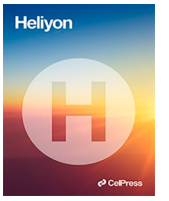
Heliyon 10 (2024) e33332

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Semi-supervised urban haze pollution prediction based on multi-source heterogeneous data

Zuhan Liu [a,*], Lili Wang [b]

[a] School of Information Engineering, Nanchang Institute of Technology, Nanchang, China
[b] College of Science, Nanchang Institute of Technology, Nanchang, China

ARTICLE INFO

Keywords:
Haze pollution
$PM_{2.5}$
Semi-supervised learning
Air quality prediction
Co-training
Tri-training

ABSTRACT

Particulate matter (PM) is defined by the Texas Commission on Environmental Quality (TCEQ) as "a mixture of solid particles and liquid droplets found in the air". These particles vary widely in size. Those particles that are less than 2.5 μm in aerodynamic diameter are known as Particulate Matter 2.5 or $PM_{2.5}$. Urban haze pollution represented by $PM_{2.5}$ is becoming serious, so air pollution monitoring is very important. However, due to high cost, the number of air monitoring stations is limited. Our work focuses on integrating multi-source heterogeneous data of Nanchang, China, which includes Taxi track, human mobility, Road networks, Points of Interest (POIs), Meteorology (e.g., temperature, dew point, humidity, wind speed, wind direction, atmospheric pressure, weather activity, weather conditions) and $PM_{2.5}$ forecast data of air monitoring stations. This research presents an innovative approach to air quality prediction by integrating the above data sets from various sources and utilizing diverse architectures in Nanchang City, China. So for that, semi-supervised learning techniques will be used, namely collaborative training algorithm Co-Training (Co-T), who further adjusting algorithm Tri-Training (Tri-T). The objective is to accurately estimate haze pollution by integrating and using these multi-source heterogeneous data. We achieved this for the first time by employing a semi-supervised co-training strategy to accurately estimate pollution levels after applying the U-air system to environmental data. In particular, the algorithm of U-Air system is reproduced on these highly diverse heterogeneous data of Nanchang City, and the semi-supervised learning Co-T and Tri-T are used to conduct more detailed urban haze pollution prediction. Compared with Co-T, which train time classifier (TC) and subspace classifier (SC) respectively from the separated spatio-temporal perspective, the Tri-T is more accurate with a and faster because of its testing accuracy up to 85.62 %. The forecast results also present the potential of the city multi-source heterogeneous data and the effectiveness of the semi-supervised learning. We hope that this synthesis will motivate atmospheric environmental officials, scientists, and environmentalists in China to explore machine learning technology for controlling the discharge of pollutants and environmental management.

1. Introduction

In the 21st century, China's economy has been developing continuously and rapidly, whose the process of industrialization and

* Corresponding author.
E-mail address: lzh512@nit.edu.cn (Z. Liu).

https://doi.org/10.1016/j.heliyon.2024.e33332
Received 24 February 2024; Received in revised form 29 May 2024; Accepted 19 June 2024
Available online 19 June 2024

urbanization has been significantly accelerated. However, the extensive economic development mode, large-scale construction and the increase of energy consumption have caused many social and environmental problems, one of which attracts particular attention is the urban air pollution [1,2]. Among them, the haze pollution with fine Particulate Matter or $PM_{2.5}$ as the main pollutant is becoming more and more serious. The coverage of haze weather continues to expand, resulting in the increasing pressure of urban atmospheric environment, which seriously restricts the sustainable development of China's economy and damages people's health [3,4]. In view of this, the Chinese government has invested a lot of resources to establish more than 1500 air pollution monitoring stations to dynamically record and release air pollutants in real time since 2012. However, it is not surprising that the number of air monitoring stations is small due to extremely high upkeep costs and land & manpower costs. There are only 9 stations in Nanchang City, about $44km^2$/station, which are far from satisfying in air monitoring. Therefore, how to give more accurate air quality prediction without increasing the number of stations has become one of the long-term goals of environmentalists.

In actual study, people have found that there are various factors affecting haze pollution, such as traffic, meteorology, industrial and human activities, which may affect the air quality in a large or small scale [5]. Actually, the complex causes and the forecast provided by air monitoring stations in small volume for which have brought great challenges to the real-time monitoring of air quality in most situations [6,7]. For haze pollution, its main components are constantly changing. In particular, the development of industrial technology has resulted in the particulate sediments, such as $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2 and CO, which produced by busy human activities and traffic flows in cities become the main force of haze pollution [8]. The methods of haze pollution prediction are mainly divided into two categories: model simulation and data-driven techniques. However, these models, such as GWR [9,10], PS-FCM [11], WRF [12] and AOD [2,13,14] require many parameters, which are difficult to obtain and hard to satisfy practical demand. As for data-driven techniques, they are also further classified into traditional numerical and machine learning methods. Typical methods of the former are spatial interpolation and Land Use Regression (LUR) model. Actual cases, such as Sampson et al. (2013) successfully predicted $PM_{2.5}$ at a fine spatial scale across the U.S. using regionalized Ordinary Kriging method [15]. Furthermore, Hasenfratz et al. (2015) utilized both time-series and simulation models to estimate vehicle contributions to pollutant levels near roadways [16]. However, spatial interpolation only consider limited geographical factors, and cannot cope with the complicated and nonlinear air quality estimation problem. And LUR model can describe the temporal and spatial changes of air quality, but it is complex and there is no standard construction method. Therefore, while LUR has been extensively used to capture the spatial distribution of air pollution, regional background and nonlinear relationships can be challenging to capture using linear approaches [17].

With the development of smart devices and cloud computing, more and more public data may be collected from various sources and analyzed in an unprecedented way [18]. Especially, Urban Big Data (UBD) is a natural data source, and many researchers naturally turn to machine learning, which has recently been used in the forecast air quality. Such as Deep Brief Network (DBN), Recurrent Neural Network (RNN), Random Forest (RF), Particle Swarm-based Fuzzy C-Means (PS-FCM) model, Online Recurrent Extreme Learning Machine (OR-ELM) technique and other machine learning methods applied to timely prevention of haze pollution [11,13,17,19–21]. For the same pollutants, many of the above works have presented that the performance of machine learning is superior to that of traditional numerical methods. At the same time, it is also proved that the latter could benefit from understanding how explanatory variables were expressed in machine learning models [20]. Furthermore, UBD has made great efforts in all aspects related to smart cities, such as taxi flow prediction [17–24], urban function division [25], traffic route planning [26], etc. Duo to different sources and structures, multi-source fusion technology of how to integrate and utilize these data is an important content for big data research. Jiang et al. (2018) made an attempt for semi-supervised urban air quality prediction based on multi-source heterogeneous data, whose prediction results also showed its potential and effectiveness of the semi-supervised learning [27]. Actually, integrating these relevant yet heterogeneous models can provide complementary predictive powers by combining the expertise of heterogeneous data. Moreover, they are used to address data sparsity issues about single infrastructures [28].

In addition, the machine learning framework naturally fits air quality prediction problem with a small amount of stations and UBD, especially semi-supervised learning. What's more, the most typical one is the Urban Air (U-Air) system developed by Zheng et al. (2013) of Microsoft Research Asia (MSRA) based on the joint training framework in 2013 [29]. Specifically, the framework uses a large number of unlabeled data related to air quality for assistance, which can obtain a prediction accuracy more than of 80 %. Essentially, the algorithm is to enable different learners to learn from each other via iterative methods based on a large number of unlabeled samples and a small number of labeled samples. However, the results of two classifiers obtained by the Co-T framework are always inconsistent, which brings difficulties to the practical application of U-Air system; actually, this embodies the Black-Box nature [30].

Rapid economic development and the sharp increasement of the population and vehicle puts forward new requirements to Nanchang's environmental protection. In the meantime, these aspects also present challenges to environmental monitoring technologies [31,32]. In view of this, we motivate and design an urban haze pollution prediction of semi-supervised learning different from the traditional prediction model based on single source and structure data in this paper. It establishes a unified prediction system based on multi-source heterogeneous data combined with specific algorithms. In the forecasting process, we implement five heterogeneous data based on track data of 10792 taxis, 24809 human mobility, 31-day meteorology, a 2470-road network and POIs including 110 intersections, 1495 bus stops and 243 gas stations in Nanchang City. The U-Air algorithm is reproduced on these highly diverse heterogeneous data of Nanchang City, and a comparative study between Co-T and Tri-T based on MLP and SVM through training TC and SC is carried out to verify and compare their prediction accuracy and training speed. The experimental simulation shows that Tri-T has higher accuracy and faster training. This work provides new insights into the data limits facing cities in terms of the predictive control strategy of haze pollution. The research significance lies in that it can integrate the characteristics of various data and make complement of various data for maximum $PM_{2.5}$ prediction accuracy, so as to provide the most accurate pollution warning for the public. The aim of the thesis is to provide a reference for people's healthy travel and social activities, and improve people's living standard.

2. Data source and feature extraction

The data set used in this paper is collected from Nanchang City, which are various from source or format. Among them, all these data in the area where the air monitoring station is located are learned as labeled samples, and other data in the area without the air monitoring station except the stations data are unlabeled samples.

2.1. District partition

This study was conducted in the mid-subtropical zone in Nanchang City (115°27′-116°35′E, 28°09′-29°11′N). As the capital of Jiangxi Province of China, Nanchang City is in the process of rapid urbanization. With the popularity of the car, the air pollution problems caused by traditional car should not be ignored. The experiment in this paper will be carried out on its central urban area of Nanchang City (28°36′-28°44′N, 115°42′-115°58′E). According to the longitude of 0.02′and latitude of 0.01′ interval, it is divided into 64 grid cells of size 32.6 × 18.5 m as shown in Fig. 1. In the Fig. 1, 9 red areas have built air quality monitoring stations, so the collected samples from these areas are labeled, and the unlabeled samples are from the other 55 areas. Therefore, there are 3888 labeled and 23760 unlabeled samples. In such cases, if done successfully, this survey would greatly improve the performance of learning by avoiding much expensive data-labeling efforts [29].

7-Petrochemical station (Suburb); 22-Xianghu station (Xihu District); 41-Wushu School station (Wanli District); 45-Construction Engineering School station (New Town District); 46-Provincial Foreign Affairs Office station (Old Town District); 47-Provincial station (Qingshanhu District); 48-Jingdong Town Government station (High-tech Development Zone); 55-Forestry Company station (Changbei Investment Company Industrial Park); 60-Institute of Economic Forestry station (Economic& Technology Development Zone).

2.2. The data of air monitoring stations

There are six types of air pollutants by Nanchang air monitoring stations in this paper, namely $PM_{2.5}$, PM_{10}, O_3, SO_2, NO_2 and CO, which were obtained from Jiangxi Provincial Environmental Monitoring Center Station, China. The increase in atmospheric pollution dominated $PM_{2.5}$ has become one of the most serious environmental hazards worldwide [10,33], o only the concentration of $PM_{2.5}$ (namely $\rho(PM_{2.5})$) is predicted in this work. The concentration limit of daily average $PM_{2.5}$ and their respective corresponding AQI are displayed in Table 1. This paper has hourly collected the $PM_{2.5}$ data of Nanchang City in July 2022, which marked the AQI level is excellent and medium, accounting for the vast majority in July alone. In other words, there rarely appears in the value of AQI or the daily average $PM_{2.5}$ smaller than 100 or 75 μg/m^3, respectively.

2.3. Traffic flow and their features F_t

With the rapid development of cities and increase in automobile ownership, traffic has become one of the main sources of $PM_{2.5}$ pollution [33]. Traffic flow data mainly refers to the data generated by moving process of vehicles. In this study, the track data of 10792 taxis is used to describe F_t in Nanchang City in July 2022, which are collected by Python to climb the Baidu map and Amap.

Traffic data of Nanchang City include taxi number, time, longitude, latitude and passenger status, etc. The sampling interval of trajectory data is from 1 to 30 s. The extracted features from traffic data include the expectation, standard deviation and distribution of the taxi speed. To elaborate a bit on that, the expectation can be inferred that the slower the average speed of vehicle, the greater the possibility of congestion, the more harmful gases emitted by vehicles in the congested space, and the air quality tends to deteriorate. The standard deviation could effectively reflect the difference in the vehicle speed as an aid to expectations. Large variance represents a scattered speed distribution, small probability of congestion and better air quality. The distribution is divided into three sections at 20 km/h and 40 km/h, and the probability that the taxi speed falls in each section is counted. For the convenience of calculation, the travel time in this speed range is used instead of the total time. Traffic flow can judge whether the urban road traffic is congested and

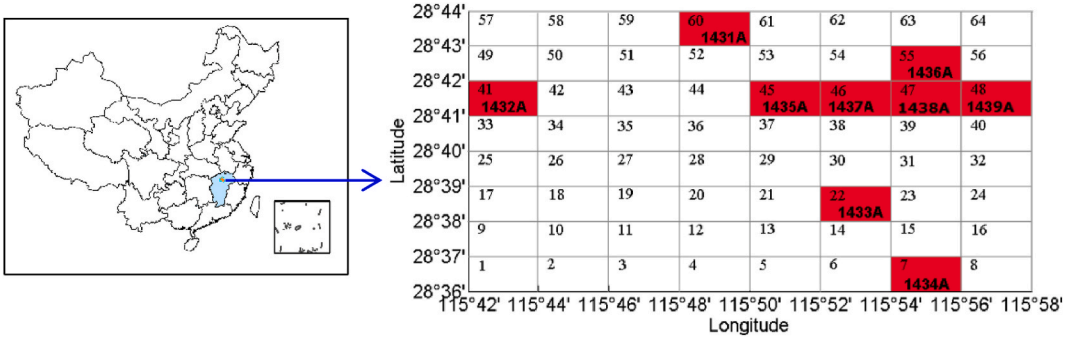


Fig. 1. Distribution of air quality monitoring station in Nanchang.

Table 1
The concentration limit of daily average $PM_{2.5}$ and their corresponding AQI.

$\rho(PM_{2.5})/\mu g/m^3$	AQI	AQI level	Health effects
$0 < \rho(PM_{2.5}) \leq 35$	$0 < AQI \leq 50$	Excellent	The air quality is satisfactory. There is basically no air pollution. A wide array of people can be normal activities
$35 < \rho(PM_{2.5}) \leq 75$	$50 < AQI \leq 100$	Good	The air quality is acceptable, but some pollutants may have a weak impact on the health for a handful of exceptionally sensitive people.
$75 < \rho(PM_{2.5}) \leq 115$	$100 < AQI \leq 150$	Slight pollution	The sensitive people are mildly aggravated. The irritation symptoms appear in healthy people.
$115 < \rho(PM_{2.5}) \leq 150$	$150 < AQI \leq 200$	Moderate pollution	Further aggravate the symptoms of susceptible people, which may affect the heart and respiratory system of healthy people.
$150 < \rho(PM_{2.5}) \leq 250$	$200 < AQI \leq 300$	Heavy pollution	The symptoms of patients with heart disease and lung disease are significantly aggravated, and the exercise endurance is reduced. The symptoms are common in healthy people.
$\rho(PM_{2.5}) > 250$	$AQI > 300$	Serious pollution	The exercise tolerance of healthy people is reduced, with obvious and strong symptoms, and some diseases appear in advance.

then mitigate it. Notwithstanding, mitigating traffic congestion on urban roads is paramount importance in urban development and reduction of energy consumption and air pollution.31.

2.4. Traffic flow and human mobility features F_h

F_h can also be extracted from taxi track, that is, human mobility. If a taxi stays at a certain point for a long time, it could be considered that human movement has occurred here. In fact, it is a mining of taxi track. Furthermore, if there are many people coming and going to some area in a certain period of time, there may be more POIs here. To be more specific, if there are many people coming and leaving, there may be a shopping center; otherwise this could be a factory or school here. We can use F_h from taxis track to calculate the number of people entering and leaving small areas. In order to simply the calculation, when passengers enter the area (namely, the passenger status is indicated by 1), the number of people entering the area plus 1; conversely, the number of departures plus 1. When the passenger status is 0 in sympathy with a taxi enters or leaves the area, the flow of people in and out has not changed.

2.5. Meteorology and their features F_m

F_m includes temperature, dew point, relative humidity, wind speed, wind direction, atmospheric pressure, weather activity and weather conditions, etc. Obviously, high wind speed will lead to easy dispersion of pollutants and low concentration, but high humidity then otherwise. High atmospheric pressure generally benefit from air quality, while the influence of temperature on it was not obvious. In particular, the air quality is generally good at high temperature and low humidity, and it is also excellent under the condition of high pressure and low temperature. In addition, the influence of adverse meteorological conditions will further intensify the accumulation of secondary aerosols and other pollutants and promote the explosive growth of $PM_{2.5}$/CO values [34].

2.6. Road networks and their features F_r

Urban road networks, one of the crucial infrastructures in cities, facilitate the people's daily commutes and maintain modern society's ability to function properly [35,36]. F_r can be obtained according to SHP format map. Different roads have a great impact on the speed limit and normal driving speed of vehicles. Generally, the vehicles can travel smoothly and at a higher speed on the road with high specifications and many lanes. It is not easy to block and generate a large amount of exhaust gas, and so what happens is the impact on the environment is small [37]. The road characteristics that could be extracted directly from the road networks include the length of roads at all levels and the length of other roads.

2.7. POIs and their features F_p

The Points of Interest (POIs) refer to places where there are usually many people, such as school, factory and park [38,39]. These POIs belong to the urban open public spaces are the areas where people tend to gather together, which may lead to impact on different air quality. For example, parks and schools with better greening can bring positive benefits to air quality, but factories are just the opposite. The POIs are classified into three types: intersection, bus stop and gas station, then their number can be counted inside these zones.

F_p comes from Baidu API, which is a developer-friendly platform and provides a series of map services for free. However, the statistical process is cumbersome, and the returned results are not consistent every time, so the types of POIs are limited.

3. Methods

The learning framework used in this study includes Co-T and Tri-T, both of which belong to semi-supervised learning method based

on the difference 0.7.

3.1. Co-Training (Co-T)

In 1998, Blum and Mitchell proposed Co-T for Webpage classification, which has attracted wide attention [40]. Many new application scenarios and expansion based on the Co-T framework emerge in endlessly [41]. However, the core framework of Co-T has never changed. The sample datasets are divided into two parts on the premise of a research problem with two sufficient perspectives; then the supervised samples are used to train the initial classifier, which classify the unlabeled samples. Thus, the data with high confidence is selected from the classification results as the training set of the classifier, and Reference 26 also did a similar study. The team took traffic flow and human activities as important causes of air quality pollution, in which they applied Co-T framework to prediction using U-Air method and achieved a breakthrough success.

There is something should be specify pointed out in the algorithm framework, that is, the TC and SC tend to use a linear -chain Conditional Random Field (CRF) [42–44] and a common Multi-Layer Perception (MLP) neural network [45], respectively. MLP neural network, for example, its input is the air quality grade of the nearest three monitoring stations from the area, the distance from the area to be predicted, and the geographical correlation. Furthermore, the correlation is measured by their correlation coefficient between road networks and human POIs.

The Co-T algorithm framework for air quality prediction is shown in Fig. 2.

3.2. Tri-Training (Tri-T)

The Tri-T algorithm was proposed by Zhou and Li (2005) [42]. Unlike Co-T algorithm, there is only need to ensure the difference between classifiers without multiple perspectives. It was originally designed to solve the binary classification problem, but not be limited to. For an n classification, it can completely expand the number of classifiers to ensure the correctness of majority voting. Due to the uniqueness of $PM_{2.5}$ data, there is no need to expand the number of classifiers, but it should be noted in other applications.

In the Tri-T algorithm framework, the conditions to be met in the loop are to ensure that the noise brought by the new pseudo labeled samples will not degrade the training results, so as to ensure the convergence and robustness of the training process [27].

How to learn from unlabeled samples with noise is the most compelling part for Tri-T algorithm. The added noise from the new samples will bring negative effects, but it can be eliminated as long as certain conditions are satisfied [46]. That is, for a data set with m samples, if

Input: labeled data $L(F_t,F_h,F_m,F_r,F_p,label)$
unlabeled data $U(F_t,F_h,F_m,F_r,F_p)$
Output: TC, SC
1. **function** Co-T (L,U)
2. **for** $F_t,F_h,F_m,label$ **do**
3. TC←CRF algorithm
4. **end for**
5. **for** F_r,F_p **do**
6. SC←MLP
7. repeat
8. **for** be well trained TC **do**
9. classify the U with TC
10. **end for**
11. **for** be well trained SC **do**
12. Classify the U with TC
13. **end for**
14. Select the pseudo labeled sample L' with the highest confidence to enter the training set
15. **for** $F_t,F_h,F_m,label$ in $L \cup L'$ **do**
16. TC←CRF algorithm
17. **end for**
18. **for** F_r,F_p in $L \cup L'$ **do**
19. SC←MLP
20. repeat
21. **if** the maximum number of cycles is reached or there is no new pseudo labeled samples
22. **End**

Fig. 2. Co-T algorithm.

$$m \geq \frac{2}{\varepsilon(1-2\eta)^2} In\left(\frac{2N}{\delta}\right) \tag{1}$$

where ε, η, N and δ are error rate of classification, sample noise rate, number of classifiers and classification confidence, respectively, there will be

$$P(d(H,H*) \geq \varepsilon)) \leq \delta \tag{2}$$

where H and H^* are approximate classified and actual plane respectively, which means that as long as the number of samples is enough, H can be infinitely close to H^*. The parameter c instead of $2In\left(2N/\delta\right)$ in formula (1) that it will make the equality in the previous inequality hold. Moreover, the equation is simply processed to obtain

$$u = \frac{c}{\varepsilon^2} = m(1-2\eta)^2 \tag{3}$$

The Tri-T algorithm framework for air quality prediction is shown in Fig. 3.

We assume that the pseudo labeled samples selected in the tth and (t-1)th round in the training process are L^t and L^{t-1}, respectively, namely $m^t = |L^t|$ and $m^{t-1} = |L^{t-1}|$. It is further assumed that L labeled samples with η_L noise rate of L. For the ith classifier ($i = 1,2,3$), the misclassification rate of the other two classifiers is e^t in round tth; for $m^t = |L \cup L^t|$ samples, there will be ($\eta_L L + |L^t| e^t$) misclassified. Therefore, the noise rate of the ith classifiers in

Round tth is

$$\eta^t = \frac{\eta_L L + |L^t| e^t}{|L \cup L^t|} \tag{4}$$

where e^t is represented by the proportion of pairs in the results of the other two classifiers that are consistent with L classification. Then the training sample noise rate of the ith classifiers in the round tth and (t-1)th are respectively

$$u^t = m^t(1-\eta^t)^2 \tag{5}$$

$$u^{t-1} = m^{t-1}\left(1-2\eta^{t-1}\right)^2 \tag{6}$$

Due to u is inversely proportional to ε, therefore, the training process of Tri-T should ensure that $u^t < u^{t-1}$, i. e

$$|L \cup L^t|\left(1-2\frac{\eta_L L + |L^t| e^t}{|L \cup L^t|}\right) > |L \cup L^{t-1}|\left(1-2\frac{\eta_L L + |L^{t-1}| e^{t-1}}{|L \cup L^{t-1}|}\right) \tag{7}$$

Input: labeled data $L(F_1,F_2,F_m,F_r,F_p,label)$;
unlabeled data U (F_1,F_2,F_m,F_r,F_p)
classifier algorithm C
m random samples from L to form S_i
Output: classifier h_1, h_2, h_3

```
function Tri-T (L,U,C)
    for each h_i do
        initial classifier h_i←S_i
    end for
    set e=(0.5,0.5,0.5), l=(0,0,0)
    repeat the following operations until L_i no longer change
    for each h_i do
        Let L_i=NULL,update_i=0
        Calculate E_i=Measure Error(h_j&h_k)
        Under the condition of E_i<e_i
        for each sample x in U do
        Select samples with the same classification of h_j and h_k to add L_i
    end if
    if l_i=0
        Put l_i in (E_i/(e_i-E_i)+1)
    if l_i<L_i
        if E_i*|L_i|<e_i*|l_i|
        update_i=1
    if l_i >E_i/(e_i-E_i)
        Random samples from E_i/(e_i-E_i)  instead of L_i
    update_i=1
    if update_i=1 after the above operation
        h_i←(S_i∪L_i) according to the corresponding algorithm
    e_i=E_i, l_i=L_i
    end
End of repeat
```

Fig. 3. Tri-T algorithm.

Obviously, if $|L^t|>|L^{t-1}|$ and $|L^t|e^t<|L^{t-1}|e^{t-1}$, the conditions are satisfied. However, in the case of $e^t < e^{t-1}$ and $|L^t|\gg|L^{t-1}|$, then it doesn't work. At this time, it needs sample L^t, then

$$s=\frac{|L^{t-1}|e^{t-1}}{e^t}-1 \tag{8}$$

when it is satisfied

$$|L^{t-1}| > \frac{e^t}{e^{t-1}-e^t} \tag{9}$$

$s>|L^{t-1}|$ still holds.

4. Experiment and results

First we will carry on different combination of features to test their testing accuracies, such as F_m, $F_t + F_h$, $F_r + F_p$, $F_m + F_t + F_h$ and $F_m + F_t + F_h + F_r + F_p$. Their testing accuracies are 68.52, 77.91, 69.78, 78.25 and 79.80, respectively. It is pretty obvious that the accuracy is increased via a combination of heterogeneous data features. Then experiment simulation of Co-T and Tri-T will be presented.

4.1. The experimental results of Co-T

This work uses a frequently-used performance evaluation methods for classifiers to test their accuracy, namely *Accuracy* [47]. That is

$$\text{Accuracy}=\frac{C}{T} \tag{10}$$

where C and T are the number of correctly classified samples and the total number of samples predicted by the system respectively.

The training process of Co-T is illustrated in Fig. 4. TC and SC are trained respectively from the separated spatio-temporal perspectives. Their test results and two classifiers trained by cyclic training using Co-T are shown in Table 2. The blue and red lines represent the training process of the spatial and temporal classifiers respectively. The best performance point occurs when the number of cycles is 13; this is, the model in this point is chosen as well trained classifiers. Compared with separated training and the single use of a certain classifier from double perspective, the performance of classification has been significantly improved.

In Tables 2 and it can also be seen that the training classifying quality of joint use of the same classifier algorithm from dual perspectives has not improved much, and the spatial classifier has even regressed. However, the testing performance of the TC has been greatly improved via the Co-T training, while the SC's performance has not regressed in spite of a small improvement. This is because the SC with higher accuracy has higher confidence in the pseudo labeled samples selected for TC in the process of joint training, and thus would be improved the classification performance of TC. However, the samples selected by TC don't significantly improve SC.

Although the TC has made great progress for $PM_{2.5}$ level prediction with multi-classification, it is not easy to prediction results of the two combined classifiers. In the experiment, we found that two results are often different. By multiplying, adding and rounding their testing accuracy of two classifiers, the further acquired testing accuracies are similar to that of the SC. The U-Air system is to multiply the probability that the two classifiers judges whether the sample belongs to a certain class, and take the class with the largest probability as the prediction result. Limited to the experimental conditions, this step cannot be realized in this work.

4.2. The experimental results of Tri-T

In this section, we provide a detailed comparison of effects on pure-supervised learning and combined Tri-T. As in Table 3, the prediction accuracy of Tri-T based on MLP network is up to 84.28 %, which is 5.02 % higher than that of MLP network with only multi-

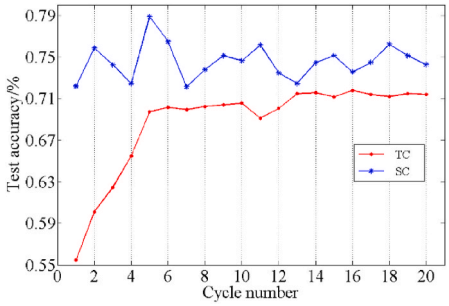


Fig. 4. Training process of Co-T.

Table 2
Comparison between Co-T and their corresponding algorithms.

Classifier algorithm	CRF	MLP
the testing accuracy under the spatio-temporal separation (%)	55.48	72.18
the testing accuracy under the spatio-temporal combination (%)	60.32	71.89
the testing accuracy under the spatio-temporal combination and after the training of Co-T (%)	67.54	73.60
the algorithm increase (%)	21.32	1.33

layer perceptron. However, Tri-T with Libsvm has the same testing accuracy in the linear kernel function. Under the polynomial and RBF kernel function, their accuracy decreased from 84.07 % to 83.18 %–83.63 % and 82.76 % respectively. While the initial classifier train same samples and kernel functions of classifiers are different from each other, its accuracy is improved by 2.22 % compared with 83.24 % of the linear kernel function.

Notice that the division of the training and test set is just the reason why the MLP with pure-supervised learning is better than that of spatio-temporal MLP algorithm. In order to divide temporal and spatial perspectives, Co-T algorithm should consider the continuity in time, or unable to divide randomly. If the training and the test datasets are uneven, this will induce a worse test effect.

Table 4 illustrates the performance of Tri-T with three SVM kernel functions and their combination. As a note here, the Tri-T's joint modal and evaluation index are voting majority and testing accuracy separately. For a Tri-T with some kernel SVM, its initial set is from labeled training set,; while for Tri-T with different kernel SVM, its initial set is formed with all labeled training set to keep the difference using different learning algorithms. It illustrations in Tables 2 and 3 that the testing accuracy of the Tri-Twith MLP network is higher than that of pure-supervised learning. After introducing the classifiers' diversity, some samples with high confidence are indeed found from the unsupervised data set, thus we can expand the supervised learning data set and improve the testing accuracy.

Meanwhile, for the Tri-T combined with SVM, the prediction accuracy gradual increases under some supervised samples training the same kernel function, directed training using all labeled samples and all samples training different kernel functions. The reason is that the difference brought by the Tri-T is not enough to offset the negative effects with the decrease of sample quantity, when the proportion of labeled samples is not high and their dimension is high. However, the different trained kernel SVM by all samples not only makes use of all labeled samples, but also introduces the difference among classifiers. Beyond that, it underscores that the training speed of the Tri-T is much faster than that of the Co-T. To explain, the Co-T uses cross validation to select samples with high confidence, while the Tri-T indirectly solves such problems via collaborative classification of several classifiers.

5. Conclusion

In this work, we design and implement semi-supervised urban haze pollution prediction model to effectively integrate heterogeneous data based on five multi-source data related to haze pollution. The algorithm of U-Air system is reproduced on these highly diverse heterogeneous data of Nanchang City. And the Tri-T was used to solve the problem of urban haze pollution level prediction, and achieved good results. The method based on different kernel functions can achieve up to 85.62 % accuracy. Compared with Co-T, the Tri-T is easier to realize in haze pollution prediction without worrying about the division of feature sets to ensure double perspectives, and the classifier algorithm is much easier to change and its training is much faster. Moreover, the theoretical basis of the Tri-T also ensures own robustness. The empirical results show that multi-source heterogeneous data based on different data related to haze pollution can provide a different yet complimentary view for the same urban phenomenon. Thus, an effective integration of them would boost the model performance.

However, there are still many difficulties and challenges ahead that (i) some indirectly relevant infrastructure data is difficult to be obtained or not detailed enough; (ii) the combination and selection of feature data are not deep enough. In the future, more data need to be collected for urban air quality prediction. Moreover, the Tri-T algorithm based on Long Short Term Memory (LSTM) [48,49] and the introduction of past time information to assist prediction or the use of Tri-net may make the prediction more accurate [50,51].

Inclusion and diversity

We support inclusive, diverse, and equitable conduct of research.

CRediT authorship contribution statement

Zuhan Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Funding acquisition, Data curation, Conceptualization. **Lili Wang:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 3
Comparison of effects on pure-supervised learning and Tri-T.

The Experiment	Combination
supervised learning MLP	80.38
combination of MLP and Tri-T	84.28
linear kernel function SVM	83.24
combination of linear SVM and Tri-T	83.24
polynomial kernel SVM	84.07
combination of polynomial SVM and Tri-T	83.63
RBF kernel function SVM	83.18
combination of RBF kernel SVM and Tri-T	82.76

Table 4
The results of Tri-T based on SVM.

Classifier	SVM1	SVM2	SVM3	Combination
linear kernel test (%)	83.14	83.14	83.14	83.14
polynomial kernel test (%)	81.27	83.49	83.37	83.49
RBF kernel test (%)	83.37	83.14	83.81	83.93
different kernel test (%)	83.49	83.14	83.52	85.62

Funding

This study was supported by the National Natural Science Foundation of China (42261077).

References

[1] L. Jin, J.S.S.L. Apte, Miller, S. Tao, S.X. Wang, G.B. Jiang, X.D. Li, Global endeavors to address the health effects of urban air pollution, Environ. Sci. Technol. 56 (2022) 6793–6798, https://doi.org/10.1021/acs.est.2c02627.
[2] L.R. Yin, L. Wang, W.Z. Huang, J.W. Tian, S. Liu, B. Yang, W.F. Zheng, Haze grading using the convolutional neural networks, Atmosphere 13 (2022) 522, https://doi.org/10.3390/atmos13040522.
[3] C. Silveira, P. Roebeling, M. Lopes, J. Ferreira, S. Costa, J.P. Teixeira, C. Borrego, A.I. Miranda, Assessment of health benefits related to air quality improvement strategies in urban areas: an impact pathway approach, J. Environ. Manag. 183 (2016) 694–702, https://doi.org/10.1016/j.jenvman.2016.08.079.
[4] Y.Y. Wang, L. Huang, C.H. Huang, J.L. Hu, M. Wang, High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city, Environ. Int. 172 (2023) 107752, https://doi.org/10.1016/j.envint.2023.107752.
[5] N.C. Chen, M.J. Yang, W.Y. Du, M. Huang, $PM_{2.5}$ Estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution MAIAC AOD in Hubei, China, ISPRS Int. J. Geo-Inf. 10 (2021) 31, https://doi.org/10.3390/ijgi10010031.
[6] G.N. Geng, Y.X. Zheng, Q. Zhang, T. Xue, H.Y. Zhao, D. Tong, B. Zheng, M. Li, F. Liu, C.P. Hong, K.B.S.J. He, Davis, Drivers of $PM_{2.5}$ air pollution deaths in China 2002–2017, Nat. Geosci. 14 (2021) 645–650, https://doi.org/10.1038/s41561-021-00792-3.
[7] H.A. Perillo, B.M. Broderick, L.W. Gill, A. McNabola, P. Kumar, J. Gallagher, Spatiotemporal representativeness of air pollution monitoring in Dublin, Ireland, Sci. Total Environ. 827 (2022) 154299, https://doi.org/10.1016/j.scitotenv.2022.154299.
[8] H.W. Lin, M. Chen, Y.J. Gao, Z.Q. Wang, F.G. Jin, Tussilagone protects acute lung injury from $PM_{2.5}$ via alleviating hif-1α/NF-κB-mediated inflammatory response, Environ. Toxicol. 37 (2022) 1198–1210, https://doi.org/10.1002/tox.23476.
[9] X.Y. Pan, J.J. Tang, T.J. Yu, J.M. Cai, Y. Xiong, F. Gao, Reposition optimization in the free-floating bike-sharing system considering transferring travels from urban rail transit, Comput. Ind. Eng. 178 (2023) 109127, https://doi.org/10.1016/j.cie.2023.109127.
[10] J.W. TuTu, S.H. Tedders, Spatial variations in the associations of term birth weight with ambient air pollution in Georgia, USA, Environ. Int. 92–93 (2016) 146–156, https://doi.org/10.1016/j.envint.2016.04.005.
[11] Z. Peng, W.Q. Liu, S.J. An, Haze Pollution causality mining and prediction based on multi-dimensional time series with PS-FCM, Inf. Sci. 523 (2020) 307–317, https://doi.org/10.1016/j.ins.2020.03.012.
[12] R.T. Liu, Z.W. Han, The effects of anthropogenic heat release on urban meteorology and implication for haze pollution in the Beijing-Tianjin-Hebei region, Adv. Meteorol. 2016 (2016) 6178308, https://doi.org/10.1155/2016/6178308.
[13] Y.J. Li, S. Cakmak, J.P. Zhu, Profiles and monthly variations of selected volatile organic compounds in indoor air in canadian homes: results of canadian national indoor air survey 2012-2013, Environ. Int. 126 (2019) 134–144, https://doi.org/10.1016/j.envint.2019.02.035.
[14] Q.Q. He, W.H. Wang, Y.M. Song, M. Zhang, B. Huang, Spatiotemporal high-resolution imputation modeling of aerosol optical depth for investigating its full-coverage variation in China from 2003 to 2020, Atmos. Res. 281 (2023) 106481, https://doi.org/10.1016/j.atmosres.2022.106481.
[15] P.D. Sampson, M. Richards, A.A. Szpiro, S. Bergen, L. Sheppard, T.V. Larson, J.D. Kaufman, A regionalized national universal Kriging model using partial least squares regression for estimating annual $PM_{2.5}$ concentrations in epidemiology, Atmos. Environ. 75 (2013) 383–392, https://doi.org/10.1016/j.atmosenv.2013.04.015.
[16] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, L. Thiele, Deriving high-resolution urban air pollution maps using mobile sensor nodes, Pervasive Mob, Comput. Times 16 (2015) 268–285, https://doi.org/10.1016/j.pmcj.2014.11.008.
[17] W.J. Wang, P.L. Tian, J.H. Zhang, E. Agathokleous, L. Xiao, T. Koike, H.M. Wang, X.Y. He, Big data-based urban greenness in Chinese megalopolises and possible contribution to air quality control, Sci. Total Environ. 824 (2022) 153834, https://doi.org/10.1016/j.scitotenv.2022.153834.
[18] T. Huang, L. Lan, X.X. Fang, P. An, J.X. Min, F.D. Wang, Promises and challenges of big data computing in health sciences, Big Data Res 2 (2015) 2–11, https://doi.org/10.1016/j.bdr.2015.02.002.
[19] L.L. Hou, Q.L. Dai, C.B. Song, B.W. Liu, F.Z. Guo, T.J. Dai, B.S. Liu, X.H. Bi, Y.F. Zhang, Y.C. Feng, Revealing drivers of gaze pollution by explainable machine learning, Environ. Sci. Technol. Lett. 9 (2022) 112–119, https://doi.org/10.1021/acs.estlett.1c00865.
[20] K.L. Shang, Z.Y. Chen, Z.X. Liu, L.H. Song, W.F. Zheng, B. Yang, S. Liu, L.R. Yin, Haze prediction model using deep recurrent neural network, Atmosphere 12 (12) (2021) 1625, https://doi.org/10.3390/atmos12121625.
[21] J.W. Tian, Y. Liu, W.F. Zheng, L.R. Yin, Smog prediction based on the deep belief-BP neural network model (DBN-BP), Urban Clim. 41 (2022) 101078, https://doi.org/10.1016/j.uclim.2021.101078.

[22] A. Wang, J.S. Xu, R. Tu, M. Saleh, M. Hatzopoulou, Potential of machine learning for prediction of traffic related air pollution, Transport, Res. D-Tr. E. 88 (2020) 102599, https://doi.org/10.1016/j.trd.2020.102599.
[23] Y. Kang, B. Yang, H. Li, T. Chen, Y.C. Zhang, Deep Spatio-temporal modified-inception with dilated convolution networks for citywide crowd flows prediction, Int. J. Pattern Recogn. 34 (2019) 2052003, https://doi.org/10.1142/S0218001420520035.
[24] J.B. Zhang, Y. Zheng, D.K. Qi, Deep Spatio-temporal residual networks for citywide crowd flows prediction, Proc. 3rd Int. AAAI Conf. Artif. Intell. (2017) 1655–1661, https://doi.org/10.48550/arXiv.1610.00081.
[25] C.Y. Ma, W.Y. Li, Discovering functional regions in modern cities by using user check-in records and POIs, IEEE Int. Conf. Robotics Biomimetics (ROBIO) 4 (2019) 509–514, https://doi.org/10.1109/ROBIO49542.2019.8961674, 2019.
[26] S. Ma, Y. Zheng, O. Wolfson, Real-time city-scale taxi ride sharing, IEEE Trans. Knowl. Data Eng. 27 (2015) 1782–1795, https://doi.org/10.1109/TKDE.2014.2334313.
[27] W.L. Jiang, C.P. Lv, H.G. Wang, Semi-supervised urban air quality prediction based on multi-source heterogeneous data. The China Automation Congress 2018 (CAC2018), 2018, pp. 235–241 (Chinese).
[28] D.S. Zhang, J.J. Zhao, H.J. Lee, S.H. Son, Heterogeneous model integration for multi-source urban infrastructure data, in: ACM Transactions on Cyber-Physical Systems, vol. 1, 2016, https://doi.org/10.1145/2967503. Article 4.
[29] Y. Zheng, F. Liu, H.P. Hsieh, U-Air: when urban air quality inference meets big data, Proc. ACM SIGKDD'13. ACM. (2013) 1436–1444, https://doi.org/10.1145/2487575.2488188.
[30] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459, https://doi.org/10.1038/nature14541.
[31] L. Hohenberger, W.W. Che, Y.X. Sun, J.C.H. Fung, A.K.H. Lau, Assessment of the impact of sensor error on the representativeness of population exposure to urban air pollutants, Environ. Int. 165 (2022) 107329, https://doi.org/10.1016/j.envint.2022.107329.
[32] L. Espeholt, S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gazen, R. Carver, M. Andrychow-icz, J. HickeyBell, N. Aaron Kalchbrenner, Deep learning for twelve hour precipitation forecasts, Nat. Commun. 13 (2022) 5145, https://doi.org/10.1038/s41467-022-32483-x.
[33] G.D. Thurston, Fossil fuel combustion and PM2.5 mass air pollution associations with mortality, Environ. Int. 160 (2022) 107066, https://doi.org/10.1016/j.envint.2021.107066.
[34] J.Y. Mao Wang, J. Y. Li, Z. Xiong, W.X. Wang, M. Perc, Predictability of road traffic and congestion in urban areas, PLoS One 10 (2015) e0121825, https://doi.org/10.48550/arXiv.1407.1871.
[35] L.Y. Wu, X.Y. Zhang, J.Y. Sun, Y. Wang, J.T. Zhong, Z.Y. Meng, Intensified wintertime secondary inorganic aerosol formation during heavy haze pollution episodes (HPEs) in Beijing, China, J. Environ. Sci.-China 114 (2022) 503–513, https://doi.org/10.1016/j.jes.2022.01.008.
[36] M. Batty, The Size, Scale, and shape of cities, Science 319 (2008) 769–771, https://doi.org/10.1126/science.1151419.
[37] L. Ferguson, J. Taylor, M. Davies, C. Shrubsole, P. Symonds, S. Dimitroulopoulou, Exposure to indoor air pollution across socio-economic groups in high-income countries: a scoping review of the literature and a modelling methodology, Environ. Int. 143 (2020) 105748, https://doi.org/10.1016/j.envint.2020.105748.
[38] W. Yang, T.H. Ai, POI information enhancement using crowdsourcing vehicle trace data and social media data: a case study of gas station, ISPRS Int. J. Geo-Inf. 7 (2018) 178, https://doi.org/10.3390/ijgi7050178.
[39] G. Lamprianidis, D. Skoutas, G. Papatheodorou, D. Pfoser, Extraction, Integration and analysis of crowdsourced points of interest from multiple web sources. Proc. 3rd ACM SIGSPATIAL Int. Workshop on Crowdsourced and Volunteered Geographic Information, 2014, pp. 16–23, https://doi.org/10.1145/2676440.2676445. Dallas, TX, USA, 4–8 November 2014.
[40] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, Proc. 11th Ann. Conf. Computational Learning Theory. (1998) 92–100, https://doi.org/10.1145/279943.279962.
[41] Y.N. Zhao, L. Wang, N.N. Zhang, X.W. Huang, L.K. Yang, W.B. Yang, Co-Training semi-supervised learning for fine-grained air quality analysis, Atmosphere 14 (2023) 143, https://doi.org/10.3390/atmos14010143.
[42] Z.H. Zhou, M. Li, Tri-Training: exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 17 (2005) 1529–1541, https://doi.org/10.1109/TKDE.2005.186.
[43] Z.H. Zhou, M. Li, Semi-supervised learning by disagreement, Knowl. Inf. Syst. 24 (2010) 415–439, https://doi.org/10.1007/s10115-009-0209-z.
[44] S. Liu, T.H. He, J.H. Dai, A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese, Mobile Network. Appl. 26 (2021) 1891–1903, https://doi.org/10.1007/s11036-020-01725-x.
[45] M. Tümay, M.E. Meral, K.A. Bayindir, Extraction of voltage harmonics using multi-layer perceptron neural network, Neural Comput. Appl. 17 (2008) 585–593, https://doi.org/10.1007/s00521-007-0154-2.
[46] M. Kovačević, M. Pasquato, M. Marelli, A.D. Luca, R. Salvaterra, A.B. Mondoni, Exploring X-Ray variability with unsupervised machine learning I. Self-organizing maps Applied to XMM-Newton data, Astron. Astrophys. 659 (2022) A66, https://doi.org/10.48550/arXiv.2202.08868.
[47] G. Balachandran, J.V.G. Krishnan, Moving scene-based video segmentation using fast convolutional neural network integration of vgg-16 net deep learning architecture, Int. J. Model. Simul. Sc. 14 (2023) 2341014, https://doi.org/10.1142/S1793962323410143.
[48] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey, IEEE T. Neur. Net. Lear. 28 (2015) 2222–2232, https://doi.org/10.48550/arXiv.1503.04069.
[49] X.Y. Wu, Z.X. Liu, L.R. Yin, W.F. Zheng, L.H. Song, J.W. Tian, B. Yang, S. Liu, A haze prediction model in Chengdu based on LSTM, Atmosphere 12 (11) (2021) 1479, https://doi.org/10.3390/atmos12111479.
[50] D.D. Chen, W. Wang, W. Gao, Z.H. Zhou, Tri-net for semi-supervised deep learning, Proc. 27th Int. Joint Conf. Artif. Intell. (2018) 2014–2020, https://doi.org/10.24963/ijcai.2018/278.
[51] Z.T. Yan, N. Yu, H. Wen, Z. Li, H.S. Zhu, L.M. Sun, Detecting internet-scale NATs for IoT devices based on Tri-net, 15th Int. Conf, Wireless Algorithms, Systems, Applications (2020) 602–614, https://doi.org/10.1007/978-3-030-59016-1_50.