Hierarchical Incomplete Multi-source Feature Learning for Spatiotemporal Event Forecasting

Liang Zhao
Virginia Tech
liangz8@vt.edu

Jieping Ye
University of Michigan
jpye@umich.edu

Feng Chen
University at Albany - SUNY
fchen5@albany.edu

Chang-Tien Lu
Virginia Tech
ctlu@vt.edu

Naren Ramakrishnan
Virginia Tech
naren@cs.vt.edu

ABSTRACT

Forecasting significant societal events is an interesting and challenging problem as it taking into consideration 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) missing values, and 3) characterization of structured feature sparsity. This paper 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. Extensive experiments on 10 datasets in different domains demonstrate the effectiveness and efficiency of the proposed model.

Keywords

Event forecasting; multiple data sources; feature selection.

1. INTRODUCTION

Significant societal events such as disease outbreaks and mass protests have a tremendous impact on our entire society, which strongly motivates anticipating their occurrences in advance. For example, according to a recent World Health Organization (WHO) report [26], seasonal influenza alone is estimated to result in around 4 million cases of severe illness and about 250,000 to 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 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 economy and social media would be

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939847

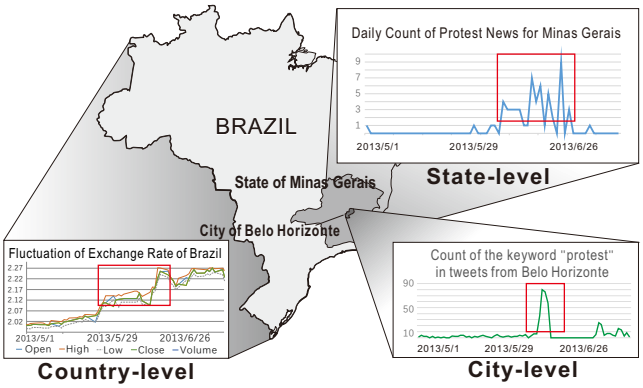


Figure 1: Predictive indicators from multiple data sources with different geographical levels during the "Brazilian Spring" civil unrest movement.

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 from different hierarchical topologies, features from higher levels influences those lower down. 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.

In order to simultaneously address all these technical challenges, this paper presents a novel model named hierarchical incomplete multi-source feature learning (HIML). 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 can not 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 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.

The main contributions of our study are summarized below. We:

- **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 on 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 multitask 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.

- **Conduct extensive experiments for performance evaluations.** The proposed method was 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 best of the existing methods along multiple metrics.

The rest of this paper is organized as follows. Section 2 reviews background and related work, and Section 3 introduces the problem setup. Section 4 presents our HIML model and an efficient model parameter optimization algorithm. The experiments on 10 real-world datasets are presented in Section 5, and the paper concludes with a summary of the research in Section 6.

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 [19] and disease outbreaks [23]. 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 [15] and stock market movements [1]. Few existing approaches can provide true spatiotemporal resolution for the predicted events [21]. For example, Gerber utilized a logistic regression model for spatiotemporal event forecasting [6]. Zhao et al. [24] designed a multitask learning framework that models forecasting tasks in related geolocations concurrently. Zhao et al. [22] also designed a new predictive model that jointly characterizes the temporal evolution of both the semantics and geographical burstiness of social media content.

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 [3]. 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 [18]. Kallus explores the predictive power of news, blogs, and social media for political event forecasting [12]. However, although these models utilize multiple data sources that can be used to indicate a number of different aspects of future events, they typically ignore the potential relationships, topology, and hierarchy among these multi-source features.

Missing values in multiple data sources. The prevention and management of missing data has been discussed and investigated in earlier work [7]. One category of work focuses on estimating missing entries based on the observed values [5]. 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 [10], but their method is restricted to non-random missing values. Yuan et al. [20] utilized multitask learning to learn a consistent feature selection pattern across different missing groups. However, none of these approaches focus specifically on missing values in hierarchical multiple data sources.

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. [9] 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 [4]. 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 [13] work with a combination of continuous and categorical variables, while Haris et al. [8] explore different types of norms. 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** $l = (l_1, l_2, \cdots, l_N)$, which is an array that configures each location l_n in each geo-level n in terms of a parent-child hierarchy such that $l_{n-1} \sqsubseteq l_n (n = 2, \cdots, N)$, 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", "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 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 $l = (l_1, \cdots, l_N)$ 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 $l = (l_1, \cdots, l_N)$ at time t, given data sources under N geographical levels $\{X_{t,l_1}, \cdots, X_{t,l_N}\}$, 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:

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

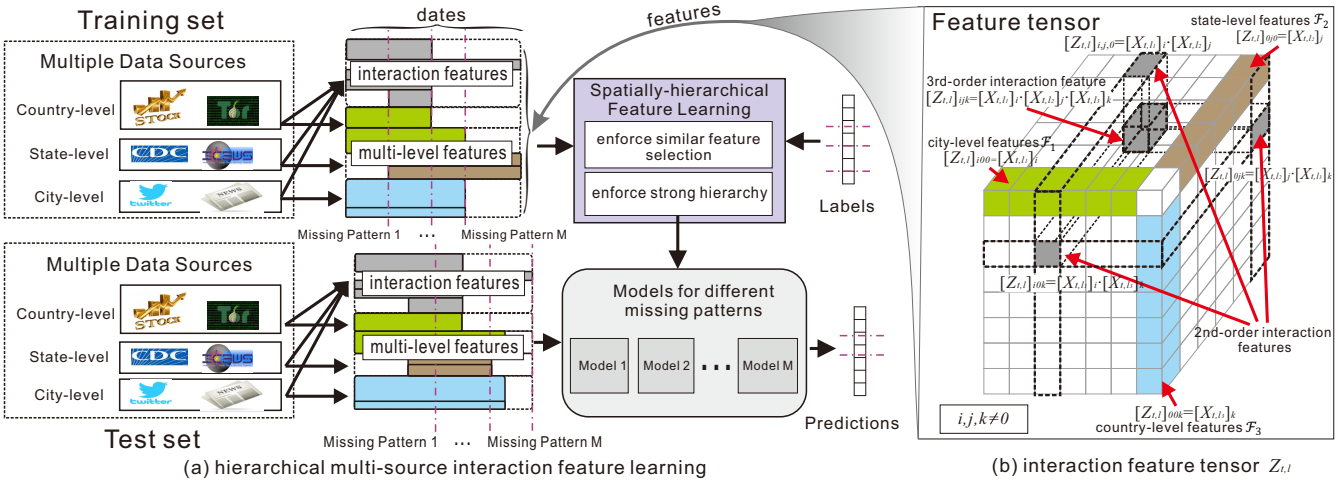
where f is the forecasting model.

In Problem (1), input variables $\{X_{t,l_1}, \cdots, X_{t,l_N}\}$ 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:

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

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



(a) hierarchical multi-source interaction feature learning (b) interaction feature tensor $Z_{t,l}$

Figure 2: A schematic view of hierarchical incomplete multi-source feature learning (HIML) model.

Equation (2), we get:

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

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

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

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 2(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, blue, and brown nodes in Figure 2(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 2(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 2(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.

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 $A\odot B = \sum_{i_1,\cdots,i_N} A_{i_1,\cdots,i_N}\cdot B_{i_1,\cdots,i_N}$. For simplicity, this can be denoted as $A\odot B = \sum_{\vec{i}} A_{\vec{i}}\cdot B_{\vec{i}}$, where $\vec{i} = \{i_1, i_2, \cdots, i_N\}$.

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:

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

where $h(\cdot)$ is a convex and differentiable mapping function. In this paper, 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. A careful and effective consideration and utilization of this structure is 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 2. 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 4.4. 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 [8, 11] as this can handle the interactions between two sets of main-effect variables. Here, we introduce their definition as follows:

Lemma 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 Nth-order interaction variable is included in the model, then all of its nth-order ($2 \leq n < N$) interactive variables and main-effect variables are included.*

Proof. According to Lemma 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 $N = 3$, Theroem 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 trade off 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. [8], which enforces enforces a hierarchical structure under a 2nd-order strong hierarchy.

4.2 Missing Features Values in the Presence of Interactions

As shown in Figure 2(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 the 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 3rd-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:

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

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 $\vec{i} = \{i_1, \cdots, i_n\}$. Also denote $\vec{i}_{-n} = \{i_1, \cdots, i_{n-1}, i_{n+1}, \cdots, i_N\}$. $\mathcal{W}_{G(\vec{i})} \equiv \bigcup_m^M [\mathcal{W}_m]_{(\vec{i})}$ 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:

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

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

4.4 Parameter Optimization

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:

$$\min_{\mathcal{W},\Phi} \mathcal{L}_M(\mathcal{W}) + \lambda_0 \sum_{\min(\vec{i})\neq 0} \|\Phi^{(0)}_{G(\vec{i})}\|_F + \sum_{n=1}^{N} \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \|\Phi^{(n)}_{G(\vec{i}_{-n})}\|_F$$
$$s.t.\ \ \mathcal{W}_m = \Phi^{(n)}_m,\ m = 1, \cdots, M;\ n = 1, \cdots, N. \quad (12)$$

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 (12) 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 (12) is:

$$L_\rho(\mathcal{W}, \Phi, \Gamma) = \mathcal{L}_M(\mathcal{W}) + \sum_{m,n}^{M,N} tr(\Gamma^{(n)}_m (\mathcal{W}_m - \Phi^{(n)}_m)) + \sum_{n=1}^{N} \lambda_n \sum_{I_{-n}\neq \vec{0}} \|\Phi^{(n)}_{G(\vec{i}_{-n})}\|_F + \rho/2 \sum_{m,n}^{M,N} \|\mathcal{W}_m - \Phi^{(n)}_m\|_F^2 + \lambda_0 \sum_{\min(\vec{i})\neq 0} \|\Phi^{(0)}_{G(\vec{i})}\|_F \quad (13)$$

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 (13) 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. [2]. 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:

$$\mathcal{W}_m = \operatorname*{argmin}_{\mathcal{W}_m} \mathcal{L}_M(\mathcal{W}) + \frac{N \cdot \rho}{2} \| \frac{1}{N} \sum_n^N \Phi^{(n)}_m - \frac{1}{N\rho} \sum_n^N \Gamma^{(n)}_m - \mathcal{W}_m \|_F^2 \quad (14)$$

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:

$$\Phi^{(n)}_m \leftarrow \operatorname*{argmin}_{\Phi^{(n)}_m} \frac{\rho}{2} \|\Phi^{(n)}_m - \mathcal{W}_m - \frac{\Gamma^{(n)}_m}{\rho}\|_F^2 + \lambda_n \sum_{\vec{i}_{-n}\neq \vec{0}} \|\Phi^{(n)}_{G(\vec{i}_{-n})}\|_F \quad (15)$$

which is a regression problem with ridge regularization. This problem can be efficiently using the proximal operator [2].

3. Update $\Phi^{(0)}_m$.

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

$$\Phi^{(0)}_m \leftarrow \operatorname*{argmin}_{\Phi^{(0)}_m} \frac{\rho}{2} \|\Phi^{(0)}_m - \mathcal{W}_m - \frac{\Gamma^{(0)}_m}{\rho}\|_F^2 + \lambda_0 \sum_{\min(\vec{i})\neq 0} \|\Phi^{(0)}_{G(\vec{i})}\|_F \quad (16)$$

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

4. Update $\Gamma^{(n)}_m$.

The Lagrangian multiplier is updated as follows:

$$\Gamma^{(n)}_m \leftarrow \Gamma^{(n)}_m + \rho(\mathcal{W}_m - \Phi^{(n)}_m) \quad (17)$$

Algorithm 1 Hierarchical Incomplete Feature Learning

Require: Z, Y, λ
Ensure: solution $\mathcal{W}$
1: Initialize $\rho = 1$, $\mathcal{W}_m, \Gamma, \Phi = \mathbf{0}$.
2: Choose $\varepsilon_s > 0, \varepsilon_r > 0$.
3: **repeat**
4: **for** $m \leftarrow 1, \ldots, M$ **do**
5: $\mathcal{W}_m \leftarrow$ Equation (14)
6: **for** $n \leftarrow 0, \ldots, N$ **do**
7: $\Phi^{(n)}_m \leftarrow$ Equation (16) # Equation (15) if $n = 0$
8: $\Gamma^{(n)}_m \leftarrow$ Equation (17)
9: **end for**
10: **end for**
11: $s = \rho \|\{\Phi^{(n)}_m - \Psi^{(n)}_m\}^{M,N}_{m,n}\|_F$ # Calculate dual residual
12: $r = \|\{\mathcal{W}^{(n)}_m - \Psi^{(n)}_m\}^{M,N}_{m,n}\|_F$ # Calculate primal residual
13: **if** $r > 10s$ **then**
14: $\rho \leftarrow 2\rho$ # Update penalty parameter
15: **else if** $10r < s$ **then**
16: $\rho \leftarrow \rho/2$
17: **else**
18: $\rho \leftarrow \rho$
19: **end if**
20: **until** $r < \varepsilon^r$ and $s < \varepsilon^s$

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 multisource 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 multisource feature learning [20]:

$$\min_W \sum_m \frac{1}{2C_m} \sum_n^{C_m} \|Y_n - W_m \cdot Z_n\|_F^2 + \lambda_0 \sum_i^{|\mathcal{F}|} \|W_{G(i)}\|_F \quad (18)$$

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 [16]:

$$\min_W \frac{1}{2C} \sum_i^C \|Y_i - W \cdot Z_i\|_F^2 + \lambda_0 \sum_i^{|\mathcal{F}|} |W_i| \quad (19)$$

where C is the count of observations.

3. **Generalization of interactive LASSO**. Let $N = 2$ and $M = 1$, which means there are only 2 hierarchical levels in data without missing value. HIML is thus reduced to a regression with regularization based on 2nd-order strong hierarchy [8]:

$$\min_W \frac{1}{2C} \sum_i^C \|Y_i - W \odot Z_i\|_F^2 + \lambda_0 \sum_{i,j\neq 0} |W_{i,j}| + \lambda_1 \sum_{j=1}^{|\mathcal{F}_1|} \|W_{\cdot,j}\|_F + \lambda_2 \sum_{i=1}^{|\mathcal{F}_2|} \|W_{i,\cdot}\|_F \quad (20)$$

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

5. 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 is 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 16.0GB memory.

5.1 Experimental Setup

5.1.1 Datasets and Labels

Table 1: Labels of different datasets. (CU=civil unrest; FLU=influenza-like-illnesses).

Dataset	Domain	Label sources [1]	#Events
Argentina	CU	Clarín; La Nación; Infobae	1306
Brazil	CU	O Globo; O Estado de São Paulo; Jornal do Brasil	3226
Chile	CU	La Tercera; Las Últimas Notícias; El Mercurio	706
Colombia	CU	El Espectador; El Tiempo; El Colombiano	1196
El Salvador	CU	El Diáro de Hoy; La Prensa Gráfica; El Mundo	657
Mexico	CU	La Jornada; Reforma; Milenio	5465
Paraguay	CU	ABC Color; Ultima Hora; La Nacíon	1932
Uruguay	CU	El País; El Observador	624
Venezuela	CU	El Universal; El Nacional; Ultimas Notícias	3105
U.S.	FLU	CDC Flu Activity Map	1027

In this paper, 10 different datasets from different domains were used for the experimental evaluations, as shown in Table 1. Among these, 9 datasets were used for event forecasting under the civil unrest domain for 9 different countries in Latin America. For these datasets, 4 data sources from different geographical levels were adopted as the model 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.

[1]In addition to the top 3 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; 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

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 [18]
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 [17]

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 [18]. The event forecasting results were validated against a labeled event set, known as the gold standard report (GSR), exclusively provided by MITRE [14]. GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin America [18], 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").

The other dataset was collected to track influenza outbreaks in the United States and consists of 3 data sources from different geographical levels, which are Twitter, ILI-Net[5], and FluSurv-NET[6], as shown in Table 3. 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, 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 [17]. The forecasting results for the flu outbreaks were validated against the corresponding influenza statistics reported by the Centers for Disease Control and Prevention (CDC)[7]. CDC publishes the weekly influenza-like illness (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 4 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").

5.1.2 Parameter Settings and Metrics

There is only one tunable parameter in the proposed HIML model, namely the regularization parameter λ. Based on

[5]ILI-NET:https://wwwn.cdc.gov/ilinet/

[6]FluSurv-NET:http://www.cdc.gov/flu/weekly/overview.htm#Hospitalization

[7]CDC: http://www.cdc.gov/flu/weekly/.

[8]Event data codebook of Conflict and mediation event observations (CAMEO): http://phoenixdata.org/description.AccessedFeb2016.

Table 3: Geographical levels and time ranges of the multiple data sources

	Civil Unrest (yyyy-mm-dd)			Influenza (yyyy-week)		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Geo-level	City	State	Country	State	Region	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	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

Table 4: Event forecasting performance in civil unrest datasets based on area under the curve (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	**0.777**	**0.8037**	0.7339	0.7264	**0.7846**
HIML	**0.5873**	**0.8353**	0.5705	**0.8169**	0.7191	0.7973	**0.7478**	**0.8537**	0.7488
Missing data ratio (30%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5035	0.7362	0.588	0.8412	0.3785	0.7896	0.478	0.6749	0.681
LASSO-INT	0.4976	0.6361	0.5912	0.8151	0.3852	0.7622	0.426	0.7177	0.6428
iMSF	0.4797	0.4611	0.4959	0.6845	0.5	0.5569	0.4811	0.4898	0.5
MTL	0.4207	0.5156	0.5023	0.5978	0.3413	0.4666	0.4318	0.347	0.4397
Baseline	0.5012	0.7724	**0.6245**	0.8032	**0.7626**	0.7598	0.738	0.8205	**0.7621**
HIML	**0.5854**	**0.8497**	0.6072	**0.8449**	0.726	**0.7907**	**0.7471**	**0.8576**	0.7378
Missing data ratio (50%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5128	0.7461	0.5301	0.8167	0.3139	0.7552	0.5285	0.5992	0.6678
LASSO-INT	0.504	0.6145	0.5537	0.7339	0.4283	0.7309	0.4745	0.5396	0.6155
iMSF	0.4796	0.4611	0.4962	0.7467	0.4899	0.5488	0.4804	0.487	0.5
MTL	0.5104	0.4818	0.4715	0.65	0.3375	0.4744	0.436	0.3578	0.3839
Baseline	0.5101	0.7717	**0.639**	0.8142	**0.7665**	**0.8079**	0.7324	0.8112	**0.7759**
HIML	**0.5795**	**0.8463**	0.548	**0.8432**	0.7126	0.7892	**0.7477**	**0.856**	0.7176
Missing data ratio (70%)									
Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5162	0.6674	0.5947	0.8344	0.2597	0.7485	0.4075	0.2652	0.6699
LASSO-INT	0.4691	0.5557	0.5469	0.7167	0.2116	0.7	0.3808	0.2256	0.6503
iMSF	0.4796	0.4611	0.5503	0.7855	0.5	0.557	0.4795	0.5221	0.5
MTL	0.4128	0.5023	0.5069	0.6195	0.3323	0.4702	0.4283	0.3569	0.6464
Baseline	0.5188	0.7741	**0.6059**	0.8121	**0.7557**	**0.8097**	0.7136	0.72	0.6993
HIML	**0.5484**	**0.7812**	0.3887	**0.8416**	0.7181	0.8001	**0.7146**	**0.8453**	**0.716**

a 10-fold cross validation on the training set, it was set as $\lambda = 0.2$. The logit function was used in Equation (11) for our HIML.

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.

5.2 Performance

In this section, the effectiveness on the AUC and ROC curves are analyzed for all the comparison methods, including LASSO [16], LASSO with Interactive Features (LASSO-INT), Incomplete Multi-Source Data Fusion (iMSF) [20], Multitask Learning (MTL) [24], and the Baseline. Their parameter settings are described in our supplementary materials[1].

5.2.1 AUC on civil unrest datasets

Table 4 summarizes the effectiveness and robustness 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.

The results shown in Table 4 demonstrate that the methods that take into account the hierarchical topology in the data sources performed better. Specifically, the performance of HIML and the baseline method outperformed the other methods for different missing data ratios. LASSO and LASSO-INT also performed competitively with AUC larger than 0.75 on four datasets. 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 the methods that were best able to handle incomplete input data achieved better robustness against missing values. The performance of LASSO dropped an average 10%, considerably more than iMSF, which dropped less than 3%, when the missing data

[1] http://people.cs.vt.edu/liangz8/materials/papers/KDD_Multi_sourceAddon.pdf

ratio increased from 3% to 70%. HIML, similar to iMSF, was able to handle the missing value problem in multiple data sources. It also achieved an outstanding 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, partially due to its ability to handle the lack of data by sharing the information across different tasks. In all, HIML outperformed all the other methods in 6 out of the 9 datasets for all the different missing data ratios by 6% on average, and achievd the second best performance on the other 3 datasets. This is because HIML effectively handles the two crucial challenges, namely hierarchical topology and interactive missing values.

5.2.2 AUC on the flu dataset

Table 5 shows the performance on the metric AUC and training runtime for forecasting influenza outbreaks.

As with the civil unrest datasets, Table 5 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 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 missing data ratios, it is apparent that the methods that were best able to handle incomplete input data not surprisingly achieved better robustness against missing values. For example, iMSF performed consistently well, with AUCs between 0.86 and 0.89 even when the missing data ratio increased from 21% to 70% because it was able to cope with the missing value problem in multiple data sources. As with iMSF, HIML also achieved a consistent performance across the full range of missing data ratios. MTL was also not quite as sensitive to changes in the missing data values, which mirrors its performance on the civil unrest datasets, shown in Table 4. The performance of the other methods, namely LASSO, LASSO-INT, and Baseline, dropped more significantly. For example, although the Baseline method achieved a good AUC of 0.9044 at a missing data ratio of 21%, this dropped to 0.4359 when the missing data ratio increased to 70% because it could not sufficiently utilize the shared knowledge across different missing patterns and thus large amounts of information were lost. As with the civil unrest datasets, when forecasting influenza outbreaks HIML once again outperformed all the other methods consistently for all the different missing data ratios by clear margins, due to its capacity to handle hierarchical topology and interactive missing data values.

5.2.3 Efficiency on running time

The rightmost column of Table 5 shows the training time efficiency comparison among HIML and the competing methods for forecasting influenza outbreaks with 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 5, the running time of the baseline method was 31.97, outperforming the other methods. LASSO, LASSO-INT, MTL, 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 of Table 5 and are not provided due to space limitations.

Table 5: Event forecasting performance in influenza datasets

	Missing data ratio				runtime
Method	21%	30%	50%	70%	(second)
LASSO	0.9180	0.9056	0.9036	0.8753	493.92
LASSO-INT	0.9142	0.9027	0.9073	0.8403	508.49
iMSF	0.8949	0.8899	0.8930	0.8628	88.90
MTL	0.6129	0.5303	0.6253	0.5568	223.78
Baseline	0.9044	0.9045	0.8562	0.4359	**31.97**
HIML	**0.9372**	**0.9368**	**0.9364**	**0.9357**	851.83

5.2.4 Event forecasting performance on ROC curves

Figure 3 illustrates the event forecasting performance ROC curves for 9 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 to save space. For the 8 civil unrest datasets in Figures 3(a)-(h), HIML performs 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, 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 best when FPR is smaller than 0.7, while HIML outperforms the other methods when $FPR > 0.7$. MTL generally achieves a limited performance, but its performance is robust against missing ratio, as can be seen in Tables 4 and 5. For the influenza outbreak dataset, as can be seen from Figure 3(i), HIML consistently outperforms the other methods with different FPR and TPR values. iMSF, LASSO, and LASSO-INT also achieve quite competitive performances, outperforming the baseline method and MTL by an apparent margin.

6. CONCLUSIONS

Significant societal events are prevalent in multiple 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 paper has proposed a novel group-Lasso-based feature learning model that characterizes the feature dependence, feature sparsity, and interactions among missing values. An efficient algorithm for parameter optimization is proposed to ensure global optima. 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.

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and dis-

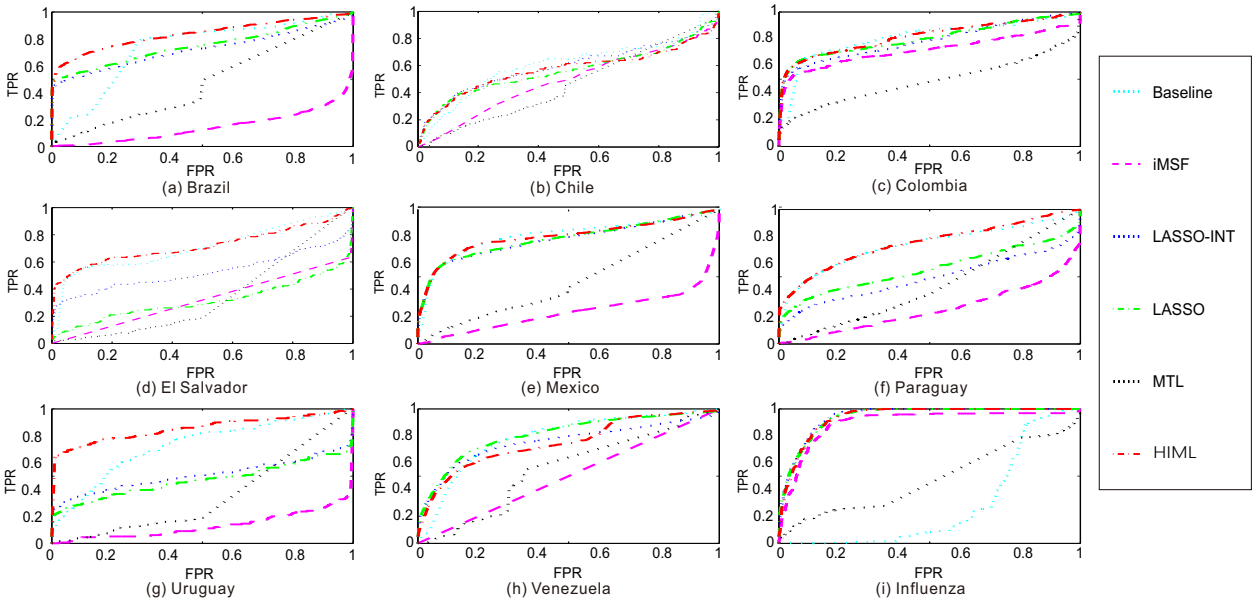


Figure 3: Receiver operating characteristic (ROC) curves for the performances on different datasets

7. REFERENCES

[1] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[3] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. Marathe, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. In *SDM 2014*, pages 262–270, 2014.

[4] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

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

[6] M. S. Gerber. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[7] S. E. Hardy, H. Allore, and S. A. Studenski. Missing data: A special challenge in aging research. *Journal of the American Geriatrics Society*, 57(4):722–729, 2009.

[8] A. Haris, D. Witten, and N. Simon. Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*, 2014.

[9] F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media, 2013.

[10] J. M. Hernandez-lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *ICDM 2014*, pages 1512–1520, 2014.

[11] V. R. Joseph. A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229, 2006.

[12] N. Kallus. Predicting crowd behavior with big public data. In *WWW 14 Companion*, pages 625–630. IW3C2, 2014.

[13] M. Lim and T. Hastie. Learning interactions through hierarchical group-lasso regularization. *arXiv preprint arXiv:1308.2719*, 2013.

[14] MITRE. http://www.mitre.org/. accessed Feb 2016.

[15] B. O'Connor, R. Balasubramanyan, and et al. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM 2010*, 11:122–129, 2010.

[16] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho. Genomic selection using regularized linear regression models: ridge regression, Lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, page S10. BioMed Central Ltd, 2012.

[17] M. J. Paul and M. Dredze. Discovering health topics in social media using topic models. *PLoS One*, 9(8):e103408, 2014.

[18] N. Ramakrishnan, P. Butler, S. Muthiah, et al. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *KDD 2014*, pages 1799–1808. ACM, 2014.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[20] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *KDD 2012*, pages 1149–1157. ACM, 2012.

[21] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one*, 9(10):e110206, 2014.

[22] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM 15*, pages 963–971. SIAM, 2015.

[23] L. Zhao, J. Chen, F. Chen, W. Wang, C.-T. Lu, and N. Ramakrishnan. Simnest: Social media nested epidemic simulation via online semi-supervised deep learning. In *ICDM 2015*, pages 639–648. IEEE, 2015.

[24] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *KDD 2015*, pages 1503–1512. ACM, 2015.