The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Distant-Supervision of Heterogeneous Multitask Learning for Social Event Forecasting with Multilingual Indicators

Liang Zhao
George Mason University
lzhao9@gmu.edu

Junxiang Wang
George Mason University
jwang40@gmu.edu

Xiaojie Guo
George Mason University
xguo7@gmu.edu

Abstract

Open-source indicators such as social media can be very effective precursors for forecasting future societal events. As events are often preceded by social indicators generated by groups of people speaking many different languages, multiple languages need to be considered to ensure comprehensive event forecasting. However, this leads to several technical challenges for traditional models: 1) high dimension, sparsity, and redundancy of features; 2) translation correlation among the multilingual features. and 3) lack of language-wise supervision. In order to simultaneously address these issues, we present a novel model capable of distant-supervision of heterogeneous multitask learning (DHML) for multilingual spatial social event forecasting. This model maps the multilingual heterogeneous features into several latent semantic spaces and then enforces a similar sparsity pattern across them all, using distant supervision across all the languages involved. Optimizing this model creates a difficult problem that is nonconvex and nonsmooth that can then be decomposed into simpler subproblems using the Alternative Direction Multiplier of Methods (ADMM). A novel dynamic programming-based algorithm is proposed to solve one challenging subproblem efficiently. Theoretical properties of the proposed algorithm are analyzed. The results of extensive experiments on multiple real-world datasets are presented to demonstrate the effectiveness, efficiency, and interpretability of the proposed approach.

Introduction

Significant social events like disease outbreaks and violent riots have a tremendous influence on the society. For instance, seasonal influenza alone is estimated to result in around 4 million cases of severe illness and hundreds of thousands of deaths each year (WHO). In the Middle East, the majority of the instabilities arise from extremism, terrorism, and civil unrest, usually resulting in major social issues with economic losses in the billions of dollars and creating millions of unemployed people. The motivation for this research is thus to mitigate these impacts by forecasting their occurrence in advance, which is now becoming feasible due to the rapid growth of web 2.0. For example, social media like Twitter have become popular platforms that serve as real-time “sensors" for social trends and incidents (Ramakrishnan et al. 2014). Every day millions of Twitter users around the globe broadcast their observations and sentiments on a wide range of topics in many different languages. Although around 50% of Twitter messages are in English, many other languages are commonly used for tweets (Hong, Convertino, and Chi 2011). Even within the same country, Twitter messages may be in multiple languages. For instance, in Brazil 73% of Twitter messages are in Portuguese, 15% are in English, and 10% are in Spanish, with the remaining 2% being in other languages. Tweets in different languages typically represent the voices of different subgroups of the population, which cannot be ignored. The collection of these observations and sentiments could provide a useful window into emerging social trends.

A considerable amount of work has been proposed for detecting historical or ongoing social events using social media (Zhang, Yuan, and Han). In addition, in order to foresee future events, a number of models for event forecasting have also been developed (Zhao et al. 2016b). Current research on social event forecasting via social media tends to share essentially similar workflows that basically learn a mapping from the observed semantic content (typically keyword frequencies) in various languages to the occurrence of future events using supervised learning techniques such as sparse learning (Zhao et al. 2015b), deep learning (Zhao et al. 2015a), or causality learning (Radinsky and Horvitz 2013). Multilingual issues in social media have generally been considered primarily for retrospective research like sentiment analysis (Bügel and Zielinski 2013; Lo et al. 2016; Becker, Moreira, and dos Santos 2017) and topic analysis (Lo, Chiong, and Cornforth 2017; Ren et al. 2016). However, due to the more challenging problem of predictive modeling, existing considerations of multilingual issues for social event forecasting remain few and problematic. For event forecasting, some works (Ramakrishnan et al. 2014) consider only the dominant language, thus losing substantial signals from the tweets in other languages. Other works simply merge all the keyword features from different languages into a single feature set for the model (Zhao et al. 2015b). These approaches suffer from several challenges. 1) High dimension, sparsity, and semantic-redundancy of the features. The number of features linearly increases with the number of languages, while most of the features are zeros because a message is typically in a single language. Moreover, the large feature set is highly redundant in semantic meaning due to the many words with same meaning across different languages. 2) Ignorance of feature translation correlation and heterogeneity. Existing work typically assumes the features in different languages are independent. However, this ignores the important semantic similarity and variety among them. For example, “protesto" in Portuguese and “protest" in English have the same meaning and significance for the topic “civil unrest", but “protesto" also has the meaning of “kick" in English. 3) Insufficiency of language-wise supervision. Different events are primarily preceded by the similar social indicators in different languages (Hong, Convertino, and Chi 2011), depending on the population bases of the people who organize and are involved in the events. Typically, it is extremely difficult to manually annotate the population base for each event accurately, because this information is implicit. Hence a lack of language-wise annotation limits the distinction among the predictive tasks across languages and challenges traditional models based on a single language.

In order to simultaneously address all these technical challenges, this paper presents a novel model named distant-supervision heterogeneous multitask learning (DHML). To effectively model the multilingual features with high dimension, sparsity, and redundancy, we treat each language as a task and can thus formulate a novel heterogeneous multitask learning problem. Unlike classic multitask learning such as that presented in (Zhao et al. 2015b; Zhou et al. 2013), where the feature sets are the same across tasks, the features (i.e., keywords) for different tasks in our problem must be different but correlated through translation relation. In order to share the correlated information among tasks, our model first maps the heterogeneous features in different tasks (i.e., languages) into their respective latent topic spaces, and then enforces a similar pattern of sparsity across all these latent spaces using a regularization term. More importantly, unlike classic multitask learning models where each task has its own label set, in our problem all the tasks share a single label set, rendering distant supervision (Hernández-González, Inza, and Lozano 2016) feasible. We adopt a multi-instance (Cheplygina, Tax, and Loog 2015) style framework to map the outputs of all the tasks into the event occurrence, in the sense that if any task can indicate the future event, then the output is positive; if they cannot, the output is negative. To optimize the proposed model, which is nonconvex and nonsmooth, we apply the Alternative Direction Methods of Multipliers (ADMM) (Boyd et al. 2011) to decompose it into simpler subproblems. To solve one quadratic subproblem embedded by a max function, we propose a new efficient dynamic programming method. To solve another biconvex subproblem, we adopt a non-monotone spectral projected gradient (Lu and Zhang 2012). Finally, we also prove a generalization bound that demonstrates both the consistency and the asymptotic behavior of the learning process of our proposed model. The major contributions of this research are as follows:

- **Develop a framework for event forecasting based on multilingual indicators**. A generic framework is proposed for spatial event forecasting that utilizes multilingual social media data with distant supervision. A number of classic approaches are shown to be special cases of our model.
- **Propose a new model for distant-supervised heterogeneous multitask learning**. Different languages are formulated into different tasks with heterogeneous features, which are then mapped to respective latent semantic spaces. Then we enforce a similar sparsity pattern across tasks in these latent spaces, using a distant supervision of all the tasks.
- **Design an efficient optimization algorithm based on ADMM and dynamic programming**. The proposed nonsmooth and nonconvex model is optimized by being decomposed into several simpler subproblems using ADMM. One subproblem coupled with a max function is solved by an efficient dynamic programming method, while another biconvex subproblem is solved by non-monotone spectral projected gradient. The theoretical properties of the proposed algorithm are also analyzed.
- **Conduct extensive experiments on several datasets**. The proposed method was evaluated on several real datasets, which demonstrates its effectiveness, efficiency, and interpretability.

Problem Formulation

Denote $X = \{X_{s,t,l}\}_{s,t,l}^{S,T,L}$ as a collection of model input data, such as Twitter messages, where T, S, and L are the sets of time intervals, spatial locations, and languages of messages, respectively. $X_{s,t,l} \in \mathbb{R}^{1\times|V_l|+1}$ denotes the feature vector, which are the keyword frequencies under language l for time t (e.g., tth date) at location s. $|V_l|$ denotes the size of feature set (e.g., vocabulary) V_l in the language l. An element $[X_{s,t,l}]_v \in X_{s,t,l},\ (v = 1, \cdots, |V_l|)$ denotes the frequency of the keyword $v \in V_l$ while $[X_{s,t,l}]_0 = 1$ is the dummy feature to provide a compact notation for bias parameter in forecasting model. Define $Y = \{Y_{s,t}\}_{s,t}^{S,T}$ as the event occurrences, where $Y_{s,t} \in \{0, 1\}$ such that $Y_{s,t} = 1$ means there is/are event(s) in location s at time t, otherwise $Y_{s,t} = 0$.

The following is our problem definition for the spatial social event forecasting using multilingual input data: For a specific location $s \in S$ at time $t \in T$, given data under different languages L, the goal is to predict the occurrence $Y_{s,\tau}$ of future event(s) where $\tau = t + p$. $p > 0$ is the lead time for forecasting. Thus, the problem is formulated as the following mapping function:

$$F : \{X_{s,t,1}, \cdots, X_{s,t,|L|}\} \rightarrow Y_{s,\tau} \qquad (1)$$

This problem is challenging in the following three aspects: 1) Input feature space has high dimension and sparsity. The size of feature set $\sum_l^L |V_l|$ linearly increases with the number of languages. Moreover, for each message, location, or time duration, the features of most of the languages are all-zeros, leading to severe sparsity. 2) Input feature space is heterogeneous and correlated. Each language has different vocabulary feature set with different size, namely $\forall\, l \neq m \rightarrow V_l \neq V_m,\ l, m \in L$. However, words in different languages could have strong semantic translation relation, leading to large information redundancy. 3) Distant supervision on high-dimension input. As can be seen in Equation (1), for each high-dimensional multilingual input, there is no language-wise supervision. Only distant supervision $Y_{s,\tau}$ upon all the languages L is available.

Models

To address the above issues, we propose a new model DHML. First, we propose to map the multilingual symbols into latent semantics and enforce the similarity among them. Second, we develop a distant-supervision multi-task learning framework.

Heterogeneous feature learning in multilingual latent spaces

In order to jointly minimize the prediction error and characterize the latent patterns in high-dimensional features, we need to optimize the penalized empirical loss: $\min \mathcal{L} + \Omega$, where $\mathcal{L}$ is the forecasting error on training set, and Ω is the regularization term that encodes structured prior knowledge embedded in the model parameters.

Due to the challenges of high dimension, sparsity, and redundancy of multilingual features, it is preferable to split the whole multilingual feature set into respective event forecasting tasks under different languages. Then each task is to forecast the events through a specific language l: $F_l : X_{s,t,l} \rightarrow y_{s,\tau,l}$, where $y_{s,\tau,l}$ is the predicted future event occurrence based on the input of language l. Moreover, the heterogeneous features of different languages correlate via translation relation. This motivates us to map the heterogeneous features into the latent semantic spaces, which have the similar patterns across different languages. The above idea can be formulated into the following objective function:

$$\min_{\Theta \geq 0, \Theta_l \Theta_l' = I, U} \frac{1}{|S||T|} \sum\nolimits_{s,t}^{S,T} \mathcal{L}(\{F_l(U_l' \Theta_l X_{s,t,l}')\}_l^L) + \lambda_1 \|U\|_{2,1} + \lambda_2 \sum\nolimits_l^L \|\Theta_l\|_1 \quad (2)$$

Throughout this paper, we use the apostrophe to denote the matrix transpose. For example, U_l' denotes the transpose of the matrix U_l. $\Theta_l \in \mathbb{R}^{d \times (|V_l|+1)}$ is the mapping from the original feature space to latent semantic space for each language l. d is the number of latent topics which can be specified by the user. Thus the multiplication $\Theta_l X_{s,t,l}'$ denotes the semantic pattern for current location s at time t for language l. $U \in \mathbb{R}^{d \times |L|}$ and $U_l \in \mathbb{R}^{d \times 1}$ denotes the weight vector for semantic feature for language l. The $\ell_{2,1}$ norm enforces different tasks for different languages to share a similar latent semantic pattern. We also enforce the sparsity of the transition matrix Θ by ℓ_1 norm. For classification problem, the loss $\mathcal{L}$ can be logistic loss, and F_l can be logit function.

Distant-supervised multi-task learning

In classic multitask learning (Zhao et al. 2015b), each task has the corresponding label set for its own output. However, in our problem, all the tasks do not have their respective labels $y_{s,\tau,l}$, but share a single label set of the final output $Y_{s,\tau}$, which neutralizes the classic multitask framework. To address this challenge, in this paper, the logical relation among the task outputs $\{F_l(X_{s,t,l})\}_l^L$ is explored and exploited. Specifically, in our problem, if any of the tasks' outputs $F_l(X_{s,\tau,l})$ is positive (e.g., indicates the occurrence of future event), then the final output $F(\{X_{s,\tau,l}\}_l^L)$ is also positive. Otherwise, the final output is negative. Because the model response $Y_{s,\tau} \in \{0, 1\}$ is binary-valued, this notion can be equivalently formulated via a "max" operation over the outputs of all the tasks:

$$\mathcal{L}(\{U_l', \Theta_l X_{s,t,l}'\}_l^L) = \mathcal{L}(\max_l F_l(U_l' \Theta_l X_{s,t,l}'), Y_{s,\tau}) \quad (3)$$

Therefore, combine Equations (2) and (7), we get the following final objective function:

$$\min_{U_l, \Theta_l \geq 0, l \in L} \frac{1}{|S||T|} \sum\nolimits_{s,t}^{S,T} \mathcal{L}(\max_l F_l(U_l' \Theta_l X_{s,t,l}'), Y_{s,\tau}) + \lambda_1 \|U\|_{2,1} + \lambda_2 \sum\nolimits_l^L \|\Theta_l\|_1 \qquad s.t. \Theta_l \Theta_l' = I, \ l \in L \quad (4)$$

The problem in Equation (4) is a nonconvex problem consisting of nonconvex and nonsmooth objective function and nonlinear equality constraint. It is very challenging to solve such a problem. In the Section "Parameter Optimization", we propose a new algorithm that can solve it efficiently and effectively.

Relation to existing methods

Several well-recognized methods are demonstrated to be the special cases of our model in Equation (4), including LASSO (Tibshirani 1996), multi-instance learning (Cheplygina, Tax, and Loog 2015), and homogeneous latent sparse feature learning (Zhou et al. 2013). The detailed deduction of their relations are presented the following.

1. **Generalization of LASSO**. Let $|L| = 1$ and Θ be an identity matrix. Also let the loss $\mathcal{L}$ be squared loss. Our DHMF is thus reduced to LASSO (Tibshirani 1996):

$$\min_u \sum\nolimits_{s,t}^{S,T} \mathcal{L}(F(u' X_{s,t,l}), Y_{s,\tau}) + \lambda_1 \|u\|_1 \quad (5)$$

where $u \in \mathbb{R}^{d \times 1}$ is the feature vector with the size of d.

2. **Generalization of homogeneous latent sparse feature learning**. Let $|L| = 1$, which removes the "max" function and $\ell_{2,1}$-norm. Also let the loss $\mathcal{L}$ be squared loss. Our DHMF is thus reduced to the homogeneous latent sparse feature learning model (Zhou et al. 2013):

$$\min_{u, \theta \geq 0} \sum\nolimits_{s,t}^{S,T} \mathcal{L}(F(u' \theta X_{s,t,l}), Y_{s,\tau}) + \lambda_1 \|u\|_1 + \lambda_2 \|\theta\|_1 \qquad s.t. \theta \theta' = I \quad (6)$$

where $\theta \in \mathbb{R}^{d \times v}$ is a matrix where v is the number of latent topics.

3. **Generalization of multi-instance learning**. Let $\lambda_1 = \lambda_2 = 0$ and $\Theta = I$, which means there is identical set of features for each instance. DHMF is thus reduced to multi-instance learning with standard assumption (Cheplygina, Tax, and Loog 2015):

$$\min_u \sum\nolimits_{s,t}^{S,T} \mathcal{L}(\max_l F(u' X_{s,t,l}), Y_{s,\tau}) \quad (7)$$

where $u \in \mathbb{R}^{d \times 1}$ is the feature vector with the size of d.

Parameter Optimization

In this section, we first propose an efficient algorithm to solve the nonconvex nonsmooth problem in Equation (4), and then discuss the theoretical properties of the proposed method.

Efficient Algorithm

It is extremely hard to directly solve the Equation (4) due to the existence of nonconvex max function inside a nonlinear loss function. An effective way to address this is to transform the original problem into the following equivalent problem by introducing auxiliary variables:

$$\begin{aligned} \min_{\Theta\geq 0,U,Z,Q} \frac{1}{|S||T|} &\sum\nolimits_{s,t}^{S,T} \mathcal{L}(f(Z_{s,t}), Y_{s,\tau}) \\ &+ \lambda_1 \|U\|_{2,1} + \lambda_2 \sum\nolimits_{l}^{L} \|\Theta_l\|_1 \\ s.t. \Theta_l \Theta'_l = I,\ & Z_{s,t} = \max_i Q_{s,t,i}, \qquad (8) \\ Q_{s,t,l} = U'_l \Theta_l X'_{s,t,l},\ & s \in S, t \in T, l \in L \end{aligned}$$

Therefore, by decoupling the max function from the nonlinear loss function $\mathcal{L}(\cdot)$ and the biconvex term $U'_l \Theta_l X'_{s,t,l}$, the original complex problem is transformed into a simpler one using two new variables Z and Q. This problem can be solved by ADMM, which requires to compute the *augmented Lagrangian* (Boyd et al. 2011) that uses additional quadratic penalty terms with the penalty parameter ρ:

$$\begin{aligned} &\frac{1}{|S|\cdot|T|} \sum\nolimits_{s,t}^{S,T} \Big(\mathcal{L}(f(Z_{s,t}), Y_{s,\tau}) + \sum\nolimits_{l}^{L} \langle \Lambda_{1,l}, \Theta_l \Theta'_l - I \rangle + \\ &+ \lambda_1 \|U\|_{2,1} + \frac{\rho}{2} \sum\nolimits_{l}^{L} \|\Theta_l \Theta'_l - I\|_F^2 + \Big\langle \Lambda_{2,s,t}, Z_{s,t} - \max_i Q_{s,t,i} \Big\rangle \\ &+ \frac{\rho}{2} \|Z_{s,t} - \max_i Q_{s,t,i}\|_F^2 + \sum\nolimits_{l}^{L} \Big\langle \Lambda_{3,s,t,l}, Q_{s,t,l} - U'_l \Theta_l X'_{s,t,l} \Big\rangle \\ &+ \lambda_2 \sum\nolimits_{l}^{L} \|\Theta_l\|_1 + \frac{\rho}{2} \sum\nolimits_{l}^{L} \|Q_{s,t,l} - U'_l \Theta_l X'_{s,t,l}\|_F^2 \Big) \qquad (9) \end{aligned}$$

where $\Lambda_1 = \{\Lambda_{1,l}\}_l^L$, $\Lambda_2 = \{\Lambda_{2,s,t}\}_{s,t}^{S,T}$, and $\Lambda_3 = \{\Lambda_{3,s,t,l}\}_{s,t,l}^{S,T,L}$ are the dual variables corresponding to the three constraints, respectively. Then, all the parameters Θ, U, Z, and Q are optimized alternately until convergence. Because Θ and U is coupled together in the constraint to form a biconvex problem, we seek to optimize them jointly in order to achieve a joint better solution for both variables.

1. Update $Q_{s,t,l},\ l \in L$.

Updating of the variable $Q_{s,t,l}$ amounts to the following optimization problem:

$$\begin{aligned} &\min_{Q_{s,t,l}} \Big\langle \Lambda_{2,s,t}, Z_{s,t} - \max_l Q_{s,t,l} \Big\rangle + \frac{\rho}{2} \|Z_{s,t} - \max_l Q_{s,t,l}\|_F^2 + \\ &\sum_{l}^{L} \Big\langle \Lambda_{3,s,t,l}, Q_{s,t,l} - U'_l \Theta_l X'_{s,t,l} \Big\rangle + \frac{\rho}{2} \sum_{l}^{L} \|Q_{s,t,l} - U'_l \Theta_l X'_{s,t,l}\|_F^2 \end{aligned}$$

It can be simplified as

$$\begin{aligned} \min_{Q_{s,t,l}} &\|Z_{s,t} - \max_l Q_{s,t,l} + \Lambda_{2,s,t}/\rho\|_F^2 \\ &+ \sum\nolimits_{l}^{L} \|Q_{s,t,l} - U'_l \Theta_l X'_{s,t,l} + \Lambda_{3,s,t}/\rho\|_F^2 \qquad (10) \end{aligned}$$

This problem is extremely challenging to solve due to the existence of a discrete operator, namely the “max" function, inside the quadratic penalty term $\|Z_{s,t} - \max_l Q_{s,t,l} + \Lambda_{2,s,t}/\rho\|_F^2$. This prevents the utilization of the traditional solution like Chebyshev approximation (Boyd and Vandenberghe 2004) for “max" function. Methods like subgradient methods (Renegar 2016) and selector variables (Andrews, Tsochantaridis, and Hofmann 2003) cannot guarantee a global solution, and are very time consuming. To solve this problem, we propose a dynamic programming based algorithm that can get the closed-form solution that has the global optimal guarantee for this problem very efficiently, as described in Theorem 1.

Theorem 1. *The solution to the problem in Equation* (10) *is:* $Q_{s,t,l} = U'_l \Theta_l X'_{s,t,l} - \Lambda_{3,s,t,l}/\rho$, *when* $l \neq \arg\min_i Q_{s,t,i}$; *Otherwise when* $l = \arg\min_i Q_{s,t,i}$, $Q_{s,t,l} = \sum_{i=1}^{k} (\tilde{Q}_{s,t,i} + U'_l \Theta_l X'_{s,t,l} + \Lambda_{2,s,t}/\rho)/(k+1)$, *where* $\tilde{Q}_{s,t}$ *is the decreasing ordered list whose elements are the set* $\{U'_l \Theta_l X'_{s,t,l} - \Lambda_{3,s,t,l}/\rho\}_l^L$, *and* k *is equal to the solution of the following problem:*

$$k = \arg\min_j\ j, \qquad (11)$$

$$s.t.,\ \sum\nolimits_{i=0}^{j} (\tilde{Q}_{s,t,i} + U'_l \Theta_l X'_{s,t,l} + \Lambda_{2,s,t}/\rho)/(j+1) > \tilde{Q}_{s,t,j-1}$$

Proof: The proof, which is very technical, is provided in the supplementary materials (Supplementary Materials). □

2. Jointly update parameters Θ and U.

Jointly optimizing Θ and U amounts to the following nonconvex subproblem:

$$\begin{aligned} &\min_{\Theta\geq 0,U} \lambda_1 \|U\|_{2,1} + \sum\nolimits_{l}^{L} \langle \Lambda_{1,l}, \Theta_l \Theta'_l - I \rangle + \frac{\rho}{2} \sum\nolimits_{l}^{L} \|\Theta_l \Theta'_l - I\|_F^2 + \\ &\lambda_2 \sum\nolimits_{l}^{L} \|\Theta_l\|_1 + \frac{\rho}{2|S|\cdot|T|} \sum\nolimits_{s,t,l}^{S,T,L} \|U'_l \Theta_l X'_{s,t,l} - (Q_{s,t,l} - \Lambda_{3,s,t,l}/\rho)\|_F^2 \end{aligned}$$

which contains a biconvex nonsmooth objective function of Θ and U as well as a quadratic equality constraint over Θ. To solve it, traditional methods like block coordinate descent (BCD) (Tseng and Yun 2009) may be easily trapped in a local minimizer in practice due to non-convexity and non-smoothness. To address this problem, we applied non-monotone strategy based on spectral projected gradient (SPG) method (Zhou et al. 2013). It is shown in (Lu and Zhang 2012) that under some suitable assumption the non-monotone SPG method has a linear convergence rate. The detailed algorithm procedures are shown in the supplementary materials (Supplementary Materials).

3. Update $Z_{s,t}$.

The optimization of Z is equal to solving the following subproblem:

$$\min_{Z_{s,t}} \mathcal{L}(f(Z_{s,t}), Y_{s,\tau}) + \frac{\rho}{2} \|Z_{s,t} - \max_l Q_{s,t,l} + \Lambda_{2,s,t}/\rho\|_F^2$$

which is a generalized linear regression with a quadratic penalty term. 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 $1/(4 \cdot I)$, where I denotes the identity matrix.

Finally, the dual variables are updated following the standard way in ADMM. The process of ADMM will terminate until convergence or specific number of iteration steps is reached.

Theoretical Properties

In this section, we first provide an equivalent reformulation of the proposed model, and then show that its generalization error is bounded and asymptotically converges to 0 when the size of training sample increases to infinity.

The original problem with $\ell_{2,1}$ and ℓ_1 norms is equivalently transformed to the following:

$$\min_{\Theta\geq 0,U}\frac{1}{|S|\cdot|T|}\sum\nolimits_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau}) \quad (12)$$
$$s.t.\|U\|_{2,1}\leq\alpha,\ \gamma_{\min}\leq\sum\nolimits_i^L\|\Theta_i\|_1\leq\gamma_{\max}^{\lambda_1},\ \Theta_l\Theta_l'=I,\ l\in L$$

where $\mathcal{L}$ is Lipschitz continuous with Lipschitz constant C. Denote $\Theta_{l,m}$ as the mth row of the matrix Θ_l, we have the following lemma.

Lemma 1. *$\gamma_{\min}$ is the lower bound of $\|\Theta_l\|_1$ such that $\gamma_{\min}=d\cdot\arg\min_{\|\Theta_{l,m}\|_2=1}\|\Theta_{l,m}\|_1$, where $\Theta_{l,m}$ is any row of Θ_l.*

Proof: The proof detail is in the supplementary material (Supplementary Materials). □

In the following, we show that the generalization error of the problem in Equation (12) is upper-bounded. First, we define the following concepts.

Definition 1 (Expected error, Empirical error)**.** *For any Θ and U, we denote the expected risk as: $\mathbb{E}(\Theta,U)=\mathbb{E}_{M\sim\mu}[\frac{1}{|S|\cdot|T|}\sum_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau})]$. Given the data $M=(X,Y)$, the empirical risk is defined as: $\hat{\mathbb{E}}(\Theta,U|M)=\frac{1}{|S|\cdot|T|}\sum_{s,t}^{S,T}\mathcal{L}(F_l(\max_l U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau})$.*

Definition 2 (global optimal solution, optimized solution)**.** *Define $\mathcal{F}_1=\{u\in\mathbb{R}^{d\times|L|}|\|u\|_2\leq\alpha\}$, $\mathcal{F}_2=\{u\in\mathbb{R}^{|L|\times d\times(|V_l|+1)}|\gamma_{\min}\leq\sum_i^L\|u_i\|_1\leq\gamma_{\max}^{\lambda_1},\ u_lu_l'=I,u_l\geq 0,\ l\in L\}$. Denote Θ^* and U^* as the global optimal solution of the expected risk:*

$$(\Theta^*,U^*)=\arg\min_{\Theta\in\mathcal{F}_2,U\in\mathcal{F}_1}\mathbb{E}(\Theta,U) \quad (13)$$
$$=\arg\min_{\Theta\in\mathcal{F}_2,U\in\mathcal{F}_1}\mathbb{E}_{M\sim\mu}[\frac{1}{|S||T|}\sum\nolimits_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau})]$$

Denote $\Theta^*_{(M)}$ and $U^*_{(M)}$ as the optimized solution by minimizing the empirical risk:

$$(\Theta^*_{(M)},U^*_{(M)})=\arg\min_{\Theta\in\mathcal{F}_2,U\in\mathcal{F}_1}\hat{\mathbb{E}}(\Theta,U|M) \quad (14)$$
$$=\arg\min_{\Theta\in\mathcal{F}_2,U\in\mathcal{F}_1}\frac{1}{|S||T|}\sum\nolimits_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau})$$

Finally, the following theorem shows the upper-bounded generalization error.

Theorem 2. *Define $\mathcal{C}_1(X)=d\cdot\|\hat{\Sigma}(X)\|_*$ and $\mathcal{C}_\infty=\sum_l^L\|\hat{\Sigma}(X_{\cdot,\cdot,l})\|_\infty/|L|$, where $\hat{\Sigma}(x)=x'x$ is the empirical co-variance matrix. Let $\epsilon>0$ and let μ be probability measure on $\mathbb{R}$. With probability of at least $1-\epsilon$ in the draw of $M\sim\mu^{|S|\cdot|T|}$. We have:*

$$\mathbb{E}(\Theta^*_{(M)},U^*_{(M)})-\mathbb{E}(\Theta^*,U^*)$$
$$=\mathbb{E}_{M\sim\mu}[\frac{1}{|S||T|}\sum\nolimits_{s,t}^{S,T}\mathcal{L}(\max_l F_l([U^*_{(M)}]_l'[\Theta^*_{(M)}]_l X'_{s,t,l}),Y_{s,\tau})]$$
$$-\inf_{\Theta\in\mathcal{F}_2,U\in\mathcal{F}_1}\mathbb{E}_{M\sim\mu}[\frac{1}{|S||T|}\sum\nolimits_{s,t}^{S,T}\mathcal{L}(\max_l F_l(U_l'\Theta_l X'_{s,t,l}),Y_{s,\tau})]$$
$$\leq 2C\alpha\sqrt{\frac{2\mathcal{C}_1(X)|L|(d+12)}{|S|\cdot|T|}}+2C|L|\alpha\sqrt{\frac{8\mathcal{C}_\infty(X)\ln(2d)}{|S|\cdot|T|}}+2\sqrt{\frac{2\ln 2/\epsilon}{|S||T|}}$$

Proof: The general proof procedure is included in supplementary material (Supplementary Materials) due to the limited space. □

The theorem provides important insights into the proposed model: 1) The more training samples utilized, the less the generalization error; 2) The generalization error converges to 0 when the training sample size approaches infinity; 3) More languages tend to require more training samples to achieve the same generalization error; and 4) If the design matrix X has a low-rank structure that leads to smaller $\mathcal{C}_1(X)$, then the generalization error converges faster.

Experiments

In this section, the experiment results for the proposed DHML are presented and discussed.

Experimental settings

In our experiment, 10 datasets were obtained by randomly sampling 10% (by volume) of the Twitter data from Jan 2013 to Dec 2014, as shown in Table 1. The data from Jan 1, 2013 to Dec 31, 2013 was utilized for training, while the remaining was used for the performance evaluation. The Latin American Twitter data are majorly in Spanish, English, and Portuguese, of which the relative percentages of tweet message amounts are shown in Table 1. The event forecasting results were validated against a labeled events set, known as the gold standard report (GSR), which is publicly available[1]. GSR is a collection of civil unrest news reports manually labeled by social science experts from the most influential newspaper outlets in Latin America (Ramakrishnan et al. 2014), 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").

This experiment adopted an extensive list of civil unrest related keywords manually defined by social science experts (Ramakrishnan et al. 2014), including 688 Spanish keywords, 569 English keywords, and 549 Portuguese keywords. The data collection was partitioned into a sequence of date-interval subcollections. The event forecasting task here was to utilize one-day tweet data to predict whether or not there will be an event in the next day for a specific city. To

Table 1: Datasets and Labels. (SPA=Spanish, ENG=English, POR=Portuguese)

Dataset	**#Tweets**	SPA (%)	ENG (%)	POR (%)	**#Events**
Argentina	160,564,890	91.6	7.3	1.1	1427
Brazil	185,286,958	10.1	16.0	73.9	3417
Chile	97,781,414	82.8	16.4	0.8	776
Colombia	158,332,002	89.8	9.4	0.8	1287
Ecuador	50,289,195	91.1	8.1	0.8	511
El Salvador	21,992,962	91.5	7.8	0.7	730
Mexico	197,550,208	83.7	15.4	0.9	5907
Paraguay	30,891,602	92.2	6.4	1.4	2114
Uruguay	10,310,514	89.7	8.8	1.4	664
Venezuela	167,411,358	92.3	6.9	0.8	3320

[1] GSR dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EN8FUW

Table 2: Event forecasting performance in AUC in each of the 10 datasets

	AR	BR	CL	CO	EC	EL	MX	PY	UY	VE
LogReg	0.594	0.686	0.677	0.644	0.599	0.618	0.661	0.6162	0.628	0.667
LASSO	0.596	0.685	0.677	0.648	0.603	0.636	0.665	0.6151	0.666	0.669
MTL	**0.733**	0.722	0.669	0.810	0.617	0.772	**0.795**	0.600	0.811	0.771
MREF	0.706	0.714	0.563	0.515	0.784	0.612	0.693	0.658	0.6812	0.588
DHML	0.704	**0.845**	**0.683**	**0.846**	**0.839**	**0.780**	0.793	**0.737**	**0.835**	**0.835**

Table 3: Comparison of runtimes (in seconds) in model training on each of the 10 datasets

	AR	BR	CL	CO	EC	EL	MX	PY	UY	VE
LogReg	12,784	30,193	2,981	8,060	312	551	17712	7,297	748	5,563
LASSO	687	1,535	242	780	295	261	2,043	527	336	1,008
MTL	**76**	**233**	**35**	108	**17**	**17**	853	**40**	**20**	**49**
MREF	3,567	25,889	6,521	14,714	4,332	4,669	31,349	9,495	5,305	5,769
DHML	202	332	852	**87**	46	33	**175**	242	82	179

achieve this, we created a training set and a test set for each city, where each data sample was the daily tweet observation with the above keyword features. The predicted events were structured as tuples of (date, city). A predicted event was matched to a real event if both the date and location attributes were matched. To validate the performance of event forecasting, the Area Under the Curve (AUC) of Receiver operating characteristic (ROC) (Zhao et al. 2016b) curve was adopted. There are three tunable parameters for our proposed model DHML, namely the regularization parameters λ_1, λ_2, and number of topics d. For each dataset, a 10-fold cross validation was utilized on training set to examine large ranges of parameter values, namely 10^{-6} to 1 for λ_1, 10^{-4} to 5 for λ_2, and 5 to 70 for d. The values with best performance on training set were selected.

Comparison Methods

The effectiveness and efficiency were compared with 4 state-of-the-art methods on spatial event forecasting, namely Logistic Regression (LogReg) (Wang, Gerber, and Brown 2012), LASSO (Ramakrishnan et al. 2014), Multi-task learning (MTL) (Zhao et al. 2015b), and Multi-resolution event forecasting (MREF) (Zhao et al. 2016a).

1. *Logistic regression (LogReg)* (Gerber 2014). LogReg utilizes a logit function to map the tweets observations into future event occurrences (“0" denotes no occurrence, “1" denotes occurrence). The input features here are union of all the keywords in all the languages. The keyword set is the same as the one being used by the proposed method.

2. *LASSO* (Ramakrishnan et al. 2014). A LASSO models is built to map the multilingual inputs to the model response, which is the future event occurrence. The feature set is the same as LogReg. To tune the regularization parameter, different values from the set $\mathcal{R}_p = \{0.01, 0.02, \cdots, 0.1, 0.2, \cdots, 1, 2 \cdots, 10\}$ were tested based on a 10-fold cross validation on the training set. To be specific, for the training set of each dataset, we partitioned the training set into 10 equal segments along the time line. We then used 9 segments for training the model and the remaining segment for validating the results, giving a total of 10 rounds by iterating the segment used for the validation. For each round a validation performance is obtained and our focus is on the average performance across all 10 rounds. The regularization parameter was set at 0.1 because this value achieved the best average performance for all 10 rounds.

3. *Multitask Learning (MTL)* (Zhao et al. 2015b). In multitask model, each task is the forecasting for each location. The feature set is the same as LASSO. As in LASSO, the values in $\mathcal{R}_p$ were tested to select the regularization parameters. Finally, the parameters were set as $\lambda_1 = 0.015$ and $\lambda = 0.001$, which were selected because they achieved the best overall performance in the 10-fold cross-validation.

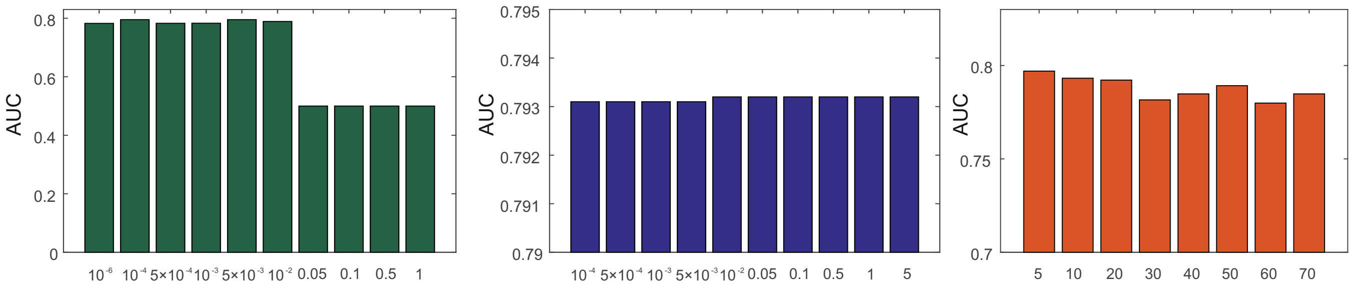
4. *Multi-resolution Event Forecasting (MREF)* (Zhao et al. 2016a). This method jointly models the prediction tasks in multiple geographical levels by utilizing their geo-hierarchical relation. The features are the same as MTL. The major parameter is the regularization parameter that is determined from the set $\mathcal{R}_p = \{0.01, 0.02, \cdots, 0.1, 0.2, \cdots, 1, 2 \cdots, 10\}$ for each dataset by 10-fold cross-validation.

Performance

In this section, the proposed DHML was evaluated quantitatively and qualitatively. The analyses of parameter sensitivity and feature selection are also presented.

As shown Table 2, the proposed DHML outperformed the competing methods in 8 out of all the 10 datasets by 15% on average, and was also competitive in the remaining 2 datasets. It exhibited outstanding performance in countries like Brazil where the multilingual phenomenon was especially obvious as shown in Table 1. In addition, MTL that considers the correlation among locations performed the second best in general. MREF is designed for event forecasting in different spatial levels, but was not the best when predicting for only city level. The performance of LogReg and LASSO was similar and less competitive.

Table 3 shows the running time of model training for each model in each dataset. It can be seen that logistic regression, LASSO, and MREF consumed the largest amounts of time, about thousands of seconds for large countries and hundreds of seconds for small ones. This is because each of Logistic



(a) sensitivity of parameter λ_1 (b) sensitivity of parameter λ_2 (c) sensitivity of number of topics d

Figure 1: Sensitivity analysis of three parameters of the model DHML

Table 4: Most important topics and keywords in each language on the Uruguay dataset. The keywords in Spanish and Portuguese have been translated into English by Google Translate (https://translate.google.com/). The table shows that similar topics were important across different languages, e.g., Topic 1, Topic 5, and Topic 8 appear in all 3 languages. And each topic tended to have coherent semantic meaning in different languages, e.g., Topic 1 is potentially about the school protests, and Topic 4 is generally about preventing the protests.

Languages	Topics	Keywords							
Spanish	Topic 8	conflict	farmer	rancher	pit	insecurity	agrarian	whistle	deforest
	Topic 5	picket	mobilize	arrest	cooperative	impoverish	Zapatista	#cnte	upsurge
	Topic 1	class	criminalize	suppress	riot	moderate	barricade	protest	teacher
	Topic 9	#cgtp	#cofecay	#snte	@eloisago.	@chertor.	@dionisio.	@morenaj.	@unt_mx
	Topic 10	power	match	Energy	warn	food	town	defending	torture
English	Topic 7	agency	community	maintain	charge	reform	discuss	loss	legalize
	Topic 1	smash	effort	proposal	invitation	arm	university	class	fight
	Topic 5	medium	report	popular	paralyze	tax	affect	danger	payment
	Topic 4	gringo	crime	investment	attack	capture	victim	protagonist	boycott
	Topic 6	mandate	striker	confront	assembly	parliament	mandatory	freedom	parade
Portuguese	Topic 2	person	president	time	class	opportunity	deputy	election	alternative
	Topic 4	ambush	plunder	warn	police	gun	convention	agreement	officer
	Topic 6	lead	national	together	change	authority	congress	labor	violence
	Topic 5	pocket	mine	shot	catch	criminal	control	enemy	upsurge
	Topic 8	hunger	hassle	fire	treatment	defeat	Medical	groom	root

Regression and LASSO utilized all the data of different locations, time, and languages to train a single model, resulting in a design matrix which was extremely huge and sparse. MREF need consider event forecasting tasks in multiple resolution with several constraints, which increased its algorithm complexity and time complexity. MTL and our DHML achieved the lowest runtimes, which were generally less than 1/10 of other methods. This was because they split the forecasting tasks into different locations and different languages, which reduced the data size and sparsity. DHML was fast also because its subproblems were solved by proximal operators and closed-form solutions.

The sensitivity of parameters of the proposed DHML on test set of the Mexico dataset is illustrated in Figure 1. Other datasets followed similar patterns. The regularization parameter λ_1 influenced the performance most while the setting of parameter λ_2 had minimal impact. When λ_1 was set to be less than 0.05 the model obtained the best performance. The number of topics is better to be set less than 20, but there was not much decrease in performance if it is larger than 30.

In addition to event forecasting, the proposed DHML can also discover the underlying semantic topics and key precursors for future events. The significance of the latent topics and keywords was ranked by their weights based on the optimized variables Θ and U. Table 4 presents the top 8 keywords of top 5 topics for each language on Uruguay Twitter dataset as an example. It shows that similar topics were important across different languages, including Topic 1, Topic 5, and Topic 8. And each topic tended to have coherent semantic meaning in different languages. This is because our model employed an $\ell_{2,1}$ to enforce a similar sparsity of semantic patterns across different languages. For example, Topic 8 in Spanish and Portuguese contained keywords like “farmer" and “hunger", which was about agricultural factors that potentially preceded future civil unrest events. Topic 1 in Spanish and English, which highlighted “riot", “fight", and “university", was about the actions of protest potentially relevant to schools.

Conclusions

In order to handle the spatial social event forecasting based on multilingual indicators, this paper proposes a novel model for distant-supervised heterogeneous multitask learning (DHML). Its generalization performance is theoretically guaranteed. Several effective models are demonstrated to be special cases of the proposed DHML. For the parameter optimization of DHML, an efficient algorithm for nonconvex and nonsmooth is proposed based on ADMM and a new dynamic programming method. Extensive experiment on real datasets demonstrated the effectiveness, efficiency, and interpretability of DHML.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2003. Support vector machines for multiple-instance learning. *Advances in neural information processing systems* 577–584.

Becker, K.; Moreira, V. P.; and dos Santos, A. G. 2017. Multilingual emotion classification using supervised learning: Comparative experiments. *Information Processing & Management* 53(3):684–704.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

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

Bügel, U., and Zielinski, A. 2013. Multilingual analysis of twitter news in support of mass emergency events. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 5(1):77–85.

Cheplygina, V.; Tax, D. M.; and Loog, M. 2015. On classification with bags, groups and sets. *Pattern Recognition Letters* 59:11–17.

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

Hernández-González, J.; Inza, I.; and Lozano, J. A. 2016. Weak supervision and other non-standard classification problems: a taxonomy. *Pattern Recognition Letters* 69:49–55.

Hong, L.; Convertino, G.; and Chi, E. H. 2011. Language matters in twitter: A large scale study. In *ICWSM 2011*.

Lo, S. L.; Cambria, E.; Chiong, R.; and Cornforth, D. 2016. A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. *Knowledge-Based Systems* 105:236–247.

Lo, S. L.; Chiong, R.; and Cornforth, D. 2017. An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications* 81:282–298.

Lu, Z., and Zhang, Y. 2012. An augmented lagrangian approach for sparse principal component analysis. *Mathematical Programming* 1–45.

Radinsky, K., and Horvitz, E. 2013. Mining the web to predict future events. In *WSDM 2013*, 255–264.

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

Ren, Z.; Inel, O.; Aroyo, L.; and de Rijke, M. 2016. Time-aware multi-viewpoint summarization of multilingual social text streams. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 387–396. ACM.

Renegar, J. 2016. "efficient" subgradient methods for general convex optimization. *SIAM Journal on Optimization* 26(4):2649–2676.

Supplementary Materials. https://github.com/zhaoliangvaio/homepage/blob/master/materials/aaai_supplementary.pdf. accessed May 2017.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tseng, P., and Yun, S. 2009. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming* 117(1):387–423.

Wang, X.; Gerber, M. S.; and Brown, D. E. 2012. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer. 231–238.

WHO. Influenza (season) fact sheet. Accessed May 15, 2015. http://www.who.int/mediacentre/factsheets/fs211/en/.

Zhang, C.; Yuan, Q.; and Han, J. Bringing semantics to spatiotemporal data mining: Challenges, methods, and applications. In *33st IEEE Intl. Conf. on Data Engineering (ICDE 2017),(Tutorial)*.

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

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

Zhao, L.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2016a. Multi-resolution spatial event forecasting in social media. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, 689–698. IEEE.

Zhao, L.; Ye, J.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2016b. 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. ACM.

Zhou, J.; Lu, Z.; Sun, J.; Yuan, L.; Wang, F.; and Ye, J. 2013. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1034–1042. ACM.