Spatio-Temporal Event Forecasting Using Incremental Multi-Source Feature Learning

LIANG ZHAO and YUYANG GAO, Emory University
JIEPING YE, University of Michigan
FENG CHEN, University of Texas
YANFANG YE, Case Western Reserve University
CHANG-TIEN LU and NAREN RAMAKRISHNAN, Virginia Tech

The forecasting of significant societal events such as civil unrest and economic crisis is an interesting and challenging problem which requires both timeliness, precision, and comprehensiveness. Significant societal events are influenced and indicated jointly by multiple aspects of a society, including its economics, politics, and culture. Traditional forecasting methods based on a single data source find it hard to cover all these aspects comprehensively, thus limiting model performance. Multi-source event forecasting has proven promising but still suffers from several challenges, including (1) geographical hierarchies in multi-source data features, (2) hierarchical missing values, (3) characterization of structured feature sparsity, and (4) difficulty in model's online update with incomplete multiple sources. This article proposes a novel feature learning model that concurrently addresses all the above challenges. Specifically, given multi-source data from different geographical levels, we design a new forecasting model by characterizing the lower-level features' dependence on higher-level features. To handle the correlations amidst structured feature sets and deal with missing values among the coupled features, we propose a novel feature learning model based on an Nth-order strong hierarchy and fused-overlapping group Lasso. An efficient algorithm is developed to optimize model parameters and ensure global optima. More importantly, to enable the model update in real time, the online learning algorithm is formulated and active set techniques are leveraged to resolve the crucial challenge when new patterns of missing features appear in real time. Extensive experiments on 10 datasets in different domains demonstrate the effectiveness and efficiency of the proposed models.

CCS Concepts: • **Information systems** → **Data mining**; • **Computing methodologies** → **Machine learning;**

Additional Key Words and Phrases: Event forecasting, multiple data sources, feature selection, online algorithm

40

This work was supported by the National Science Foundation grant: 1755850. It was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon.

Authors' addresses: L. Zhao and Y. Gao, Emory University; emails: {liang.zhao, yuyang.gao}@emory.edu; J. Ye, University of Michigan; F. Chen, University of Texas, Dallas; Y. Ye, Case Western Reserve University; C.-T. Lu and N. Ramakrishnan, Virginia Tech.

1556-4681/2021/09-ART40 $15.00
https://doi.org/10.1145/3464976

ACM Reference format:
Liang Zhao, Yuyang Gao, Jieping Ye, Feng Chen, Yanfang Ye, Chang-Tien Lu, and Naren Ramakrishnan. 2021. Spatio-Temporal Event Forecasting Using Incremental Multi-Source Feature Learning. *ACM Trans. Knowl. Discov. Data.* 16, 2, Article 40 (September 2021), 28 pages.
https://doi.org/10.1145/3464976

1 INTRODUCTION

Significant societal events have a tremendous impact on our entire society, such as disease outbreaks and mass protests, which strongly motivate anticipating their occurrences accurately in real time. For instance, according to a recent **World Health Organization** (**WHO**) report [58], seasonal influenza alone is estimated to result in around 4 million cases of severe illness and about 250,000–500,000 deaths each year. In regions such as the Middle East and Latin America, the majority of instabilities arise from extremism or terrorism, while others are the result of civil unrest. Population-level uprisings by disenchanted citizens are generally involved, usually resulting in major social problems that may involve economic losses that run into the billions of dollars and create millions of unemployed people. Significant societal events are typically caused by multiple social factors. For example, civil unrest events could be caused by economic factors (e.g., increasing unemployment), political factors (e.g., a presidential election), and educational factors (e.g., educational reform). Moreover, societal events can also be driven and orchestrated through social media and news reports. For example, in a large wave of mass protests in the summer of 2013, Brazilian protesters calling for demonstrations frequently used Twitter as a means of communication and coordination. Therefore, to fully characterize these complex societal events, recent studies have begun to focus on utilizing indicators from multiple data sources to track different social factors and public sentiment that jointly indicate or anticipate the potential future events.

These multi-source-based methods share essentially similar workflows. They begin with collecting and preprocessing each single data source individually, from which they extract meaningful features such as ratios, counts, and keywords. They then aggregate these feature sets from all different sources to generate the final input of the forecasting model. The model response, in this case, predicting the occurrence of future events, is then mapped to these multi-source input features by the model. Different data sources commonly have different time ranges. For example, Twitter has been available since 2006, but **Centers for Disease Control and Prevention (CDC**) data dates back to the 1990s. When the predictive model utilizes multiple data sources, of which some are incomplete, typically the samples with missing values in any of these data sources are simply removed, resulting in substantial information loss.

Multi-source forecasting of significant societal events is thus a complex problem that currently still faces several important challenges. **1. Hierarchical topology**. When features in different data sources come from different topological levels, they cannot normally be treated as independent and homogeneous. For example, Figure 1 shows multiple indicators during the "Brazilian Spring", the name given to a large wave of protest movements in Brazil in June 2013 caused by economic problems and spread by social media. Here, indicators in the economy and social media would be the precursors of the protests. Some of these indicators are country-level, such as the exchange rate; some are state-level, such as news reports specific to a state; and some are city-level, such as the Twitter keyword count for chatter geolocated to a specific city. When forecasting city-level protest events, however, it is unrealistic to simply treat the union of all these multi-level features directly as city-level features for prediction. Moreover, it is unreasonable to assume that all cities across the country are equally influenced by the higher-level features and are completely independent of each other. **2. Interactions involving missing values**. When features are drawn

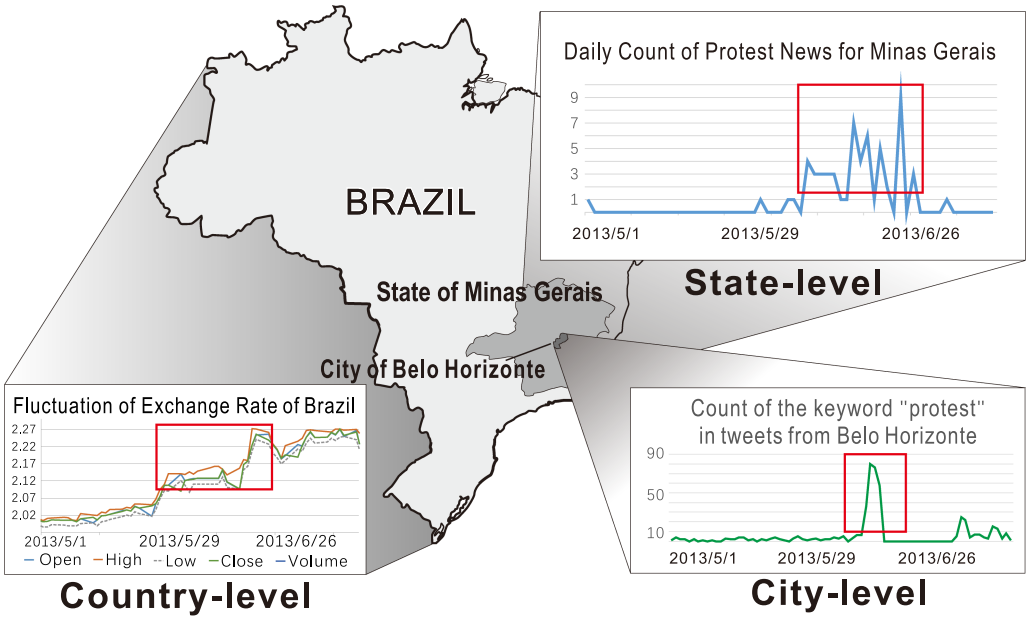

Fig. 1. Predictive indicators from multiple data sources with different geographical levels during the "Brazilian Spring" civil unrest movement.

from different hierarchical topologies, features from higher levels influence those from lower levels. Thus, the missing value in such feature sets will also influence other features. This means that simply discarding the missing values is not an ideal strategy as its interactions with other features also need to be considered. **3. Geo-hierarchical feature sparsity**. Among the huge number of features from multiple data sources, only a portion of them will actually be helpful for predicting the response. However, due to the existence of hierarchical topology among the features, as mentioned earlier, features are not independent of each other. It is thus clearly beneficial to discover and utilize this hierarchically structured pattern to regulate the feature selection process. **4. Incremental model update with new missing patterns of features.** In multi-source models, the availability of the multiple sources usually changes and the model need to adapt swiftly to the new missing patterns of sources, as shown in Figure 2. Retraining with the whole historical data is typically prohibitive and thus incremental learning is preferred. However, this problem cannot be addressed by conventional online learning, because the set of available feature changes when the current missing pattern changes. Furthermore, it also cannot be addressed by existing methods on incremental feature selection because it requires respective feature selection for each corresponding missing pattern. Moreover, the predictive model on new missing patterns also needs to learn from the existing missing patterns in order to quickly gain good generalizability with few initial samples.

In order to simultaneously address all these technical challenges, this article presents a novel model named **hierarchical incomplete multi-source feature learning** (**HIML**) and its incremental-learning version, named **online-HIML** (**oHIML**). HIML is capable of handling the features' hierarchical correlation pattern and secure the model's robustness against missing values and their interactions. To characterize the hierarchical topology among the features from multi-source data, we build a multi-level model that cannot only handle all the features' impacts on the response, but also take into account the interactions between higher- and lower-level features. Under the assumption of feature sparsity, we characterize the hierarchical structure among the

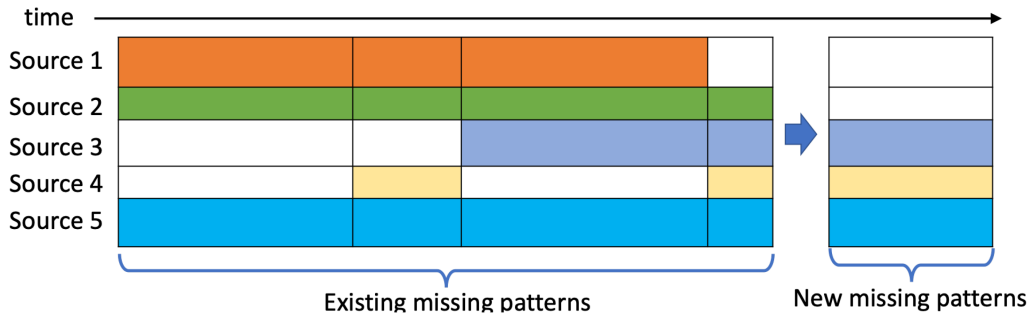

Fig. 2. Not all the data sources (and thus their respective data features) are available all the time. Along with the time, there will be new missing patterns and the model needs to be able to incrementally update along with data with new missing patterns of features.

features and utilize it to regulate a proper hierarchical pattern. Our HIML model can also handle missing values among multiple data sources by incorporating a multitask strategy that treats each missing pattern as a task. Both batch-learning-based and incremental-learning-based methods have been proposed with theoretical analyses.

The main contributions of our study are summarized below:

— **Design a framework for event forecasting based on hierarchical multi-source indicators**. A generic framework is proposed for spatial event forecasting that utilizes hierarchically topological multiple data sources and is based on a generalized multi-level model. A number of classic approaches to related research are shown to be special cases of our model.
— **Propose a robust model for geo-hierarchical feature selection**. To model the structured inherent in geo-hierarchical features across multiple data sources, we propose an N-level interactive group Lasso based on strong hierarchy. To handle interactions among missing values, the proposed model adopts a multi-task framework that is capable of learning the shared information among the tasks corresponding to all the missing patterns.
— **Develop an efficient algorithm for model parameter optimization.** To learn the proposed model, a constrained overlapping group Lasso problem needs to be solved, which is technically challenging. By developing an algorithm based on the **alternating direction method of multipliers** (**ADMM**) and introducing auxiliary variables, we ensure a globally optimal solution to this problem.
— **Propose an incremental multi-source feature learning algorithm**. To quickly learn the new missing patterns in real time without retraining the whole model, an efficient incremental learning method has been proposed based on active set techniques [22]. A theoretical equivalence of the objective function of the existing model is presented, based on which a stochastic ADMM is developed to update the model (with new missing patterns) incrementally.
— **Conduct extensive experiments for performance evaluations.** The proposed batch and incremental-based methods were evaluated on 10 different datasets in two domains: Forecasting civil unrest in Latin America and influenza outbreaks in the United States. The results demonstrate that the proposed approach runs efficiently and consistently outperforms the existing methods in multiple metrics.

This paper is an extension of the article [62] in the following aspects: (1) *A new model and theoretical analysis.* We extend the original model, which is based on a fixed number of multi-source missing patterns, to a new one which can accommodate all the possible multi-source missing

patterns. Moreover, we provide the theoretical proof of the equivalence between the solution of these two model formulations. (2) *A new optimization algorithm.* Based on the newly-proposed extended model, we develop new incremental learning algorithm, which innovatively updates the model in real time when both the samples and new missing patterns of (new) features can be incrementally added. Time complexity analysis of the new algorithm has also been added. (3) *Additional experiments.* We added the results on all the 10 datasets for the newly proposed incremental-learning-based algorithm, named oHIML. We also added the scalability validation and analysis on proposed batch-learning-based and incremental-learning-based algorithms. Discussions on the results are also provided. Moreover, we added a new comparison method, namely AutoInt, which is a multi-head self-attentive neural network with residual connections that maps the numerical, categorical features, and their interactions into the same low-dimensional space. We also added the analysis of the performance of this method. (4) *More comprehensive literature reviews.* We have added relevant literature survey in online models and incremental feature selection. In addition, we also surveyed recent related work in relevant topics on event detection and forecasting, multi-source event forecasting, missing values in multiple data sources, and feature selection in presence of interactions.

The rest of this article is organized as follows: Section 2 reviews background and related work, and Section 3 introduces the problem setup. Section 4 presents our model while Section 5 both batch-learning-based and incremental-learning-based parameter optimization algorithm. The experiments on 10 real-world datasets are presented in Section 6, and the article concludes with a summary of the research in Section 7.

2 RELATED WORK

This section introduces related work in several research areas.

Event detection and forecasting in social media. There is a large body of work that focuses specifically on the identification of ongoing events, such as earthquakes [43] and disease outbreaks [14, 15, 47, 58]. Unlike these approaches, which typically uncover events only after their occurrence, event forecasting methods predict the incidence of such events in the future. Most event forecasting methods focus on temporal events, with no interest in the geographical dimension, such as elections [35] and stock market movements [2]. Few existing approaches can provide true spatiotemporal resolution for the predicted events [56]. For example, Gerber utilized a logistic regression model for spatiotemporal event forecasting [16]. Zhao et al. [59] designed a **multi-task learning** (**MTL**) framework that models forecasting tasks in related geo-locations concurrently. Zhao et al. [57] also designed a new predictive model that jointly characterizes the temporal evolution of both the semantics and geographical burstiness of social media content. Shi et al. [45] focus on jointly detect the events and discover the social media users' interests during the events. By generalizing the spatial locations as nodes in the network, Shao et al. formulate the event detection problems as subgraph detection methods [44], but they are not able to consider the attributes in different hierarchical levels. Cui et al. [8] utilize Weibo data to detect the foodborne disease outbreaks for restaurant regulation while Brown et al. utilize Twitter to forecast the outcomes of sport match results based on prediction markets [4]. Tensor-completion-based techniques have been also applied for event prediction. They usually first learn the underlying factors of different modes (e.g., along spatial and temporal dimensions), and then use time series forecasting techniques to predict the future values for those underlying factors [32, 33, 65]. Zhao et al. [62] has proposed a multi-source event forecasting methods that can handle block-wise missing patterns of the different data sources. These types of methods are difficult to handle multi-level attributes and the model parameters cannot be updated online. For a more comprehensive survey, please refer to recent survey papers such as [55].

Multi-source event forecasting. In recent years, a few researchers have begun to utilize multiple data sources as surrogates to forecast future significant societal events such as disease outbreaks and civil unrest. Chakraborty et al. proposed an ensemble model to forecast **Influenza-like Illness (ILI)** ratios based on seven different data sources [5]. Focusing on civil unrest events, Ramakrishnan et al. employ a LASSO model as the event predictor, where the inputs are the union of feature sets from different data sources [41]. Kallus explores the predictive power of news, blogs, and social media for political event forecasting [24]. Huang et al. [21] propose a multi-modal recurrent framework to jointly detect abnormal events based on citywide spatiotemporal data. Li et al. [26] propose to handle the survival analysis problem, when there are block-wise missing data when using multiple data sources by leveraging MTL strategies. Zhao et al. [61] consider social media data in multiple languages and learn the correspondence among the features in different languages by matrix decomposition. Hetero-ConvLSTM [54] leverages ConvLSTM to merge multiple spatial data sources across a time window to forecast the patterns for the future window, although it cannot handle hierarchical sources and missing patterns in the multi-source data. Wang et al. [48] and Zhao et al. [60] propose new methods that can fuse hierarchical features in different geographical levels. However, although these models utilize multiple data sources that can be used to indicate a number of different aspects of future events, they cannot jointly handle the potential relationships, hierarchy, and missing values among these multi-source features.

Missing values in multiple data sources. The prevention and management of missing data have been discussed and investigated in existing work [17, 50]. One category of work focuses on estimating missing entries based on the observed values [13]. These methods work well when missing data are rare, but are less effective when a significant amount of data is missing. To address this problem, Hernandez et al. utilized probabilistic matrix factorization [20], but their method is restricted to non-random missing values. Yuan et al. [53] utilized MTL to learn a consistent feature selection pattern across different missing groups. Li et al. [28] focus on the multi-source block-wise missing data in survival analysis, modeling, and leverage MTL in both aspects in different features and sources. In order to alleviate the missing values and complement the information across different data sources, Que et al. focus on learning the similarity among different sources by non-negative matrix factorization and similarity constraints on the patterns of different sources [40]. However, none of these approaches focus specifically on missing values in hierarchical multiple data sources. Moreover, none of the above approaches can incrementally update the model for new-coming missing values in real time without retraining the whole model.

Online models and online feature selection. Most of the existing online learning methods assume the set of features does not change but the new samples can update the model in real time [10, 31]. This track of research has been extensively investigated and here several representative works are presented. For example, Duchi et al. [10] dynamically incorporated the geometric knowledge of the data observed in earlier iterations, in order to achieve a new and informative subgradient method. In addition, to capture time-varying characteristics, some adaptive linear regression methods such as recursive least squares [11] and online passive-aggressive algorithms [7] are leveraged to provide an incremental update on the regression model. More recently, online learning for deep neural networks has been attracting fast-increasing attention [49]. For example, Roy et al. [42] develop an adaptive hierarchical network structure composed of deep convolutional neural networks that can grow and learn as new data becomes available. Another research track is to assume that the set of features can change during model training. There are typically two categories of them: heuristic-based [30, 52, 64] and optimization-based [9, 39, 51]. For example, Liu et al. [30] proposed an online multi-label streaming feature selection framework that includes importance selection and redundancy update of the features under online fashion. Assuming that the new features can come sequentially, Zhou et al. [64] presented an adaptive-complexity-penalty

method named α-investing, for online feature selection that dynamically tunes the threshold on the error reduction required when adding each new feature. Li et al. [27] propose novel methods for semi-supervised incremental learning on streaming data by first learning the non-stationary latent feature representation, which is then input into the layers for classification. Ditzler et al. [30] constructed ensemble models using variants of online bagging and boosting to achieve better model generalizability yet similar complexity to single models. However, these approaches focus on the new features, instead of new multi-source missing patterns of feature sets. And they cannot effectively learn the corresponding model for new feature missing patterns in real time.

Feature selection in the presence of interactions. Feature selection by considering feature interactions has been attracting research interest for some time. For example, to enforce specific interaction patterns, Peixoto et al. [19] employed conventional step-wise model selection techniques with hierarchical constraints. Unfortunately, such approaches are expensive for high-dimensional data. Choi et al. proposed a more efficient LASSO-based non-convex problem with re-parametrized coefficients [6]. To obtain globally optimal solutions, more recent research has utilized interaction patterns such as strong or weak hierarchy that are enforced via convex penalties or constraints. Both of these apply a group-lasso-based framework; Lim and Hastie [29] work with a combination of continuous and categorical variables, while Haris et al. [18] explore different types of norms. More recently, kernel-based methods [37] and deep learning techniques [12] have been leveraged to learn feature interactions. For instance, Song et al. [46] invent a multi-head self-attentive neural network with residual connections to map the numerical, categorical features, and their interactions into the same low-dimensional space. However, none of these approaches considers missing values in the feature sets.

3 PROBLEM SETUP

In this section, the problem addressed by this research is formulated. Specifically, Section 3.1 poses the hierarchical multi-source event forecasting problem and introduces the multi-level model formulation. Section 3.2 discusses the problem generalization and challenges.

3.1 Problem Formulation

Multiple data sources could originate at different geographical levels, for example, city-level, state-level, or country-level, as shown in Figure 1. Before formally stating the problem, we first introduce two definitions related to geographical hierarchy.

Definition 1 (Subregion). Given two locations q_i and s_j under the ith and jth ($i < j$) geographical levels, respectively, if the whole spatial area of the location q_i is included by location s_j, we say q_i is a **subregion** of s_j, denoted as $q_i \sqsubseteq s_j$ or equally $s_j \sqsupseteq q_i$ ($i < j$).

Definition 2 (Location Tuple). The location of a tweet or an event is denoted by a **location tuple** $l = (l_1, l_2, \ldots, l_N)$, which is an array that configures each location l_n in each geo-level n in terms of a parent-child hierarchy such that $l_{n-1} \sqsubseteq l_n (n = 2, \ldots, N)$, where l_n is the **parent** of l_{n-1} and l_{n-1} is the **child** of l_n.

For example, for the location "San Francisco", its location tuple could be ("San Francisco", "California", and "USA") that consists of this city, its parent, and the parent's parent.

Suppose X denotes the set of multiple data sources coming from N different geographical levels. These can be temporally split into fixed time intervals t (e.g., "date") and denoted as $X = \{X_{t,l}\}_{t,l}^{T,L} = \{X_{t,l_n}\}_{t,l,n}^{T,L,N}$, where $X_{t,l_n} \in \mathbb{N}^{|\mathcal{F}_n| \times 1}$ refers to the feature vector for the data at time t in location l_n under nth geo-level. Specifically, the element $[X_{t,l_n}]_i$ ($i \neq 0$) is the value for ith feature while $[X_{t,l_n}]_0 = 1$ is a dummy feature to provide a compact notation for bias parameter in the forecasting model. T denotes all the time intervals, L denotes the set of all the locations, and N denotes the set

of all the geographical levels. $\mathcal{F}_n$ denotes the feature set for Level n and $\mathcal{F} = \{\mathcal{F}_n\}_{n=1}^{N}$ denotes the set of features in all the geo-levels. We also utilize a binary variable $Y_{t,l} \in \{1, 0\}$ for each location $l = (l_1, \ldots, l_N)$ at time t to indicate the occurrence ("yes" or "no") of a future event. We also define $Y = \{Y_{t,l}\}_{t,l}^{T,L}$. Thus, the hierarchical multi-source event forecasting problem can be formulated as below:

Problem Formulation: For a specific location $l = (l_1, \ldots, l_N)$ at time t, given data sources under N geographical levels $\{X_{t,l_1}, \ldots, X_{t,l_N}\}$, the goal is to predict the occurrence of future event $Y_{\tau,l}$, where $\tau = t + p$ and $p > 0$ is the lead time for forecasting. Thus, the problem is formulated as the following mapping function:

$$f : \{X_{t,l_1}, \ldots, X_{t,l_N}\} \rightarrow Y_{\tau,l}, \tag{1}$$

where f is the forecasting model.

In Problem (1), input variables $\{X_{t,l_1}, \ldots, X_{t,l_N}\}$ are not independent of each other because the geographical hierarchy among them encompasses hierarchical dependence. Thus, classical single-level models such as linear regression and logistic regression cannot be utilized here.

As generalizations of the single-level models, multi-level models are commonly used for problems where input variables are organized at more than one level. The variables for the locations in Level $n-1$ are dependent on those of their *parents*, which are in Level n ($2 \leq n \leq N$). The highest level (i.e., Level N) variables are independent variables. Without loss of generality and for convenience, here we first formulate the model with $N = 3$ geographical levels (e.g., city-level, state-level, and country-level) and then generalize it to $N \in Z^+$ in Section 3.2. The multi-level models for hierarchical multi-source event forecasting are formulated as follows:

$$\begin{aligned}
&(level-1)\ \ Y_{\tau,l} = \alpha_0 + \sum_{i=1}^{|\mathcal{F}_1|} \alpha_i^T \cdot [X_{t,l_1}]_i + \varepsilon,\\
&(level-2)\ \ \alpha_i = \beta_{i,0} + \sum_{j=1}^{|\mathcal{F}_2|} \beta_{i,j}^T \cdot [X_{t,l_2}]_j + \varepsilon_i,\\
&(level-3)\ \ \beta_{i,j} = W_{i,j,0} + \sum_{k=1}^{|\mathcal{F}_3|} W_{i,j,k}^T \cdot [X_{t,l_3}]_k + \varepsilon_{i,j},
\end{aligned} \tag{2}$$

where α_i, $\beta_{i,j}$, and $W_{i,j,k}$ are the coefficients for models of Level-1, Level-2, and Level-3, respectively. Each Level-1 parameter α_i is linearly dependent on Level-2 parameters $\beta_{i,j}$ and each Level-2 parameter $\beta_{i,j}$ is again linearly dependent on Level-3 parameters $W_{i,j,k}$. ε, ε_i, and $\varepsilon_{i,j}$ are the noise terms for Levels 1, 2, and 3. Combining all the formulas in Equation (2), we get

$$Y_{\tau,l} = \sum_{i=0}^{|\mathcal{F}_1|} \sum_{j=0}^{|\mathcal{F}_2|} \sum_{k=0}^{|\mathcal{F}_3|} W_{i,j,k} \cdot [X_{t,l_1}]_i \cdot [X_{t,l_2}]_j \cdot [X_{t,l_3}]_k + \varepsilon, \tag{3}$$

where ε is noise term. Utilizing tensor multiplication, Equation (3) can be expressed in the following compact notation:

$$Y_{\tau,l} = W \odot Z_{t,l} + \varepsilon, \tag{4}$$

where $W = \{W_{i,j,k}\}_{i,j,k=0}^{|\mathcal{F}_1|,|\mathcal{F}_2|,|\mathcal{F}_3|}$ and $Z_{t,l}$ are two $(|\mathcal{F}_1|+1) \times (|\mathcal{F}_2|+1) \times (|\mathcal{F}_3|+1)$ tensors, and an element of $Z_{t,l}$ is defined as $[Z_{t,l}]_{i,j,k} = [X_{t,l_1}]_i \cdot [X_{t,l_2}]_j \cdot [X_{t,l_3}]_k$. The operator $\odot$ is the summation of the Hadamard product of two tensors such that $A \odot B = \sum_{i,j,k} A_{ijk} \cdot B_{ijk}$ for 3rd-order tensors A and B.

The tensor $Z_{t,l}$ is illustrated in Figure 3(b). Specifically, the terms $[Z_{t,l}]_{i,0,0} = [X_{t,l_1}]_i$, $[Z_{t,l}]_{0,j,0} = [X_{t,l_2}]_j$, and $[Z_{t,l}]_{0,0,k} = [X_{t,l_3}]_k$ are the main-effect variables shown, respectively as green,

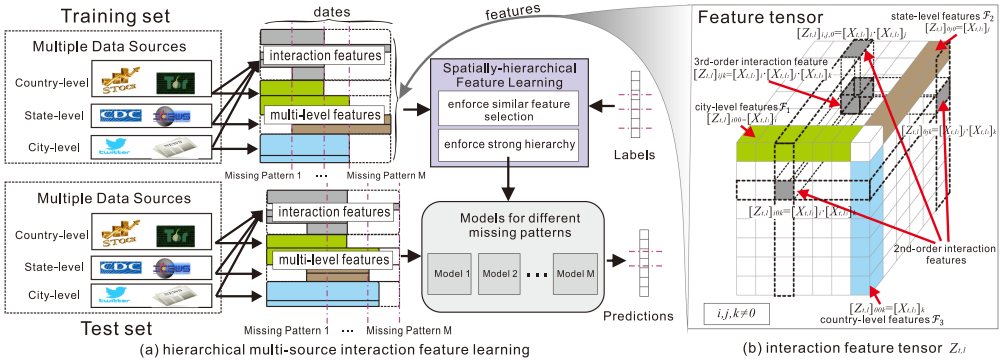

Fig. 3. A schematic view of HIML model.

blue, and brown nodes in Figure 3(b). Main-effect variables are independent variables. The terms $[Z_{t,l}]_{i,j,0} = [X_{t,l_1}]_i \cdot [X_{t,l_2}]_j$, $[Z_{t,l}]_{i,0,k} = [X_{t,l_1}]_i \cdot [X_{t,l_3}]_k$, and $[Z_{t,l}]_{0,j,k} = [X_{t,l_2}]_j \cdot [X_{t,l_3}]_k$ are 2nd-order interactive variables and are shown as nodes on the surfaces formed by the lines of the main-effect variables in Figure 3(b). Their values are dependent on both of their two main-effect variables. The terms $[Z_{t,l}]_{i,j,k} = [X_{t,l_1}]_i \cdot [X_{t,l_2}]_j \cdot [X_{t,l_3}]_k$ are called 3rd-order interactions because their values are dependent on 2nd-order interactive variables, as shown in Figure 3(b). Finally, denote $Z = \{Z_{t,l}\}_{t,l}^{T,L}$ as the set of feature tensors for all the locations L and time intervals T.

3.2 Problem Generalization

Here, the 3-level model in Equation (4) is generalized into an N-level model. Moreover, the linear function in Equation (4) is generalized into nonlinear setting.

3.2.1 1. N-Level Geo-Hierarchy. In Equation (4), we assumed that the number of geographical levels is $N = 3$. Now we extend this by introducing the generalized formulation where the integer $N \geq 2$. We retain the formulation in Equation (4), and generalize the operator $\odot$ into a summation of the Nth-order Hadamard product such that $A \odot B = \sum_{i_1,\ldots,i_N} A_{i_1,\ldots,i_N} \ldots B_{i_1,\ldots,i_N}$. For simplicity, this can be denoted as $A \odot B = \sum_{\vec{i}} A_{\vec{i}} \cdots B_{\vec{i}}$, where $\vec{i} = \{i_1, i_2, \ldots, i_N\}$.

3.2.2 2. Generalized Multi-Level Linear Regression. In Equation (4), we assumed a linear relation between input variable $Z_{t,l}$ and the response variable $Y_{t,l}$. However, in many situations, a more generalized relation could be necessary. For example, we may need a logistic regression setup when modeling a classification problem. Specifically, the generalized version of our multi-level model adds a nonlinear mapping between the input and response variables:

$$Y_{t,l} = h(W \odot Z_{t,l}) + \varepsilon, \tag{5}$$

where $h(\cdot)$ is a convex and differentiable mapping function. In this article, the standard logistic function $h(x) = 1/(1 + e^{-x})$ is considered (see Section 4.3).

Although the models proposed in Equations (4) and (5) are capable of modeling the features coming from different geo-hierarchical levels, they suffer from three challenges: (1) The weight tensor W is typically highly sparse. This is because the main effects could be sparse, meaning that their interaction (i.e., multiplication) will be even more sparse. Without considering this sparsity, the computation will be considerably more time-consuming. (2) The pattern of W is structured. There is a geo-hierarchy among the multi-level features, which causes their interactions in W to follow specific sparsity patterns. Careful and effective consideration and utilization of this structure are both vital and beneficial. (3) The models do not consider missing values, whereas

these are actually quite common in practical applications that use multi-source data. A model that is capable of handling missing values is therefore imperative. In the next section, we present HIML, a novel hierarchical feature learning approach based on constrained overlapping group Lasso, to address all three challenges.

4 HIERARCHICAL INCOMPLETE MULTI-SOURCE FEATURE LEARNING

Without loss of generality and for convenience, Section 4.1 first proposes our hierarchical feature learning model for $N = 3$ geographical levels, and then Section 4.2 generalizes it to handle the problem of missing values, as shown in Figure 3. Section 4.3 then takes the model further by generalizing it to $N \in \mathbb{Z}^+$ geographical levels and incorporating nonlinear loss functions. The algorithm for the model parameter optimization is proposed in Section 5. The relationship of our HIML model to existing models is discussed in Section 4.5.

4.1 Hierarchical Feature Correlation

In fitting models with interactions among variables, a 2nd-order strong hierarchy is widely utilized [18, 23] as this can handle the interactions between two sets of main-effect variables. Here, we introduce their definition as follows:

Lemma 4.1 (2nd-order Strong Hierarchy). *If a 2nd-order interaction term is included in the model, then both of its product factors (i.e., main effect variables) are present. For example, if* $W_{i,j,0} \neq 0$, *then* $W_{i,0,0} \neq 0$ *and* $W_{0,j,0} \neq 0$.

Here we generalize the 2nd-order Strong Hierarchy to Nth-order Strong Hierarchy ($N \in \mathbb{Z}^+ \wedge N \geq 2$) as follows:

Theorem 1 (Nth-order Strong Hierarchy). *If an* N*th-order interaction variable is included in the model, then all of its* n*th-order* ($2 \leq n < N$) *interactive variables and main-effect variables are included.*

Proof. According to Lemma 4.1, if an nth-order interaction variable ($2 \leq n \leq N$) is included, then its product-factor pairs, (n–1)th-order interaction factor and main effect, must also be included. Similarly, if an $(n-k)$th-order interaction variable ($1 \leq k \leq n-2$) is included, then so must its pairs of $(n-k-1)$th-order interaction factor and main effect. By varying k from 1 to $N-2$, we immediately know that any nth-order ($2 \leq n < N$) interactive variables and main effects must be included. □

When $N = 3$, Theorem 1 becomes the *3rd-order strong hierarchy*. Specifically, if $W_{i,j,k} \neq 0$, then we have $W_{i,j,0} \neq 0$, $W_{i,0,k} \neq 0$, $W_{0,j,k} \neq 0$, $W_{i,0,0} \neq 0$, $W_{0,j,0} \neq 0$, and $W_{0,0,k} \neq 0$, where $i, j, k \neq 0$. In the following we propose a general convex regularized feature learning approach that enforces the 3*rd-order strong hierarchy*.

The proposed feature learning model minimizes the following penalized empirical loss:

$$\min_{W} \mathcal{L}(W) + \Omega(W), \tag{6}$$

where $\mathcal{L}(W)$ is the loss function such that $\mathcal{L}(W) = \sum_{t,l} \|Y_{\tau,l} - W \odot Z_{t,l}\|_F^2$. $\Omega(W)$ is the regularization term that encodes task relatedness:

$$\begin{aligned}\Omega(W) = \lambda_0 \sum_{i,j,k\neq 0} |W_{i,j,k}| + \lambda_1 \sum_{j+k\neq 0} \|W_{\cdot,j,k}\|_F \\ + \lambda_2 \sum_{i+k\neq 0} \|W_{i,\cdot,k}\|_F + \lambda_3 \sum_{i+j\neq 0} \|W_{i,j,\cdot}\|_F,\end{aligned} \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. λ_0, λ_1, λ_2, and λ_3 are regularization parameters such that $\lambda_0 = \lambda/(|\mathcal{F}_1| \cdot |\mathcal{F}_2| \cdot |\mathcal{F}_3|)$, $\lambda_1 = \lambda/(\sqrt{|\mathcal{F}_1|} \cdot |\mathcal{F}_2| \cdot |\mathcal{F}_3|)$, $\lambda_2 = \lambda/(|\mathcal{F}_1| \cdot \sqrt{|\mathcal{F}_2|} \cdot |\mathcal{F}_3|)$, and $\lambda_3 = \lambda/(|\mathcal{F}_1| \cdot |\mathcal{F}_2| \cdot \sqrt{|\mathcal{F}_3|})$, where λ is a regularization parameter that balances the tradeoff between the loss function $\mathcal{L}(W)$ and the regularization terms. Equation (7) is a higher-order generalization of the ℓ_2 penalty proposed by Haris et al. [18], which enforces a hierarchical structure under a 2nd-order strong hierarchy.

4.2 Missing Features Values in the Presence of Interactions

As shown in Figure 3(a), multiple data sources usually have different time durations, which result in incomplete data in multi-level features and about the feature interactions among them. Before formally describing the proposed generalized model for missing values, we first introduce two related definitions.

Definition 3 (Missing Pattern Block). A **missing pattern block (MPB)** is a block of multi-source data $\{X_{t,l}\}_{t,l}^{T_m,L}$ ($T_m \subseteq T$) that share the same missing pattern of feature values. Define $\mathcal{M}(X_{t,l})$ as the set of missing-value features of the data $X_{t,l}$. Assume the total number of MPBs is M, then they must satisfy the following three criteria:

— (completeness) : $T = \bigcup_m^M T_m$
— (coherence) : $\forall t_i, t_j \in T_m : \mathcal{M}(X_{t_i,l}) = \mathcal{M}(X_{t_j,l})$
— (exclusiveness) : $\forall t_i \in T_m, t_j \in T_n, m \neq n : \mathcal{M}(X_{t_i,l}) \neq \mathcal{M}(X_{t_j,l})$

Therefore, *completeness* indicates that the whole time period of dataset is covered by the union of all MPB's. *Coherence* expresses the notion that any time points in the same MPB have identical set of missing features. Finally, *Exclusiveness* suggests that time points in different MPB's must have different sets of missing features.

Definition 4 (Feature Indexing Function). We define $\mathcal{W}_m$ as the weight tensor learned by the data for MPB $\{X_{t,l}\}_{t,l}^{T_m,L}$. A feature indexing function $\mathcal{W}_{G(\cdot)}$ is defined as follows:

$$\mathcal{W}_{G(\cdot)} \equiv \bigcup_m^M [\mathcal{W}_m]_{(\cdot)}.$$

For example, $\mathcal{W}_{G(i,j,k)} \equiv \bigcup_m^M [\mathcal{W}_m]_{i,j,k}$ and $\mathcal{W}_{G(i,\cdot,k)} \equiv \bigcup_m^M [\mathcal{W}_m]_{i,\cdot,k}$.

According to Definitions 3 and 4, the feature learning problem based on a third-order strong hierarchy is then formalized as

$$\begin{aligned}\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) &+ \lambda_0 \sum_{i,j,k\neq 0} \|\mathcal{W}_{G(i,j,k)}\|_F + \lambda_1 \sum_{j+k\neq 0} \|\mathcal{W}_{G(\cdot,j,k)}\|_F \\ &+ \lambda_2 \sum_{i+k\neq 0} \|\mathcal{W}_{G(i,\cdot,k)}\|_F + \lambda_3 \sum_{i+j\neq 0} \|\mathcal{W}_{G(i,j,\cdot)}\|_F, \end{aligned} \tag{8}$$

where the loss function $\mathcal{L}(\mathcal{W})$ is defined as follows:

$$\mathcal{L}(\mathcal{W}) = \sum\nolimits_{T_m \subseteq T} \frac{1}{|T_m|} \sum\nolimits_{t,l}^{T_m,L} \|Y_{\tau,l} - \mathcal{W}_m \odot Z_{t,l}\|_F^2, \tag{9}$$

where $|T_m|$ is the total time period of the MPB T_m.

4.3 Model Generalization

We can now extend the above 3rd-order strong hierarchy-based incomplete feature learning to Nth-order and prove that the proposed objective function satisfies the Nth-order strong hierarchy.

The model is formulated as follows:

$$\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \left\| \mathcal{W}_{G(\vec{i})} \right\|_F + \sum_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \left\| \mathcal{W}_{G(\vec{i}_{-n})} \right\|_F, \tag{10}$$

where $\mathcal{W} = \{\mathcal{W}_m\}_m^M$, and $\mathcal{W}_m \in \mathbb{R}^{|\mathcal{F}_1|\times\cdots\times|\mathcal{F}_N|}$ is an Nth-order tensor whose element index is $\vec{i} = \{i_1, \ldots, i_n\}$. Also denote $\vec{i}_{-n} = \{i_1, \ldots, i_{n-1}, i_{n+1}, \ldots, i_N\}$. $\mathcal{W}_{G(\vec{i})} \equiv \bigcup_m^M [\mathcal{W}_m]_{(\vec{i})}$ according to Definition 4. $\lambda_0 = \lambda/(\prod_i^N |\mathcal{F}_i|)$, $\lambda_n = \lambda/(\sqrt{|\mathcal{F}_n|} \cdot \prod_{i\neq n} |\mathcal{F}_i|)$.

Theorem 2. *The regularization in Equation (10) enforces a hierarchical structure under an Nth-order strong hierarchy. The objective function in Equation (10) is convex.*

Proof. First, $\mathcal{L}(\mathcal{W})$ is convex because the Hessian matrix for $\|Y_{\tau,l} - \mathcal{W}_m \odot Z_{t,l}\|_F^2$ is semidefinite. Second, according to Definition 4 and the properties of the norm, $\|\mathcal{W}_{G(\vec{i})}\|_F = \|\bigcup_m^M [\mathcal{W}_m]_{\vec{i}}\|_F$ is convex. Similarly, $\|\mathcal{W}_{G(\vec{i}_{-n})}\|$ is also convex. Therefore, the objective function is convex. □

Our model is not restricted to a linear regression and can be extended to generalized linear models, such as logistic regression. The loss function is as follows:

$$\begin{aligned}\mathcal{L}_M(\mathcal{W}) = -\sum_{T_m \subseteq T} \frac{1}{|T_m|} \sum_{t,l}^{T_m, L} \{ & Y_{\tau,l} \log h(\mathcal{W}_m \odot Z_{t,l}) \\ & \cdot (1 - Y_{\tau,l}) \log (1 - h(\mathcal{W}_m \odot Z_{t,l}))\},\end{aligned} \tag{11}$$

where $h(\cdot)$ could be a nonlinear convex function such as the standard logistic function $h(x) = 1/(1 + e^{-x})$.

4.4 Exponentially Many Possible Missing Patterns

This section considers the situation when there are new missing patterns appearing in real time when the model is updated incrementally. Theoretically equivalent problem is presented and proved based on active set techniques.

Because the missing patterns come in sequentially in time order, there could be new missing patterns along with time and thus our model framework should be able to accommodate all the possible missing patterns that could appear in the future, in order to achieve incremental learning.

However, the number of all the possible missing patterns is exponentially many. Specifically, recall that the number of primitive features for each n-th layer is $|\mathcal{F}_n|$, then the total number of possible missing patterns is $2^{\sum_n^N |\mathcal{F}_n|}$. Assume among all of them, there are M missing patterns that have already been seen, and thus there are another $M' = 2^{\sum_n^N |\mathcal{F}_n|} - M$ unseen missing patterns. Therefore, any unseen missing pattern(s) can be denoted as $\mathcal{W}'_{G(\cdot)} \subseteq \bigcup_{m=M+1}^{M'+M} [\mathcal{W}_m]_{(\cdot)}$. Then the objective function which also includes the unseen missing patterns is as follows:

$$\begin{aligned}\min_{\mathcal{W},\mathcal{W}'} \mathcal{L}(\mathcal{W}) + \lambda_0 & \sum\nolimits_{\min(\vec{i})\neq 0} \left\| \mathcal{W}_{G(\vec{i})} \bigcup \mathcal{W}'_{G(\vec{i})} \right\|_F \\ + \sum\nolimits_{n=1}^{N} \lambda_n & \sum\nolimits_{\vec{i}_{-n}\neq\vec{0}} \left\| \mathcal{W}_{G(\vec{i}_{-n})} \bigcup \mathcal{W}'_{G(\vec{i}_{-n})} \right\|_F,\end{aligned} \tag{12}$$

where it should be noted that there is no $\mathcal{W}'$ in the loss function, $\mathcal{L}(\mathcal{W})$ because the other M' missing patterns have not been seen in the historical data.

However, Equation (12) could be prohibitively more time-consuming to be solved than Equation (10) because $M' > M$ and sometimes $M' \gg M$. To address this problem, in the following, we present a theorem which shows the equivalence between these two problems.

Theorem 3. *The solutions of $\mathcal{W}$ in the objective functions in Equation (10) and Equation (12) are identical.*

Proof. We first prove that the solution to the variables $\mathcal{W}'$ is all-zeros by contradiction. Specifically, assume there exist solution $\mathcal{W}$ and $\mathcal{W}''$ to the objective function value such that $\mathcal{W}'' \neq \mathbf{0}$. Then there must be a corresponding solution $\mathcal{W}$ and $\mathcal{W}' \neq \mathbf{0}$ which will achieve an even lower objective function value because $\sum_{\min(\vec{i})\neq 0} \|\mathcal{W}_{G(\vec{i})} \bigcup \mathbf{0}\|_F \leq \sum_{\min(\vec{i})\neq 0} \|\mathcal{W}_{G(\vec{i})} \bigcup \mathcal{W}''_{G(\vec{i})}\|_F$, as well as $\sum_{n=1}^{N} \lambda'_n \sum_{\vec{i}_{-n}\neq\vec{0}} \|\mathcal{W}_{G(\vec{i}_{-n})} \bigcup \mathbf{0}\|_F \leq \sum_{n=1}^{N} \lambda'_n \sum_{\vec{i}_{-n}\neq\vec{0}} \|\mathcal{W}_{G(\vec{i}_{-n})} \bigcup \mathcal{W}'_{G(\vec{i}_{-n})}\|_F$.

Then it is easy to see that $\sum_{\min(\vec{i})\neq 0} \|\mathcal{W}_{G(\vec{i})} \bigcup \mathbf{0}\|_F = \sum_{\min(\vec{i})\neq 0} \|\mathcal{W}_{G(\vec{i})}\|_F$ and $\sum_{n=1}^{N} \lambda'_n \sum_{\vec{i}_{-n}\neq\vec{0}} \|\mathcal{W}_{G(\vec{i}_{-n})} \bigcup \mathbf{0}\|_F = \sum_{n=1}^{N} \lambda'_n \sum_{\vec{i}_{-n}\neq\vec{0}} \|\mathcal{W}_{G(\vec{i}_{-n})}\|_F$. The proof is completed. □

The above proof indicates that the variables involved in Equation (12) are the active set among all the variables in Equation (10). Here *active set* means the nontrivial solution typically consisting of non-zero feature weights while all the remaining feature weights outside the active set in Equation (10) are trivially zeros, following the definitions from [1, 22]. The equivalence between them shows us an efficient way which only needs to solve the small problem on the active set [22] (i.e., Equation (10)), and then involve newly-seen missing patterns incrementally instead of directly involving all the possible missing patterns. More detailed descriptions of the executive algorithm will be given in Section 5.2.

4.5 Relations to Other Approaches

In this section, we show that several classic previous models are actually special cases of the proposed HIML model.

1. **Generalization of block-wise incomplete multi-source feature learning**. Let $N = 1$, which means there is only one hierarchical level in the multisource data. Our model in Equation (10) is thus reduced to an incomplete multi-source feature learning [53]:

$$\min_{W} \sum_{m} \frac{1}{2C_m} \sum_{n}^{C_m} \|Y_n - W_m \cdot Z_n\|_F^2 + \lambda_0 \sum_{i}^{|\mathcal{F}|} \|W_{G(i)}\|_F, \tag{13}$$

where C_m is the count of observations in the mth MPB and $\mathcal{F}$ is the feature set.

2. **Generalization of LASSO**. Let $N = 1$ and $M = 1$, which means there is only one level and there are no missing values. Our HIML model is thus reduced to a regression with ℓ_1-norm regularization [36]:

$$\min_{W} \frac{1}{2C} \sum\nolimits_{i}^{C} \|Y_i - W \cdot Z_i\|_F^2 + \lambda_0 \sum\nolimits_{i}^{|\mathcal{F}|} |W_i|, \tag{14}$$

where C is the count of observations.

3. **Generalization of interactive LASSO**. Let $N = 2$ and $M = 1$, which means there are only two hierarchical levels in data without missing value. HIML is thus reduced to a regression with regularization based on 2nd-order strong hierarchy [18]:

$$\begin{aligned}\min_{W} \frac{1}{2C} \sum\nolimits_{i}^{C} \|Y_i - W \odot Z_i\|_F^2 + \lambda_0 \sum\nolimits_{i,j\neq 0} |W_{i,j}| \\ + \lambda_1 \sum\nolimits_{j=1}^{|\mathcal{F}_1|} \|W_{\cdot,j}\|_F + \lambda_2 \sum\nolimits_{i=1}^{|\mathcal{F}_2|} \|W_{i,\cdot}\|_F,\end{aligned} \tag{15}$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ are the feature sets for the two levels, respectively.

5 PARAMETER OPTIMIZATION

In this section, we propose the optimization algorithms for the objective functions developed in the last section. The batch and incremental-based algorithms as well as their analyses are elaborated in Sections 5.1 and 5.2, respectively.

5.1 Batch Learning Algorithm

The problem in Equation (10) contains an overlapping group Lasso which makes it difficult to solve. To decouple the overlapping terms, we introduce an auxiliary variable Φ and reformulate Equation (10) as follows:

$$\begin{aligned} \min_{\mathcal{W},\Phi} \ & \mathcal{L}_M(\mathcal{W}) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \left\| \Phi^{(0)}_{G(\vec{i})} \right\|_F + \sum\nolimits_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \left\| \Phi^{(n)}_{G(\vec{i}_{-n})} \right\|_F \\ & s.t. \ \ \mathcal{W}_m = \Phi^{(n)}_m, \ m = 1, \ldots, M; \ n = 1, \ldots, N., \end{aligned} \tag{16}$$

where the parameter $\Phi^{(n)}_m \in \mathbb{R}^{|\mathcal{F}_1|\times\cdots\times|\mathcal{F}_N|}$ is the auxiliary variable for the mth MPB for Level n. $\Phi_{G(\cdot)}$ then follows Definition 4 such that $\Phi_{G(\cdot)} = \bigcup_m^M [\Phi_m]_{(\cdot)}$. M is defined in Definition 3 and N is the number of levels of the features.

It is easy to see that Equation (16) is still convex using Theorem 2. We propose to solve this constrained convex problem using the **alternative direction method of multipliers** (**ADMM**) framework. The augmented Lagrangian function of Equation (16) is

$$\begin{aligned} L_\rho(\mathcal{W}, \Phi, \Gamma) = & \ \mathcal{L}_M(\mathcal{W}) + \sum\nolimits_{m,n}^{M,N} tr(\Gamma^{(n)}_m (\mathcal{W}_m - \Phi^{(n)}_m)) \\ & + \sum\nolimits_{n=1}^{N} \lambda_n \sum_{I_{-n}\neq \vec{0}} \left\| \Phi^{(n)}_{G(\vec{i}_{-n})} \right\|_F + \rho/2 \sum\nolimits_{m,n}^{M,N} \left\| \mathcal{W}_m - \Phi^{(n)}_m \right\|_F^2 \\ & + \lambda_0 \sum\nolimits_{\min(\vec{i})\neq 0} \left\| \Phi^{(0)}_{G(\vec{i})} \right\|_F, \end{aligned} \tag{17}$$

where ρ is a penalty parameter. $tr(\cdot)$ denotes the trace of a matrix. $\Gamma^{(n)}_m$ is a Lagrangian multiplier for the constraint $\mathcal{W}_m - \Phi^{(n)}_m = 0$.

To solve the objective function in Equation (17) with multiple unknown parameters $\mathcal{W}$, Φ, and Γ, we propose the hierarchical incomplete feature learning algorithm as in Algorithm 1. It alternately optimizes each of the unknown parameters until convergence is achieved. Lines 11–12 show the calculation of residuals and Lines 13–19 illustrate the updating of the penalty parameter, which follows the updating strategy proposed by Boyd et al. [3]. Lines 4–10 show the updating of each of the unknown parameters by solving the subproblems described in the following.

1. Update $\mathcal{W}_m$.

The weight tensor $\mathcal{W}_m$ is learned as follows:

$$\mathcal{W}_m = \operatorname*{argmin}_{\mathcal{W}_m} \mathcal{L}_M(\mathcal{W}) + \frac{N\rho}{2} \left\| \frac{1}{N} \sum\nolimits_n^N \Phi^{(n)}_m - \frac{1}{N\rho} \sum\nolimits_n^N \Gamma^{(n)}_m - \mathcal{W}_m \right\|_F^2, \tag{18}$$

which is a generalized linear regression with least-squares loss functions. A second-order Taylor expansion is performed to solve this problem, where the Hessian is approximated using a multiple of the identity with an upper bound of $1/(4 \cdot I)$. I denotes the identity matrix.

2. Update $\Phi^{(n)}_m$ $(n \geq 1)$.

The auxiliary variable $\Phi^{(n)}_m$ is learned as follows:

$$\Phi^{(n)}_m \leftarrow \operatorname*{argmin}_{\Phi^{(n)}_m} \frac{\rho}{2} \left\| \Phi^{(n)}_m - \mathcal{W}_m - \frac{\Gamma^{(n)}_m}{\rho} \right\|_F^2 + \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \left\| \Phi^{(n)}_{G(\vec{i}_{-n})} \right\|_F, \tag{19}$$

which is a regression problem with ridge regularization. This problem can be efficiently using the proximal operator [3].

3. Update $\Phi_m^{(0)}$.

The auxiliary variable $\Phi_m^{(0)}$ is learned as follows:

$$\Phi_m^{(0)} \leftarrow \underset{\Phi_m^{(0)}}{\operatorname{argmin}} \frac{\rho}{2} \left\| \Phi_m^{(0)} - \mathcal{W}_m - \frac{\Gamma_m^{(0)}}{\rho} \right\|_F^2 + \lambda_0 \sum_{\min(\vec{i})\neq 0} \left\| \Phi_{G(\vec{i})}^{(0)} \right\|_F, \tag{20}$$

which is also a regression problem with ridge regularization and can be again efficiently solved by utilizing the proximal operator.

4. Update $\Gamma_m^{(n)}$.

The Lagrangian multiplier is updated as follows:

$$\Gamma_m^{(n)} \leftarrow \Gamma_m^{(n)} + \rho \left(\mathcal{W}_m - \Phi_m^{(n)} \right). \tag{21}$$

ALGORITHM 1: Hierarchical Incomplete Feature Learning

Require: Z, Y, λ
Ensure: solution $\mathcal{W}$
Initialize $\rho = 1$, $\mathcal{W}_m$, Γ, $\Phi = \mathbf{0}$.
Choose $\varepsilon_s > 0$, $\varepsilon_r > 0$.
repeat
  for $m \leftarrow 1, \ldots, M$ **do**
    $\mathcal{W}_m \leftarrow$ Equation (18)
    for $n \leftarrow 0, \ldots, N$ **do**
      $\Phi_m^{(n)} \leftarrow$ Equation (20) # Equation (19) if $n \neq 0$
      $\Gamma_m^{(n)} \leftarrow$ Equation (21)
    end for
  end for
  $s = \rho \| \{\Phi_m^{(n)} - \Psi_m^{(n)}\}_{m,n}^{M,N} \|_F$ # Calculate dual residual
  $r = \| \{\mathcal{W}_m^{(n)} - \Psi_m^{(n)}\}_{m,n}^{M,N} \|_F$ # Calculate primal residual
  if $r > 10s$ **then**
    $\rho \leftarrow 2\rho$ # Update penalty parameter
  else if $10r < s$ **then**
    $\rho \leftarrow \rho/2$
  else
    $\rho \leftarrow \rho$
  end if
until $r < \varepsilon^r$ and $s < \varepsilon^s$

Algorithm Analyses: As shown in Theorem 2, the objective function in Equation (16) is convex. In addition, the constraint is simple linear equality. Thus, the ADMM algorithm guarantees to converge to global optima, following the proof process in [3].

For the time complexity, the subproblem for calculating $\mathcal{W}_m$ requires $O(|Z| \cdot T \cdot L)$ thanks to the utilization of Hessian matrix approximation introduced above. $|Z|$ is the number of interaction features. The subproblem for $\Phi^{(n)}$ is dominated by group soft-thresholding, with the time complexity of $O(M \cdot |Z|)$. In all, the total time complexity is $O(l_0 \cdot (l_1 \cdot M \cdot |Z| \cdot T \cdot L + M \cdot |Z|))$, where $C = l_0 \cdot l_1$ and l_0 and l_1 are the number of iterations for the loops of ADMM and the first subproblem, respectively.

5.2 Incremental Learning Algorithm

To improve the time and memory efficiency and achieve incremental updating of the learned model, here we propose an online version of the parameter optimization algorithm. Assume current time is denoted as t, therefore we have:

$$\min_{\mathcal{W}_t} \mathcal{L}_t(\mathcal{W}_t) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \left\| \mathcal{W}_{G(\vec{i}),t} \right\|_F + \sum_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \left\| \mathcal{W}_{G(\vec{i}_{-n}),t} \right\|_F, \tag{22}$$

where $\mathcal{W}_t$ denotes the feature weight at time t and $\mathcal{L}_t$ denotes the loss function at time t. Similar to the above, we assume there are m missing patterns in $\mathcal{W}_t$.

Now a new sample comes, and there are two possible situations: When this new sample follows an existing missing pattern, then there is no change to the above objective function. Otherwise, when this sample brings a new missing pattern and its feature weight is denoted as $\mathcal{W}_{\text{new}}$, we extend our feature set $\mathcal{W}_t$ by $\mathcal{W}_t \leftarrow \mathcal{W}_t \bigcup \{\mathcal{W}_{\text{new}}\}$. According to Theorem 3, the objective function based on this extended feature set is still equivalent to the original one in Equation (22). After building up Lagrangian forms in a way similar to that in Section 5.1, the problem can be solved and illustrated in Algorithm 2. In general, for each new-coming sample, in Lines 3–5 we first identify if it brings new missing patterns, and extend our objective function if so. Then, different parameters are updated iteratively through their corresponding subproblems in Lines 6–20, which will also be described in detail in the following. Note that any parameter that has a subscript "t" denotes that it is at time t.

1. Update $\mathcal{W}_{m,t}$.

Denotes m_t as the missing pattern of the current time t, the weight tensor $\mathcal{W}_{m,t+1}$ $(m = m_t)$ is learned as follows:

$$\begin{aligned}\mathcal{W}_{m,t+1} \leftarrow & \operatorname*{argmin}_{\mathcal{W}_m} \sum\nolimits_l^L \log(1+\exp(-Y_{t,l}(W_m X_{t,l}))) \\ & + \sum\nolimits_n^N tr\left(\Gamma_m^{(n)}\left(\mathcal{W}_m - \Phi_{m,t}^{(n)}\right)\right) + \rho/2 \sum\nolimits_n^N \left\|\mathcal{W}_m - \Phi_{m,t}^{(n)}\right\|_F^2 \\ & + D(\mathcal{W}_{m,t+1}, \mathcal{W}_{m,t})/\eta_{t+1},\end{aligned} \tag{23}$$

where the function $D(x, y)$ denotes the Bregman divergence [63] between x and y to keep the smoothness of the parameter value update in consecutive time points. And $\eta_t \propto 1/\sqrt{t}$ is the step-size. When $m \neq m_t$, the weight tensor $\mathcal{W}_{m,t+1}$ is updated as follows:

$$\begin{aligned}\mathcal{W}_{m,t+1} \leftarrow & \sum\nolimits_n^N tr\left(\Gamma_m^{(n)}\left(\mathcal{W}_m - \Phi_{m,t}^{(n)}\right)\right) + \rho/2 \sum\nolimits_n^N \left\|\mathcal{W}_m - \Phi_{m,t}^{(n)}\right\|_F^2 \\ & + D(\mathcal{W}_{m,t+1}, \mathcal{W}_{m,t})/\eta_{t+1}.\end{aligned} \tag{24}$$

2. Update $\Phi_{\cdot,t}^{(n)}$ $(n \geq 1)$.

The auxiliary variable $\Phi_m^{(n)}$ is learned as follows:

$$\begin{aligned}\Phi_{\cdot,t+1}^{(n)} \leftarrow & \sum_m^M \operatorname*{argmin}_{\Phi_m^{(n)}} \frac{\rho}{2}\left\|\Phi_m^{(n)} - \mathcal{W}_m - \frac{\Gamma_m^{(n)}}{\rho}\right\|_F^2 + \lambda_n \sum_{\vec{i}_{-n}\neq\vec{0}} \left\|\Phi_{G(\vec{i}_{-n})}^{(n)}\right\|_F \\ & + D\left(\Phi^{(n)}, \Phi_t^{(n)}\right)/\eta_{t+1}.\end{aligned} \tag{25}$$

3. Update $\Phi_{\cdot,t}^{(0)}$.

The auxiliary variable $\Phi_m^{(0)}$ is learned as follows:

$$\begin{aligned}\Phi_{\cdot,t+1}^{(0)} \leftarrow & \operatorname*{argmin}_{\Phi^{(0)}} \frac{\rho}{2}\sum_m^M \left\|\Phi_m^{(0)} - \mathcal{W}_m - \frac{\Gamma_m^{(0)}}{\rho}\right\|_F^2 + \lambda_0 \sum_{\min(\vec{i})\neq 0} \left\|\Phi_{G(\vec{i})}^{(0)}\right\|_F \\ & + D\left(\Phi^{(0)}, \Phi_t^{(0)}\right)/\eta_{t+1}.\end{aligned} \tag{26}$$

4. Update $\Gamma_{m,t}^{(n)}$.

The Lagrangian multiplier is updated as follows:

$$\Gamma_{m,t+1}^{(n)} \leftarrow \Gamma_{m,t}^{(n)} + \rho\left(\mathcal{W}_{m,t+1} - \Phi_{m,t+1}^{(n)}\right). \tag{27}$$

ALGORITHM 2: Online Hierarchical Incomplete Feature Learning

Require: $Z^{(t)}$, $Y^{(t)}$ $(t = 1, 2, \cdots)$, λ
Ensure: solution $\mathcal{W}$
Initialize $\rho = 1$, $\mathcal{W}_m^{(t)}$, $\Gamma^{(t)}$, $\Phi^{(t)} = \mathbf{0}$.
for A new sample **do**
  if The sample has new missing pattern **then**
    Equivalently extend the current objective function using Theorem 3
  end if
  for $m \leftarrow 1, \ldots, M$ **do**
    if $m = m_t$ **then**
      $\mathcal{W}_m^{(t)} \leftarrow$ Equation (23)
    else
      $\mathcal{W}_m^{(t)} \leftarrow$ Equation (24)
    end if
    for $n \leftarrow 0, \ldots, N$ **do**
      if n=0 **then**
        $\Phi_m^{(n),(t)} \leftarrow$ Equation (26)
      else
        $\Phi_m^{(n),(t)} \leftarrow$ Equation (25)
      end if
      $\Gamma_m^{(n),(t)} \leftarrow$ Equation (27)
    end for
  end for
  $s = \rho \| \{\Phi_m^{(n),(t)} - \Psi_m^{(n),(t)}\}_{m,n}^{M,N} \|_F$ # Calculate Dual residual
  $r = \| \{\mathcal{W}_m^{(n)} - \Psi_m^{(n),(t)}\}_{m,n}^{M,N} \|_F$ # Calculate primal residual
  if $r > 10s$ **then**
    $\rho \leftarrow 2\rho$ # Update penalty parameter
  else if $10r < s$ **then**
    $\rho \leftarrow \rho/2$
  else
    $\rho \leftarrow \rho$
  end if
  $t \leftarrow t + 1$
end for

For the time complexity of each update of the above online algorithm, the subproblem for calculating $\mathcal{W}_m$ requires $O(|Z|)$ thanks to the utilization of the Hessian matrix approximation introduced above, where $|Z|$ denotes the number of interaction features. The subproblem for $\Phi^{(n)}$ is dominated by group soft-thresholding, with the time complexity of $O(M \cdot |Z|)$. In all, the total time complexity is $O(l_0 \cdot (l_1 \cdot M \cdot |Z| + M \cdot |Z|)) = O(C \cdot M \cdot |Z|)$, where l_0 and l_1 are the number of iterations for the loops of ADMM and first subproblem, respectively. Thus, $C = l_0 \cdot l_1$ is a variable independent of M and $|Z|$.

6 EXPERIMENT

In this section, the performance of the proposed model HIML is evaluated using 10 real datasets from different domains. First, the experimental setup is introduced. The effectiveness and efficiency of HIML are then evaluated against several existing methods for a number of different data missing ratios. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU@ 3.40GHz) and 16.0 GB memory.

6.1 Experimental Setup

6.1.1 Datasets and Labels. In this article, 10 different datasets from different domains were used for the experimental evaluations, as shown in Table 1. Among these, nine datasets were used for event forecasting under the civil unrest domain for nine different countries in Latin America. For these datasets, four data sources from different geographical levels were adopted as the model

Table 1. Labels of Different Datasets (CU=civil unrest; FLU=ILI)

Dataset	Domain	Label sources[1]	#Events
Argentina	CU	Clarín; La Nación; Infobae	1,306
Brazil	CU	O Globo; O Estado de São Paulo; Jornal do Brasil	3,226
Chile	CU	La Tercera; Las Últimas Notícias; El Mercurio	706
Colombia	CU	El Espectador; El Tiempo; El Colombiano	1,196
El Salvador	CU	El Diáro de Hoy; La Prensa Gráfica; El Mundo	657
Mexico	CU	La Jornada; Reforma; Milenio	5,465
Paraguay	CU	ABC Color; Ultima Hora; La Nacíon	1,932
Uruguay	CU	El País; El Observador	624
Venezuela	CU	El Universal; El Nacional; Ultimas Notícias	3,105
U.S.	FLU	CDC Flu Activity Map	1,027

inputs, which are Twitter, **The Onion Router** (**Tor**) network traffic statistics,[2] Currency Exchange,[3] and **Integrated Crisis Early Warning System** (**ICEWS**) counts,[4] as shown in Table 3. The features of each data source are shown in Table 2. The data collected for each source was partitioned into a sequence of date-interval subcollections. The data for the period from April 1, 2013 to December 31, 2013 was used for training, while the data from January 1, 2014 to December 31, 2014, was used for the performance evaluation. The locations of the tweets were all geocoded by the EMBERS geocoder [41]. The event forecasting results were validated against a labeled event set, known as the **gold standard report** (**GSR**), exclusively provided by MITRE [34]. GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin America [41], as shown in Table 1. An example of a labeled GSR event is given by the tuple: (CITY = "Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20"). On the other hand, for the input data in each time segment in each data source, it can be formulated as a feature vector along its corresponding features listed in Table 2. Notice that for the keyword features shown in Table 2, they are just formulated as keyword counts, namely they are in the form of "bag of words", and thus organized as part of the feature vector for the corresponding data sources.

The other dataset was collected to track influenza outbreaks in the United States and consists of three data sources from different geographical levels, which are Twitter, ILI-Net,[5] and FluSurv-NET,[6] as shown in Table 4. These data sources all have different geographical levels. The features of each data source are shown in Table 2. In this case, the data collection for each source was partitioned into a sequence of week-interval subcollections. The data for the period from January 1,

[1]In addition to the top three domestic news outlets, the following news outlets are included: The New York Times, The Guardian, The Wall Street Journal, The Washington Post, The International Herald Tribune, The Times of London, and Infolatam.
[2]Tor: https://www.torproject.org/.
[3]Currency Exchange: http://finance.yahoo.com/currency-converter/.
[4]ICEWS project: http://www.lockheedmartin.com/us/products/W-ICEWS.html.
[5]ILI-NET:https://wwwn.cdc.gov/ilinet/.
[6]FluSurv-NET:http://www.cdc.gov/flu/weekly/.

Table 2. Features of Multiple Data Sources

domain	data sources	features
Civil Unrest	CURRENCY	Open,High,Low,Close
	Tor	Tor network traffic statistics
	ICEWS	CAMEO Codes[8] of event news article content
	Twitter	Volume time series of 982 keywords from [41]
FLU	FluSurv-NET	Influenza Hospitalization Ratio by age groups: 0-4 yr, 5-17 yr, 18-49 yr, 50-64 yr, and 65+ yr
	ILI-Net	weighted/unweighted ILI ratios,positive perctentage, #cases of flu types: A(H1N1), A(N1), A(H3), A, B, H3N2v
	Twitter	Volume time series of 522 keywords from [38]

Table 3. Geographical Levels and Time Ranges of the Multiple Data Sources for Civil Unrest Forecasting

	Level 1	Level 2	Level 3
Geo-level	City	State	Country
data sources: training period	Twitter: 2013-04-01~2013-12-31	ICEWS: 2013-04-01~2013-07-10 2013-10-21~2013-12-31	CURRENCY: 2013-04-01~2013-10-21 TOR: 2013-04-01~2013-10-21

Table 4. Geographical Levels and Time Ranges of the Multiple Data Sources for Influenza Forecasting

	Level 1	Level 2	Level 3
Geo-level	State	Region	Country
data sources: training period	Twitter: 2011-1~2013-52	ILI-Net: 2009-35~2013-52	FluSurv-NET: 2009-1~2011-12 2011-36~2012-13 2012-36~2013-52

2011 to December 31, 2013 was used for training, while the data from January 1, 2014 to December 31, 2014, was used for the performance evaluation. The locations of the tweets were geocoded by the Carmen geocoder [38]. The forecasting results for the flu outbreaks were validated against the corresponding influenza statistics reported by the CDC.[7] CDC publishes the weekly ILI activity level for each state in the United States based on the proportional level of outpatient visits to healthcare providers for ILI. There are four ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to a salient flu outbreak and is effectively the target when forecasting. An example of a CDC flu outbreak event is: (STATE = "Virginia", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

6.1.2 Parameter Settings and Metrics. There is only one tunable parameter in the proposed HIML, namely the regularization parameter λ. Based on a 10-fold cross validation on the training

[7]CDC: http://www.cdc.gov/flu/weekly/.
[8]Event data codebook of **Conflict and mediation event observations (CAMEO)**: http://phoenixdata.org/description.AccessedFeb2016.

set, it was set as $\lambda = 0.2$. For oHIML, the batch size was set as 200. The logit function was used in the objective functions of our HIML and oHIML.

In the experiment, the event forecasting task was to predict whether or not there would be an event during the next time step for a specific location. For civil unrest datasets, a time step is one day and the location is a city. For disease outbreaks, a time step is one week and the location is a state. A predicted event was matched to a GSR event if both the time and location attributes were matched; otherwise, it was considered a false forecast. To validate the prediction performance, different metrics were adopted: The **True Positive Ratio** (**TPR**) designates the percentage of positive predictions that successfully matched the events that truly happened, while the **False Positive Ratio** (**FPR**) denotes the percentage of positive predictions that were actually false alarms. In addition, a **Receiver operating characteristic** (**ROC**) curve was utilized to evaluate the forecasting performance as its discrimination threshold for each predictive model was varied. Finally, the use of **Area Under ROC Curve** (**AUC**) was also examined as a comprehensive measure of forecasting performance.

6.1.3 Comparison Methods. The following methods were included in the performance comparison:

1. *LASSO [36].* The feature set was the union of the features from different data sources. Only the time period with all the data sources available was retained; samples with any missing value were discarded. The regularization parameter was set as 0.01 based on a 10-fold cross validation on the training set.

2. *LASSO with Interactive Features (LASSO-INT).* The feature set here consisted of two parts: (1) the union of the features from different data sources; and (2) the interactive features among all the features from different data sources. Only the time period with all the data sources available was retained; samples with any missing value were discarded. The regularization parameter was set as 0.01 based on a 10-fold cross validation on the training set.

3. *Incomplete Multi-Source Data Fusion (iMSF) [53].* The classification error here was minimized while a similar selection for the same features across different samples was enforced through a group Lasso over all the samples of each feature. The regularization parameter was set as 0.2 based on a 10-fold cross validation.

4. *Multitask Learning (MTL) [59].* Each task was the event forecasting for the location being predicted. This model utilized a feature set that was the same as that of LASSO-INT. The regularization parameters ρ_1 and ρ_{L2} were set based on a 10-fold cross-validation for each dataset.

5. *AutoInt [46].* A self-attentive neural method to automatically learning representations of high-order combination features. The features are projected into the low-dimensional space and further into stacked multiple interacting layers implemented by self-attentive neural network. The output of the final interacting layer is the low-dimensional representation of learnt combinatorial features, which is further used for the prediction task. Only the time period with all the data sources available was retained; samples with any missing value were discarded. For the hyper-parameters, we used two heads for the self-attention layer and the embedding size was set to 16 following its default setting. The model was trained by the Adam [25] optimizer for three epochs with a batch size of 1,024.

6. *Baseline.* We built a corresponding classifier based on logistic regression for each geographical level. When predicting, the prediction results from these classifiers were comprehensively considered using a voting strategy. Specifically, if the majority of the predictions were "occurrence", then the final prediction was "occurrence", otherwise it was "no-occurrence". The regularization parameters for all the logistic regression models were tuned by a 10-fold cross validation on the training set.

6.2 Performance

In this section, the effectiveness in terms of the AUC and ROC curves are analyzed for all the comparison methods.

6.2.1 AUC on Civil Unrest Datasets. Table 5 summarizes the effectiveness comparison for forecasting civil unrest events for different missing data ratios. The AUC measure has been adopted to quantify the performance. The original percentage of missing data in our data sources was 3%. We manually enlarged this to 30%, 50%, and 70% by randomly reducing the number of dates with complete multiple sources. We used 10 random seeds, and reported the mean and standard deviation of the performance for each method and dataset in Table 5.

The results shown in Table 5 demonstrate that the methods that take into account the hierarchical topology in the data sources performed better. Specifically, the performance of HIML and the online version of HIML (oHIML outperformed the other methods for different missing data ratios in general, with small standard deviations, which indicate the consistency on the performance. The baseline method, LASSO, and LASSO-INT also performed competitively with AUC larger than 0.70 on over four datasets. AutoInt is a method that can also consider the interactions among the features in different geographical hierarchical levels, which also achieved good performance in all the datasets in general. Compared with the other methods, iMSF and MTL had only limited performance for a missing data ratio of 3%. When looking across different missing data ratios, it can be seen that for some datasets and some methods, when missing ratio increases from 3% to 70% the drops may not very obvious, and sometimes there are even slight increase in the performance. The general reason is because using more training data may not necessarily lead to better performance in the situation when training set and test set could have different distributions in practice, which is very common in societal event forecasting [55]. In our experiment, in general there are already sufficient data for model training even with 70% missing ratio. So reducing the missing ratio from 70% to 50%, 30%, and 3% may either improve or decrease the performance, depending on how close the distribution of training set is to that of the test set. HIML and oHIML, similar to iMSF, were able to handle the missing value problem in multiple data sources. They achieved good model robustness against missing values, dropping on average less than 3% when the missing data ratio increased from 3% to 70%. MTL was also not particularly sensitive to the change in missing values. In all, oHIML outperformed the other methods in all of the nine datasets for most data missing ratios, because it is able to effectively consider hierarchical topology and sufficiently leverage the information in different missing patterns.

6.2.2 AUC on the Flu Dataset. Table 6 shows the performance on the metric AUC and training runtime for forecasting influenza outbreaks. Similar to the experiments on civil unrest dataset, the mean and standard deviation of the performance have been obtained and reported in Table 6.

As with the civil unrest datasets, Table 6 shows that for the influenza dataset, the methods that take into account the hierarchical topology in the data sources still perform competitively for the missing data ratio of 21% that was present in the real-world dataset. Specifically, the performance of HIML, oHIML, and the baseline method outperformed both iMSF and MTL. LASSO and LASSO-INT also performed competitively, with AUC surpassing 0.85 for different missing data ratios. Compared with the other methods, MTL suffered from a limited performance on a missing data ratio of 21%. When looking across the different data missing ratios, it is apparent that the methods that cannot handle incomplete input data achieved worse performance against larger missing values. iMSF, HIML, and oHIML achieved a more consistent performance across the full range of missing data ratios. The performance of the other methods, namely LASSO, LASSO-INT, and MTL, dropped more significantly. For example, although the LASSO method achieved a good AUC of 0.9180 at

Table 5. Event Forecasting Performance in Civil Unrest Datasets Based on AUC of ROC

Missing data ratio (3%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5267	0.7476	0.5624	0.8032	0.3148	0.7823	0.5572	0.4693	0.8073
LASSO-INT	0.5268	0.7191	0.5935	0.7861	0.5269	0.777	0.4887	0.5069	0.7543
iMSF	0.4795	0.4611	0.5033	0.7213	0.5	0.5569	0.4486	0.4904	0.5
MTL	0.3885	0.5017	0.5011	0.4334	0.3452	0.4674	0.4313	0.3507	0.5501
Baseline	0.5065	0.7317	0.6148	0.8084	**0.7770**	0.8037	0.7339	0.7264	**0.7846**
AutoInt	0.5310	0.6581	**0.6967**	**0.8307**	0.4604	**0.8046**	0.6827	0.8228	0.7052
HIML	**0.5873**	**0.8353**	0.5705	0.8169	0.7191	0.7973	**0.7478**	**0.8537**	0.7488
oHIML	**0.5601**	**0.8539**	**0.7417**	**0.8230**	**0.8045**	**0.8069**	**0.7484**	**0.8708**	**0.8289**
Missing data ratio (30%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5240	0.7463	0.5842	0.8019	0.3221	0.7812	0.5506	0.4934	0.8043
	± 0.0031	± 0.0015	± 0.014	± 0.0017	± 0.0297	± 0.0021	± 0.0166	± 0.0394	± 0.0086
LASSO-INT	0.5198	0.7155	0.6148	0.7857	0.5032	0.7747	0.4799	0.5367	0.7491
	± 0.0078	± 0.0095	± 0.0158	± 0.0030	± 0.0343	± 0.0035	± 0.0157	± 0.0455	± 0.0137
iMSF	0.4796	0.4611	0.4962	0.76	0.5000	0.5565	0.481	0.4909	0.5000
	± 0.0000	± 0.0000	± 0.0000	± 0.0007	± 0.0000	± 0.0002	± 0.0001	± 0.0003	± 0.0000
MTL	0.5180	0.4478	0.5420	0.5113	0.6111	0.6185	0.681	0.6648	0.4823
	± 0.1181	± 0.0879	± 0.0742	± 0.0694	± 0.1610	± 0.0241	± 0.0334	± 0.2456	± 0.0481
Baseline	0.5150	0.8307	0.3694	**0.8502**	**0.7770**	0.7909	0.7327	**0.8666**	0.6117
	± 0.0047	± 0.0015	± 0.0000	± 0.0012	± 0.0282	± 0.0003	± 0.0084	± 0.0038	± 0.1154
AutoInt	0.5464	0.7153	**0.6780**	0.8312	0.5514	**0.8214**	0.6553	0.8100	0.6576
	± 0.0025	± 0.0306	± 0.0482	± 0.0009	± 0.0355	± 0.0119	± 0.0268	± 0.0114	± 0.0235
HIML	**0.5859**	**0.8334**	0.5622	0.8183	0.7164	0.7955	**0.7459**	0.851	**0.7586**
	± 0.0021	± 0.0018	± 0.0067	± 0.001	± 0.0034	± 0.0003	± 0.0043	± 0.0011	± 0.0146
oHIML	**0.5765**	**0.8492**	**0.6783**	**0.8475**	**0.7948**	**0.7990**	**0.7423**	**0.8618**	**0.7387**
	± 0.0095	± 0.0049	± 0.0279	± 0.0056	± 0.0149	± 0.0015	± 0.0079	± 0.0046	± 0.0243
Missing data ratio (50%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.527	0.7453	0.5967	0.7992	0.3163	0.7828	0.5582	0.5286	0.8027
	± 0.0099	± 0.0021	± 0.0233	± 0.0032	± 0.0086	± 0.0031	± 0.0221	± 0.0709	± 0.0057
LASSO-INT	0.5163	0.7051	0.6189	0.7837	0.4693	0.7738	0.4654	0.5797	0.7541
	± 0.0097	± 0.0139	± 0.0207	± 0.0038	± 0.0622	± 0.0066	± 0.0111	± 0.0762	± 0.0152
iMSF	0.4798	0.4611	0.4961	0.753	0.4901	0.5494	0.4808	0.4865	0.5000
	± 0.0001	± 0.0000	± 0.0009	± 0.0011	± 0.0014	± 0.0009	± 0.0000	± 0.0003	± 0.0000
MTL	0.5005	0.4056	0.5365	0.4975	0.6742	0.6508	0.6272	0.7131	0.5182
	± 0.0867	± 0.1547	± 0.087	± 0.0334	0.1550	± 0.0153	± 0.0675	0.2681	± 0.0745
Baseline	0.5226	**0.8339**	0.5628	**0.8373**	**0.7975**	0.7919	0.7341	**0.8699**	0.5808
	± 0.0122	± 0.0035	0.1277	± 0.0035	± 0.0037	± 0.0010	± 0.0042	± 0.0011	± 0.1046
AutoInt	0.5544	0.7248	**0.6905**	0.8319	0.5365	**0.8058**	0.6745	0.8195	0.6257
	± 0.0146	± 0.0209	± 0.0110	± 0.0032	± 0.0137	± 0.002	± 0.0303	± 0.0245	± 0.1619
HIML	**0.5782**	0.8318	0.5649	0.8170	0.7103	0.7933	**0.7418**	0.8487	**0.7743**
	± 0.0040	± 0.0027	± 0.0115	± 0.0020	± 0.0087	± 0.0002	± 0.0066	± 0.0044	± 0.0167
oHIML	**0.5673**	**0.8423**	**0.6598**	**0.8378**	**0.7837**	**0.7943**	**0.7443**	**0.8648**	**0.7280**
	± 0.0089	± 0.0064	± 0.0235	± 0.0046	± 0.0126	± 0.0025	± 0.0092	± 0.0067	± 0.0179
Missing data ratio (70%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5281	0.7454	0.5970	0.7997	0.3352	0.7859	0.5428	0.6499	0.7812
	± 0.0124	± 0.0060	± 0.0271	± 0.0030	± 0.0494	± 0.0087	± 0.0274	± 0.0296	± 0.0084
LASSO-INT	0.5054	0.6852	0.6113	0.7826	0.4616	0.771	0.4704	0.6394	0.7481
	± 0.0093	± 0.0291	± 0.0136	± 0.0031	± 0.0683	± 0.0132	± 0.0403	± 0.0326	± 0.0166
iMSF	0.4797	0.4482	0.4959	0.7719	0.5000	0.5521	0.4827	0.5222	0.5000
	± 0.0000	± 0.0118	± 0.0026	± 0.0018	± 0.0000	± 0.0043	± 0.0005	± 0.0000	± 0.0000
MTL	0.5212	0.4325	0.4776	0.5829	0.7334	0.6483	0.6101	0.7372	0.4951
	± 0.1035	± 0.0818	± 0.0699	± 0.0741	± 0.0437	± 0.0188	± 0.1286	± 0.1922	± 0.0687
Baseline	0.5324	**0.8313**	0.4636	**0.8414**	**0.8027**	0.7912	**0.7467**	**0.8742**	0.6551
	± 0.0038	± 0.0040	± 0.1575	± 0.0018	± 0.0065	± 0.0007	± 0.0052	± 0.0017	± 0.0283
AutoInt	**0.5557**	0.7637	**0.6795**	**0.8309**	0.4918	**0.8012**	0.6549	0.8046	0.5612
	± 0.0071	± 0.0213	± 0.0286	± 0.0011	± 0.0816	± 0.0073	± 0.0541	± 0.0364	± 0.1167
HIML	**0.5641**	0.8264	0.5610	0.8152	**0.7045**	0.7908	**0.7282**	0.8352	**0.7707**
	± 0.0094	± 0.0068	± 0.0200	± 0.0026	± 0.0123	± 0.0009	± 0.0101	± 0.0073	± 0.0193
oHIML	0.5496	**0.8396**	**0.7322**	0.8237	0.6910	**0.7998**	0.7276	**0.8716**	**0.7034**
	± 0.0119	± 0.0079	± 0.0076	± 0.0054	± 0.0109	± 0.0023	± 0.0175	± 0.0034	± 0.0184

Bold underlined font denotes a best performer while a second best performer is denoted as bold font.

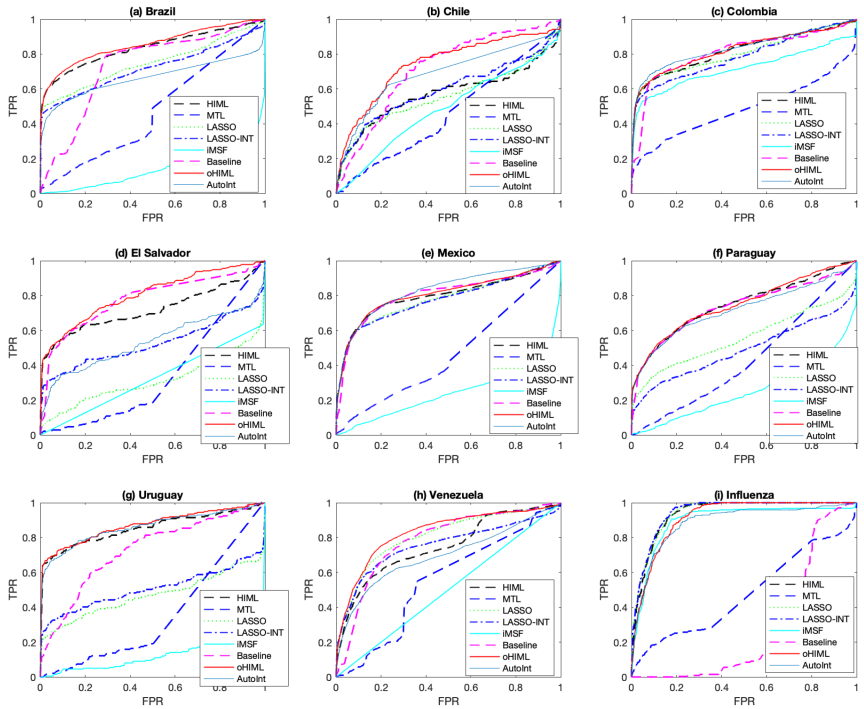

Fig. 4. ROC curves for the performances on different datasets.

a missing data ratio of 21%, this dropped to 0.7640 when the missing data ratios increased to 70% because it could not sufficiently utilize the shared knowledge across different missing patterns and thus large amounts of information were lost. HIML achieved the highest AUC in general. As with the civil unrest datasets, when forecasting influenza outbreaks oHIML once again outperformed all the other methods except HIML consistently for all the different missing data ratios by clear margins, due to its capacity to handle hierarchical topology and interactive missing data values.

6.2.3 Efficiency in Running Time. The rightmost column of Table 6 shows the training time efficiency comparison among HIML, oHIML, and the competing methods for forecasting influenza outbreaks with a 21% missing ratio. The running times on the test set for all the comparison methods are instant (i.e., less than 0.01 second for one prediction) so that are not provided here. According to Table 6, the running time of the baseline method was 31.97 seconds, outperforming the other methods. Online version HIML achieves the second-best runtime of 47.85 seconds, which is more than 18 times speed up comparing with regular full batch HIML and is nearly competitive with the baseline method. LASSO, LASSO-INT, MTL, AutoInt, and HIML were hundreds of seconds on the whole training set. However, the running times achieved by all these methods were only a maximum of 15 minutes for a 4-year-long huge training set for week-wise event forecasting tasks, making this eminently practical for real-world applications. The efficiency evaluation results on civil unrest datasets follow a similar pattern to Table 6 and are not provided due to space limitations.

6.2.4 Event Forecasting Performance on ROC Curves. Figure 4 illustrates the event forecasting performance ROC curves for nine datasets in two domains, namely civil unrest and influenza outbreaks. The Argentina dataset follows a similar pattern to that of Chile and is thus not shown here

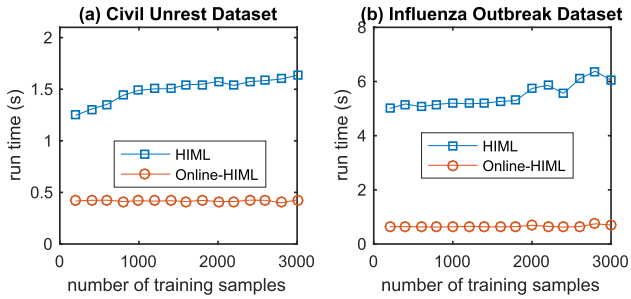

Fig. 5. Scalability of the proposed models for civil unrest dataset and influenza outbreak dataset.

Table 6. Event Forecasting Performance in Influenza Datasets

	Missing data ratio				runtime
Method	21%	30%	50%	70%	(second)
LASSO	**0.9180**	0.9019 ± 0.0099	0.8754 ± 0.0158	0.7640 ± 0.0970	493.92
LASSO-INT	0.9142	0.8887 ± 0.0101	0.8572 ± 0.0390	0.7475 ± 0.1082	508.49
iMSF	0.8949	0.8818 ± 0.0067	0.8616 ± 0.0383	0.7743 ± 0.0770	88.90
MTL	0.6129	0.5920 ± 0.0528	0.5397 ± 0.0250	0.4882 ± 0.0280	223.78
Baseline	0.9044	0.9086 ± 0.0067	0.9025 ± 0.0134	0.9055 ± 0.0128	**31.97**
AutoInt	0.8885	0.8779 ± 0.0157	0.8721 ± 0.0041	0.8680 ± 0.0062	286.00
HIML	**0.9372**	**0.9362 ± 0.0003**	**0.9370 ± 0.0008**	**0.9367 ± 0.0007**	851.83
oHIML	0.9145	**0.9149 ± 0.0032**	**0.9150 ± 0.0083**	**0.9170 ± 0.0140**	**47.85**

Bold underlined font denotes a best performer while a second best performer is denoted as bold font.

to save space. For the eight civil unrest datasets in Figures 4(a)–(h), HIML and oHIML perform the best overall, with ROC curves covering the largest area above the axis. Moreover, the ROC curves for HIML are consistently above those of the other methods in datasets including Brazil, Colombia, El Salvador, Paraguay, and Uruguay as FPR and TPR vary from 0 to 1. For the datasets for Chile and Mexico, HIML, oHIML, AutoInt, LASSO, LASSO-INT, and the Baseline perform similarly, all outperforming the other methods. For the dataset for Venezuela, LASSO, LASSO-INT, and the Baseline method perform better than HIML and AutoInt when FPR is smaller than 0.7, while HIML outperforms the competing methods when $FPR > 0.7$. MTL generally achieves a limited performance, as can be seen in Tables 5 and 6. For the influenza outbreak dataset, as can be seen from Figure 4(i), HIML and oHIML consistently outperforms the other methods with different FPR and TPR values. iMSF, LASSO, AutoInt, and LASSO-INT also achieve quite competitive performances, outperforming MTL by an apparent margin.

6.2.5 Scalability. The training times for the batch-based models are typically sensitive to the size of the training set. Figure 5 illustrates the impact of scalability on the number of training samples needed by the proposed approach for the civil unrest dataset and flu dataset, respectively. Here we use the Mexico dataset to represent the civil unrest dataset, the rest datasets follow the same trend.

As shown in Figure 5(a), for the civil unrest dataset, the runtime for training HIML is linear in the number of training samples, starting from only 1.2 seconds with 1,000 samples and rising to 1.6 seconds with 15,000 samples. Unlike batch-based models, the training times for the oHIML were not sensitive to the number of training samples utilized, with a relatively constant runtime of around 0.4 seconds.

On the flu dataset, shown in Figure 5(b), the runtimes for both HIML and oHIML were longer than for the civil unrest dataset due to the larger scale of the data. The runtime for training HIML once again increased linearly with the number of training samples, starting from 5 seconds with 100 samples and rising to 6 seconds with 3,000 samples. The runtimes of the online versions of the proposed models were consistently around 0.7 seconds when the number of training samples was varied from 100 to 3,000.

7 CONCLUSIONS

The occurrence of significant societal events are influenced and determined by various aspects of society, e.g., economics, politics, and culture. To accommodate all the intricacies involved in the underlying domain, event forecasting should be based on multiple data sources but existing models still suffer from several challenges. This article has proposed novel hierarchical and incremental multi-source feature learning models that characterize the feature dependence, feature sparsity, and interactions among missing values. Efficient batch- and incremental-based algorithm for parameter optimization are proposed. Extensive experiments on 10 real-world datasets with multiple data sources demonstrated that the proposed model outperforms other comparison methods in different ratios of missing values.

REFERENCES

[1] Francis Bach. 2008. Exploring large feature spaces with hierarchical multiple kernel learning. arXiv:0809.1493. Retrieved from https://arxiv.org/abs/0809.1493.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3, 1 (2011), 1–122.

[4] Alasdair Brown, Dooruj Rambaccussing, J. James Reade, and Giambattista Rossi. 2018. Forecasting with social media: Evidence from tweets on soccer matches. *Economic Inquiry* 56, 3 (2018), 1748–1763.

[5] Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhuo Chen, Patrick Butler, Elaine O. Nsoesie, Sumiko R. Mekaru, John S. Brownstein, Madhav Marathe, and N. Ramakrishnan 2014. Forecasting a moving target: Ensemble models for ILI case count predictions. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. 262–270.

[6] Nam Hee Choi, William Li, and Ji Zhu. 2010. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* 105, 489 (2010), 354–364.

[7] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7, Mar (2006), 551–585.

[8] Wenjuan Cui, Pengfei Wang, Yi Du, Xin Chen, Danhuai Guo, Jianhui Li, and Yuanchun Zhou. 2017. An algorithm for event detection based on social media data. *Neurocomputing* 254 (2017), 53–58.

[9] G. Ditzler, J. LaBarck, J. Ritchie, G. Rosen, and R. Polikar. 2018. Extensions to online feature selection using bagging and boosting. *IEEE Transactions on Neural Networks and Learning Systems* 29, 9 (Sept 2018), 4504–4509. DOI: https://doi.org/10.1109/TNNLS.2017.2746107

[10] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, Jul (2011), 2121–2159.

[11] Yaakov Engel, Shie Mannor, and Ron Meir. 2004. The kernel recursive least squares algorithm. *IEEE Transactions on Signal Processing* 52, 8 (2004), 2275–2285.

[12] Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2020. Cross-GCN: Enhancing graph convolutional network with k-Order feature interactions. arXiv:2003.02587. Retrieved from https://arxiv.org/abs/2003.02587.

[13] Sujuan Gao. 2004. A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. *Statistics in Medicine* 23, 2 (2004), 211–219.

[14] Yuyang Gao and Liang Zhao. 2018. Incomplete label multi-task ordinal regression for spatial event scale forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[15] Yuyang Gao, Liang Zhao, Lingfei Wu, Yanfang Ye, Hui Xiong, and Chaowei Yang. 2019. Incomplete label multi-task deep learning for spatio-temporal event subtype forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3638–3646.

[16] Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.
[17] Susan E. Hardy, Heather Allore, and Stephanie A. Studenski. 2009. Missing data: A special challenge in aging research. *Journal of the American Geriatrics Society* 57, 4 (2009), 722–729.
[18] Asad Haris, Daniela Witten, and Noah Simon. 2014. Convex modeling of interactions with strong heredity. arXiv:1410.3517. Retrieved from https://arxiv.org/abs/1410.3517.
[19] Frank E. Harrell. 2013. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Science & Business Media.
[20] Jose M. Hernandez-lobato, Neil Houlsby, and Zoubin Ghahramani. 2014. Probabilistic matrix factorization with non-random missing data. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*. 1512–1520.
[21] Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Dawei Yin, and Nitesh Chawla. 2019. MiST: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *The World Wide Web Conference*. 717–728.
[22] Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan. 2015. Generalized hierarchical kernel learning. *The Journal of Machine Learning Research* 16, 1 (2015), 617–652.
[23] V. Roshan Joseph. 2006. A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics* 48, 2 (2006), 219–229.
[24] Nathan Kallus. 2014. Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*. IW3C2, 625–630.
[25] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980.
[26] Yan Li, Lu Wang, Jiayu Zhou, and Jieping Ye. 2019. Multi-task learning based survival analysis for multi-source block-wise missing data. *Neurocomputing* 364 (2019), 95–107.
[27] Yanchao Li, Yongli Wang, Qi Liu, Cheng Bi, Xiaohui Jiang, and Shurong Sun. 2019. Incremental semi-supervised learning on streaming data. *Pattern Recognition* 88 (2019), 383–396.
[28] Yan Li, Tao Yang, Jiayu Zhou, and Jieping Ye. 2018. Multi-task learning based survival analysis for predicting alzheimer's disease progression with multi-source block-wise missing data. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 288–296.
[29] Michael Lim and Trevor Hastie. 2013. Learning interactions through hierarchical group-lasso regularization. arXiv:1308.2719. Retrieved from https://arxiv.org/abs/1308.2719.
[30] Jinghua Liu, Yaojin Lin, Yuwen Li, Wei Weng, and Shunxiang Wu. 2018. Online multi-label streaming feature selection based on neighborhood rough set. *Pattern Recognition* 84 (2018), 273–287. DOI: https://doi.org/10.1016/j.patcog.2018.07.021
[31] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11, Jan (2010), 19–60.
[32] Yasuko Matsubara, Yasushi Sakurai, Christos Faloutsos, Tomoharu Iwata, and Masatoshi Yoshikawa. 2012. Fast mining and forecasting of complex time-stamped events. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 271–279.
[33] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Tozammel Hossain, et al. 2019. Tensor-based method fortemporal geopolitical event forecasting. *ICML Workshop on Learning and Reasoning with Graph-Structured Data* (2019).
[34] MITRE. [n.d.]. Retrieved from Feb 2016 from http://www.mitre.org/.
[35] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *International AAAI Conference on Weblogs and Social Media* 11 (2010), 122–129.
[36] Joseph O. Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. 2012. Genomic selection using regularized linear regression models: Ridge regression, Lasso, elastic net and their extensions. In *BMC Proceedings*, Vol. 6. S10.
[37] Hristo Paskov, Alex Paskov, and Robert West. 2020. Learning high order feature interactions with fine control kernels. arXiv:2002.03298. Retrieved from https://arxiv.org/abs/2002.03298.
[38] Michael J. Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PLoS One* 9, 8 (2014), e103408.
[39] Simon Perkins and James Theiler. 2003. Online feature selection using grafting. In *Proceedings of the 20th International Conference on International Conference on Machine Learning*. 592–599.
[40] Xiaofan Que, Yazhou Ren, Jiayu Zhou, and Zenglin Xu. 2017. Regularized multi-source matrix factorization for diagnosis of alzheimer's disease. In *International Conference on Neural Information Processing*. Springer, 463–473.

[41] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang-Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, and David Mares. 2014. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *Knowledge Discovery and Data Mining* . ACM, 1799–1808.

[42] Deboleena Roy, Priyadarshini Panda, and Kaushik Roy. 2020. Tree-CNN: A hierarchical deep convolutional neural network for incremental learning. *Neural Networks* 121 (2020), 148–160.

[43] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. 851–860.

[44] Minglai Shao, Jianxin Li, Feng Chen, Hongyi Huang, Shuai Zhang, and Xunxun Chen. 2017. An efficient approach to event detection and forecasting in dynamic multivariate social media networks. In *Proceedings of the 26th International Conference on World Wide Web*. 1631–1639.

[45] Lei-Lei Shi, Lu Liu, Yan Wu, Liang Jiang, and James Hardy. 2017. Event detection and user interest discovering in social media data streams. *IEEE Access* 5 (2017), 20953–20964.

[46] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1161–1170.

[47] Zichun Su and Jialin Jiang. 2020. Hierarchical gated recurrent unit with semantic attention for event prediction. *Future Internet* 12, 2 (2020), 39.

[48] Dawei Wang and Wei Ding. 2015. A hierarchical pattern learning framework for forecasting extreme weather events. In *2015 IEEE International Conference on Data Mining*. IEEE, 1021–1026.

[49] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large scale incremental learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 374–382.

[50] Fei Xue and Annie Qu. 2019. Integrating multi-source block-wise missing data in model selection. arXiv:1901.03797. Retrieved from https://arxiv.org/abs/1901.03797.

[51] Haichuan Yang, Ryohei Fujimaki, Yukitaka Kusumura, and Ji Liu. 2016. Online feature selection: A limited-memory substitution algorithm and its asynchronous parallel variation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1945–1954.

[52] Kui Yu, Xindong Wu, Wei Ding, and Jian Pei. 2014. Towards scalable and accurate online feature selection for big data. In *2014 IEEE International Conference on Data Mining*. IEEE, 660–669.

[53] Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. 2012. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Knowledge Discovery and Data Mining*. ACM, 1149–1157.

[54] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 984–992.

[55] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.

[56] Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS One* 9, 10 (2014), e110206.

[57] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Spatiotemporal event forecasting in social media. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 963–971.

[58] Liang Zhao, Jiangzhuo Chen, Feng Chen, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2015. SimNest: Social media nested epidemic simulation via online semi-supervised deep learning. In *2015 IEEE International Conference on Data Mining*. IEEE, 639–648.

[59] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1503–1512.

[60] Liang Zhao, Junxiang Wang, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2017. Spatial event forecasting in social media with geographically hierarchical regularization. *Proceedings of the IEEE* 105, 10 (2017), 1953–1970.

[61] Liang Zhao, Junxiang Wang, and Xiaojie Guo. 2018. Distant-supervision of heterogeneous multitask learning for social event forecasting with multilingual indicators. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 4498–4505.

[62] Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2085–2094.

[63] Wenliang Zhong and James Kwok. 2014. Fast stochastic alternating direction method of multipliers. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*. 46–54.
[64] Jing Zhou, Dean Foster, Robert Stine, and Lyle Ungar. 2005. Streaming feature selection using alpha-investing. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 384–393.
[65] Lihua Zhou, Guowang Du, Ruxin Wang, Dapeng Tao, Lizhen Wang, Jun Cheng, and Jing Wang. 2019. A tensor framework for geosensor data forecasting of significant societal events. *Pattern Recognition* 88 (2019), 27–37.

Received March 2020; revised March 2021; accepted May 2021