Rev Res 4:033055

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146

Wei Q, Jiang Y, Chen JZY (2018) Machine-learning solver for modified diffusion equations. Phys Rev E 98:053304

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.