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

Incomplete Label Multi-Task Ordinal Regression for Spatial Event Scale Forecasting

Yuyang Gao
George Mason University
Fairfax, VA 22030

Liang Zhao
George Mason University
Fairfax, VA 22030

Abstract

Event scales are commonly used by practitioners to gauge subjective feelings on the magnitude and significance of social events. For example, the Centers for Disease Control and Prevention (CDC) utilizes a 10-level scale to distinguish the severity of flu outbreaks and governments typically categorize violent outbreaks based on their intensity as reflected in multiple aspects. Effective forecasting of future event scales can be used qualitatively to determine reasonable resource allocations and facilitate accurate proactive actions by practitioners. Existing spatial event forecasting methods typically focus on the occurrence of events rather than their ordinal event scales as this is very challenging in several respects, including 1) the ordinal nature of the event scale, 2) the spatial heterogeneity of event scaling in different geo-locations, 3) the incompleteness of scale label data for some spatial locations, and 4) the spatial correlation of event scale patterns. In order to address all these challenges concurrently, a MultI-Task Ordinal Regression (MITOR) framework is proposed to effectively forecast the scale of future events. Our model enforces similar feature sparsity patterns for different tasks while preserving the heterogeneity in their scale patterns. In addition, based on the first law of geography, we proposed to enforce spatially-closed tasks to share similar scale patterns with theoretical guarantees. Optimizing the proposed model amounts to a new non-convex and non-smooth problem with an isotonicity constraint, which is then solved by our new algorithm based on ADMM and dynamic programming. Extensive experiments on ten real-world datasets demonstrate the effectiveness and efficiency of the proposed model.

Introduction

Societal events that are spatially based, such as disease outbreaks and organized crime, have a significant impact on society. The ability to successfully forecast future spatial events of this nature would thus be extremely beneficial for decision makers seeking to avoid, control, or alleviate the associated social upheaval and risks. Spatial social event forecasting is a fast-growing research area that typically forecasts the *occurrence* of future spatial events, namely whether or not the spatial events will happen. However, in many applications simply forecasting the *occurrence* of an event is not enough. Knowledge regarding the *scale* or *level* of a future event is vital if decision makers are to achieve optimal resource allocation. For example, as shown in Figure 1, the Centers for Disease Control and Prevention (CDC) rank the severity of ongoing disease outbreaks using five scale points. The successful prediction of the scale of future disease outbreaks enables practitioners to allocate appropriate levels of resources for vaccination and isolation. For organized crime event forecasting, the significance of the future crime events is a valuable reference for government agencies seeking to set realistic alert levels and allocate sufficient resources for local and national law enforcement. However, as yet little research has focused specifically on spatial social event scale forecasting.

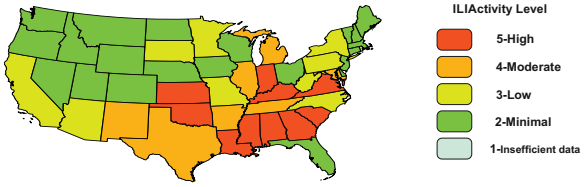

Figure 1: ILI data for one week of the 2016-17 influenza season (CDC)

Traditional event forecasting methods typically predict a binary output (i.e., the event either occurs or it doesn't), and hence cannot be directly applied to forecast event scales, which are ordinal variables. Spatial event scale forecasting is a new problem that poses several unique and interesting challenges that are yet to be addressed. **1) Spatial heterogeneity of event scales**: Different locations have different characteristics, such as population, weather and administrative structures. These factors often lead to population discrepancies in the social media users at different locations. For example, for influenza outbreak forecasting, the same number of mentions of the keyword 'flu' in tweets can lead to quite different actual scales in California and Nebraska; since California has a far larger population than Nebraska, the same number of mentions of 'flu' in tweets will signify a far lower influenza activity level in California than in Nebraska. **2) Incomplete labels in spatial event scales**: For any given location, there will often be some missing scale levels within the training data as some event scales were

not represented within the period of time chosen as training data. For example, for influenza outbreak forecasting, in Nebraska there were no level 3 or level 5 events during the year of 2011, which means that any model trained based on this data will lose the power to forecast these scales in the future.
3) Spatial correlation of event scale patterns: Spatial locations are not independent of one another, but instead are correlated following spatial topology and this therefore needs to be considered. According to the well-recognized "first law of geography" (Cressie 2015), the event scale pattern should be more similar in nearby locations than those faraway. For example, in influenza epidemic outbreaks, nearby states typically have similar activity level, as shown in Figure 1.

In this paper, we propose a new MultI-Task Ordinal Regression (MITOR) framework for spatial event scale forecasting that concurrently addresses all the above challenges. The main contributions of our study are summarized as follows:

1. **Developing the MITOR framework for event scale forecasting.** We formulate event scale forecasting for multiple locations as a multitask ordinal regression problem. We enforces similar feature sparsity patterns for different locations while preserving heterogeneity in their scale patterns.
2. **Proposing a model that enforces structured scale patterns.** Based on the first law of geography, we propose to enforce similar event scale patterns among spatially-closer tasks. This is achieved by a newly proposed regularization term that is proved to be equivalent to the divergence of the scale distribution patterns among nearby locations.
3. **Developing an efficient algorithm for solving a new non-convex problem.** To solve the proposed model's objective function that is non-convex and non-smooth problem with an isotonicity constraint, we propose a new algorithm based on the Alternating Direction Method of Multipliers (ADMM) and dynamic programming that is capable of solving the proposed model efficiently and is guaranteed to converge to a local optimal solution.
4. **Conducting comprehensive experiments to validate the effectiveness and efficiency.** Extensive experiments on 10 datasets from civil unrest and influenza outbreaks domains demonstrate that the proposed models outperform other comparison methods. In addition, sensitivity analysis and qualitative analysis are provided to demonstrate the effectiveness of our regularization term.

Related Work

Spitial event forecasting. Most research that has been reported focuses on temporal events and ignores the underlying geographical information related to the forecasting of elections (O'Connor et al. 2010), stock market movements (Argyriou, Evgeniou, and Pontil 2007), disease outbreaks (Achrekar et al. 2011), and box office ticket sales (Arias, Arratia, and Xuriguera 2013). There are a few existing approaches that provide true spatiotemporal resolution for predicted events. For example, (Gerber 2014) utilized a logistic

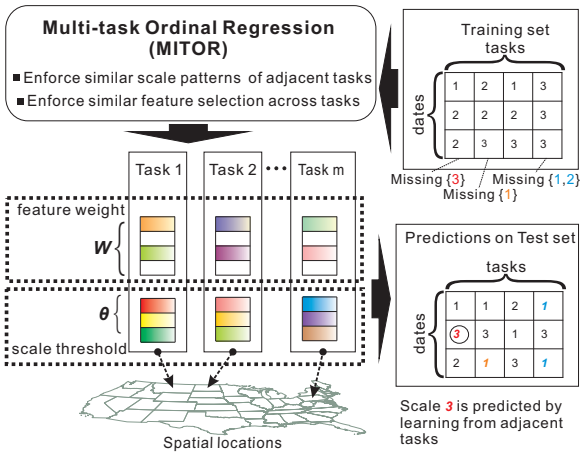

Figure 2: Flowchart of the proposed MITOR model

regression model for spatiotemporal events forecasting using topic-related tweet volumes as features. (Ramakrishnan et al. 2014) built separate LASSO models for different locations to predict the occurrence of civil unrest events. (Zhao et al. 2015a) designed a new predictive model based on a topic model that jointly characterizes the temporal evolution in terms of both the semantics and geographical burstiness. However, all of them focus on the occurrence only, but not able to handle the scales of future events.

Multi-task learning. Multi-task learning (MTL) refers to models that learn multiple related tasks simultaneously to improve their generalization performance (Arias, Arratia, and Xuriguera 2013; Thrun and OSullivan 1998). Many MTL approaches have been proposed (Tutz 2003). For example (Evgeniou and Pontil 2004) proposed a regularized MTL framework that constrained all task models to be close to each other. The task relatedness can also be modeled by constraining multiple tasks to share a common underlying structure, e.g., a common set of features (Argyriou, Evgeniou, and Pontil 2007), or a common subspace (Ando and Zhang 2005). (Zhao et al. 2015b) demonstrated the utility of applying a Multi-Task Learning framework for spatiotemporal event forecasting.

Ordinal regression. Ordinal regression is a point-wise approach to classifying data where the labels exhibit a natural order. Threshold-based methods assume that the unobserved continuous variables underlie the ordinal response (Gutiérrez et al. 2016; Verwaeren, Waegeman, and De Baets 2012). As a member of cumulative link models (CLMs) (Agresti and Kateri 2011), the proportional odds model (POM) is specifically designed for threshold-based ordinal regression (McCullagh 1980). Non-proportional alternatives like generalized ordered logit model simply assume a different weight for each class (Williams and others 2006). Another alternative applies the proportional odds assumption only to a subset of variables, known as the partial proportional odds model (Peterson and Harrell Jr 1990). (Tutz 2003) presented a general framework that extends generalized additive models to incorporate non-parametric parts.

Problem Setup and Preliminary Setups

Problem Setup

Suppose there are $\mathcal{S}$ spatial locations (e.g., cities, states) in a country of interest. Given a time interval t (e.g., hour, day), the spatio-temporal social media data for location $s \in \mathcal{S}$ and time t can be formulated as $X_{s,t} \in \mathbb{R}^{n\times 1}$, which denotes a feature vector whose i-th element is term frequency - inverse document frequency (TF-IDF).

The event scale at location s and time t is defined as an ordinal response $Y_{s,t} \in \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\}$, where $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k$ are ordinal class labels and k is the total number of ordinal event scales. A natural label ordering is included, denoted as $\mathcal{C}_1 \prec \mathcal{C}_2 \prec ... \prec \mathcal{C}_k$, where $\prec$ is the ascending order relation.

Given the tweet data $X_{s,t}$ in a specific location s and a time interval t, the goal is to predict the scale of a future event denoted by $Y_{l,\tau}$ for the same location s and a future time interval τ, where $\tau = t + p$ and $p > 0$ is the lead time. In this paper, we set the time intervals t as per day and the lead time p is set to one day ahead. Formally, this problem is equivalent to learning a mapping from social media data to future event scale predictions $f : X_{s,t} \rightarrow Y_{s,\tau}$.

Preliminaries

Because the response variable $Y_{s,t}$ is an ordinal variable, the assumption of an order between event scales makes it inappropriate to apply conventional methods such as multi-class classification and regression directly. Specifically, conventional regression models like linear regression require continuous values and thus cannot handle the categorical variable $Y_{s,t}$ in our problem. Classification models, although they focus on categorical variables, only address nominal variables and ignore the ordinal information in our problem.

To predict the ordinal variable $Y_{s,t}$, ordinal regression models such as the proportional odds model (POM) (McCullagh 1980) are commonly used to effectively leverage and address the ordinal nature of the problem. Compared to multi-class logistic regression, POM adds the constraint that the hyper-planes that separate different classes are parallel for all classes, that is, the *weight co-efficient* vector w is common across classes. The model also assumes that a latent variable underlies the ordinal response, which will be estimated by *threshold* vector θ in the model in order to separate different class labels.

In the logistic ordinal regression we model the cumulative probability as the logistic function. Thus, we can formulate the objective function of our problem as a negative log-likelihood:

$$\underset{W,\Theta}{\operatorname{argmin}} -\sum_{s,t=1}^{\mathcal{S},\mathcal{T}} \log(\sigma(w^T X_{s,t} + \theta_{Y_{s,t}}) - \sigma(w^T X_{s,t} + \theta_{Y_{s,t}-1}))$$
$$\text{s.t.} \theta_1 \leq \theta_2 \leq \theta_3 \leq ... \leq \theta_{k-1} \quad (1)$$

Where $w \in \mathbb{R}^{n\times 1}$, and $\theta \in \mathbb{R}^{(k-1)\times 1}$ are the two parameter sets to be estimated in the model, with $\theta_0 = -\infty$ and $\theta_k = \infty$ to represent extremal classes, $X_{s,t}$ denotes t-th sample of the s-th location, $Y_{s,t}$ denotes its corresponding scale. The function $\sigma(x)$ is the logistic sigmoid function denoted as $\sigma(x) = 1/(1 + e^{-x})$. Notice that our problem and proposed models are generic and can also accommodate other ordinal regression models. In this paper, we focus on POM.

The model proposed in Equation (1) suffers from two challenges: 1) all the locations share a single *weight coefficient* vector w and *threshold* vector θ, therefore cannot handle any spatial heterogeneity in the event scale for different locations; and 2) Equation (1) assumes all the locations are independent even though some spatial correlations exist among locations regarding the event scale pattern, as shown in Figure 1. In order to jointly handle these challenges, in the next section, we present our MITOR framework.

Models

In this section, we propose a new model, MITOR-I, that enforces similar feature sparsity patterns for different tasks while preserving heterogeneity in their scale patterns. We then move on to propose MITOR-II, which introduces a novel regularization term that utilizing geo-information to restrict adjacent locations by sharing their event scale labels, building on MITOR-I.

MITOR-I

To handle the spatial heterogeneity of event scale criteria for different locations, we need to build an exclusive model for each individual locations, all of which have their own *thresholds*. Although these *thresholds* are different, different locations share similar feature *weight coefficients* patterns because people generally share a common language and speak in a similar way, so the keywords for a topic of interest will be similar across different locations, for example, "influenza" and "cough",would both refer to the topic 'flu'. Therefore, we propose to leverage multitask learning in ordinal regression to enforce different tasks that share a similar *weight coefficients* pattern but reserve their own *thresholds*. We define a feature weight coefficient matrix $W \in \mathbb{R}^{n\times\mathcal{S}}$ and a threshold matrix $\Theta \in \mathbb{R}^{\mathcal{S}\times k}$, where each column of W, denoted as $W_{\cdot,s}$, is the feature weight coefficient vector for task s, while each row of Θ, denoted as $\Theta_{s,\cdot}$, is the threshold vector for task s. Learning multiple related tasks simultaneously effectively increases the sample size for each task, since when we learn a model for a specific task, we use information from all other tasks.

The similar pattern of W across different tasks is achieved by enforcing a similar sparsity pattern among tasks. Specifically, we can add $\ell_{2,1}$ norm regularization over the W matrix, which sums the ℓ_2 norms for each feature, and each ℓ_2 norm is enforced for all the tasks for each feature. Thus, the i-th feature, which corresponds to the i-th element in each model, is likely to be selected or not by all models simultaneously.

Mathematically, we propose the first MultI-Task Ordinal Regression model (MITOR-I) as follows:

$$\operatorname{argmin}_{W,\Theta} \mathcal{L}(W, \Theta) + \alpha||W||_{2,1}$$
$$\text{s.t. } \Theta_{s,1} \leq \Theta_{s,2} \leq \Theta_{s,3} \leq ... \leq \Theta_{s,k-1},\ s \in \{1, 2, ..., \mathcal{S}\}$$

Where we define $\mathcal{L}(W, \Theta)$ as follows for simplicity and for later use:

$$-\sum_{s,t=1}^{\mathcal{S},\mathcal{T}} \log\left(\sigma\left(W_{\cdot,s}^T X_{s,t} + \Theta_{s,Y_{s,t}}\right) - \sigma\left(W_{\cdot,s}^T X_{s,t} + \Theta_{s,Y_{s,t}-1}\right)\right)$$

$||W||_{2,1}$ is the group sparsity term for matrix W which encourages all tasks to select a common set of features; it can be computed as the sum of ℓ_2-norm for each row in W. The regularization hyper-parameter α controls the sparsity.

MITOR-II

In a multi-task learning setting for ordinal regression, each task has only a limited number of samples and thus not every task has a complete set of labels in the training set. For example, in Figure 2 the top right box contains an example of a set of training data labels (event scales). Only task 3 has a complete set of event scales: all the other tasks are missing one or more labels. The threshold associated with the missing labels cannot be learned during the training phase and hence the model is not capable of predicting the missing labels. Note that this issue becomes more severe as the number of class labels increases. For example, in the U.S. influenza outbreak dataset, among all the CDC data from 2011 to 2014, for each year only around 50% of the states include all the scale points.

Different tasks each have their own missing labels, so we propose to allow correlated tasks to adaptively complement each other's missing labels. This means that we need to determine the correlation among tasks. Based on the first law of geography "everything is related to everything else, but near things are more related than distant things"(Cressie 2015), we know nearby locations will tend to be more similar to each other. For example, in a disease outbreak, nearby states typically have similar levels of flu activity, as shown in Figure 1. For a time interval t, given two locations i and j that are close in geo-spatial distance, the probability of that the event scale at location i being equal or under event scale $\mathcal{C}_q$, which is denoted as $P(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})$, is similar to that of location j, leading to the following equation:

$$\frac{P(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})}{P(Y_{i,t} \succ \mathcal{C}_q|X_{i,t})} \approx \frac{P(Y_{j,t} \preceq \mathcal{C}_q|X_{j,t})}{P(Y_{j,t} \succ \mathcal{C}_q|X_{j,t})}$$

The *odds* of being equal or under scale $\mathcal{C}_q$ is defined as the fraction of the probability of being equal or under scale $\mathcal{C}_q$ over the probability of being above scale $\mathcal{C}_q$. Mathematically, the *odds* can be expressed as:

$$odds(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})) = \frac{P(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})}{P(Y_{i,t} \succ \mathcal{C}_q|X_{i,t})} \quad (2)$$

Therefore, the ratio of the *odds* of being equal or under two adjacent scales q and $q+1$ of two tasks close in geo-spatial distance should also be similar. Mathematically, this can be expressed as:

$$\frac{odds(Y_{i,t} \preceq \mathcal{C}_{q+1}|X_{i,t})}{odds(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t}))} \approx \frac{odds(Y_{j,t} \preceq \mathcal{C}_{q+1}|X_{j,t})}{odds(Y_{j,t} \preceq \mathcal{C}_q|X_{j,t}))} \quad (3)$$

The similarity pattern in Equation (2) can thus be equivalently denoted by thresholds, as shown in Lemma 1.

Lemma 1. *Equation* (3) *is theoretically equivalent to the following:*

$$\Theta_{i,\mathcal{C}_{q+1}} - \Theta_{i,\mathcal{C}_q} \approx \Theta_{j,\mathcal{C}_{q+1}} - \Theta_{j,\mathcal{C}_q} \quad (4)$$

where i and j are two tasks that are close in geo-spatial distance and $\mathcal{C}_q$ and $\mathcal{C}_{q+1}$ are two adjacent event scales.

Proof. We can derive the lemma from the following equations:

$$ln\left(\frac{P(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})}{P(Y_{i,t} \succ \mathcal{C}_q|X_{i,t})}\right) = W_{\cdot,i}^T X_{i,t} + \Theta_{i,\mathcal{C}_q} \quad (5)$$

Equation (5) is the definition of POM. From this, we can derive an equivalent expression with $\mathcal{C}_{q+1}$ and subtract one from the other to omit the input vector $X_{i,t}$ on the right, as shown in the following equation:

$$ln\left(\frac{P(Y_{i,t} \preceq \mathcal{C}_{q+1}|X_{i,t})}{P(Y_{i,t} \succ \mathcal{C}_{q+1}|X_{i,t})}\right) - ln\left(\frac{P(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t})}{P(Y_{i,t} \succ \mathcal{C}_q|X_{i,t})}\right) = \Theta_{i,\mathcal{C}_{q+1}} - \Theta_{i,\mathcal{C}_q}$$

Combining above equation with Equation (2), where the term *odds* is defined, we obtain the ratio of odds with Θ as:

$$\frac{odds(Y_{i,t} \preceq \mathcal{C}_{q+1}|X_{i,t})}{odds(Y_{i,t} \preceq \mathcal{C}_q|X_{i,t}))} = e^{\Theta_{i,\mathcal{C}_{q+1}} - \Theta_{i,\mathcal{C}_q}} \quad (6)$$

Thus, combining Equation (3) and Equation (6), we reach the conclusion that given two tasks i and j that are close in geo-spatial distance, the difference between threshold $\Theta_{i,\mathcal{C}_{q+1}} - \Theta_{i,\mathcal{C}_q}$ and $\Theta_{j,\mathcal{C}_{q+1}} - \Theta_{j,\mathcal{C}_q}$ should be similar, as shown in Equation (4). □

Therefore, we propose a new model to encourage the difference between threshold parameter $\Theta_{i,\mathcal{C}_{q+1}} - \Theta_{i,\mathcal{C}_q}$ to be similar among adjacent tasks. This is done by introducing a new regularization term for Θ which makes use of the spatial information of the tasks, given by the adjacent matrix of tasks. Mathematically, the new model is as follows:

$$\begin{aligned} &\text{argmin}_{W,\Theta}\, \mathcal{L}(W,\Theta) + \alpha\|W\|_{2,1} + \\ &\frac{\beta}{2}\sum_{i=1}^{\mathcal{S}}\sum_{j=2}^{k-1} \| \left(\Theta_{i,j} - \Theta_{i,j-1}\right) - \frac{1}{N_i}\sum_{z \in adj(i)} \left(\Theta_{z,j} - \Theta_{z,j-1}\right) \|_2^2 \\ &\text{s.t. } \Theta_{i,1} \leq \Theta_{i,2} \leq \Theta_{i,3} \leq ... \leq \Theta_{i,k-1},\ i \in \{1,2,...,\mathcal{S}\} \end{aligned} \quad (7)$$

Where the function $adj(i)$ returns the set of tasks that is adjacent to task i and N_i is the total number of its neighbors. This term will encourage adjacent tasks to have a similar ratio for the *odds* between two consecutive scales by encouraging the difference between threshold parameter $\Theta_{i+1} - \Theta_i$ to be similar among adjacent tasks. The regularization hyper-parameter β controls the importance of this term.

Algorithm

In this section, we propose a new algorithm to optimize the parameters of MITOR models in Equation (7). Since MITOR-I is a special case of MITOR-II when $\beta = 0$, we focus on MITOR-II here.

The problem in Equation (7) is nonconvex and nonsmooth. Moreover, the parameter Θ has isotonicity constraints (namely $\Theta_{i,1} \leq \Theta_{i,2} \leq \Theta_{i,3} \leq ... \leq \Theta_{i,k-1}$). These challenges neutralize the existing methods like subgradient and coordinate descent methods (Bishop 2006). In order to solve the challenges, we propose a new algorithm based on ADMM (Boyd et al. 2011) that first decomposes the original problem into several simpler subproblems and then solve them iteratively. To handle the isotonicity constraints, quadratic penalties of non-smooth functions have been introduced, which are solved by our newly proposed methods

based on dynamic programming that can ensure global optimal solution for this subproblem.

The pseudo-code of the proposed algorithm is summarized in **Algorithm 1**. The parameter set $\{W, \Theta, U, V, y^{(1)}, y^{(2)}, y^{(3)}\}$ is alternately solved by the proposed algorithm until convergence is achieved. Lines 3-7 show the alternating optimization of each of the variables. The detailed optimization for all the variables are described in more detail below.

Algorithm 1: The Proposed Algorithm

Require: $X, Y, \rho, \alpha, \beta, \lambda_W, \lambda_\Theta$
Ensure: solution W, Θ
1: initialize $W^0, \Theta^0, U^0, V^0, y^{(1)0}, y^{(2)0}, y^{(3)0}, i = 0$
2: **repeat**
3: $W^i, \Theta^i \Leftarrow$ Equation (9)
4: $U^i \Leftarrow$ Equation (10)
5: $V^i \Leftarrow$ calculation following Theorem 1
6: $y^{(1)i}, y^{(2)i}, y^{(3)i} \Leftarrow$ Equation (11)
7: $i \Leftarrow i + 1$
8: **until** convergence

Base on ADMM formulation, the original objective function of MITOR-II can be re-written as follows:

$$\mathcal{L}(W, \Theta) + \alpha||U||_{2,1} + \frac{\beta}{2}\sum_{i=1}^{\mathcal{S}}\sum_{j=2}^{k-1} \| (V_{i,j} - V_{i,j-1}) - \frac{1}{N_i}\sum_{z \in adj(i)} (V_{z,j} - V_{z,j-1}) \|_2^2$$
$$\text{s.t. } W = U, \Theta = V \quad (8)$$
$$V_{i,1} \leq V_{i,2} \leq V_{i,3} \leq ... \leq V_{i,k-1} \text{ for } i \in \{1, 2, ..., \mathcal{S}\}$$

Thus, the augmented Lagrangian is:

$$\underset{W,\Theta,U,V}{\text{argmin}}\ \mathcal{L}(W, \Theta) + \alpha||U||_{2,1} + trace(y^{(1)}(W - U)^T) + \rho/2||W - U||_2^2 + trace(y^{(2)}(\Theta - V)^T) + \rho/2||\Theta - V||_2^2 + \frac{\beta}{2}\sum_{i=1}^{\mathcal{S}}\sum_{j=2}^{k-1}\|(V_{i,j} - V_{i,j-1}) - \frac{1}{N_i}\sum_{z \in adj(i)} (V_{z,j} - V_{z,j-1})\|_2^2 + \sum_{i=2}^{k-1} y_{\cdot,i}^{(3)}(V_{\cdot,i-1} - V_{\cdot,i})^T + \rho/2\sum_{i=2}^{k-1}||max(V_{\cdot,i-1} - V_{\cdot,i}, 0)||_2^2$$

Notice that the max operator here acts as a vector max which will set the element of the vector to 0 when it is less than 0.

Update W, Θ

The sub-problem of updating W and Θ is as follows:

$$\text{argmin}_{W,\Theta}\ \mathcal{L}(W, \Theta) + trace(y^{(1)}(W - U)^T) + \rho/2||W - U||_2^2 + trace(y^{(2)}(\Theta - V)^T) + \rho/2||\Theta - V||_2^2 \quad (9)$$

Since $\mathcal{L}(W, \Theta)$ is a non-convex function with respect to W and Θ, we will use a traditional gradient descent algorithm, carefully choosing the step size λ_W and λ_Θ for W and Θ to jointly update them to local optima.

Update U

The sub-problem of updating U is as follows:

$$\text{argmin}_U\ \alpha||U||_{2,1} + trace(y^{(1)}(W - U)^T) + \rho/2||W - U||_2^2 \quad (10)$$

This can be solved by proximal gradient descent using the proximal operator on the $\ell_{2,1}$ norm (Bach et al. 2012).

Update V

The sub-problem of updating V is as follows:

$$\underset{V}{\text{argmin}}\ \frac{\beta}{2}\sum_{i=1}^{\mathcal{S}}\sum_{j=2}^{k-1}\| (V_{i,j} - V_{i,j-1}) - \frac{1}{N_i}\sum_{z \in adj(i)} (V_{z,j} - V_{z,j-1}) \|_2^2 + \rho/2||\Theta - V||_2^2 + trace(y^{(2)}(\Theta - V)^T) + \sum_{i=2}^{k-1} y_{\cdot,i}^{(3)}(V_{\cdot,i-1} - V_{\cdot,i})^T + \rho/2\sum_{i=2}^{k-1}||max(V_{\cdot,i-1} - V_{\cdot,i}, 0)||_2^2$$

The $adj()$ function introduces some difficulties for updating V, since every pair of consecutive class level thresholds for the same task show in the same term. In addition, the same class level threshold among all tasks will also lead to recursive relationships. This makes elemental-wise updating of V impossible in pratice.

In order to address this problem, we can treat the $adj()$ function as the matrix representation $R \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$, and reformulate the problem as matrix multiplication:

$$\underset{V}{\text{argmin}}\ \frac{\beta}{2}\sum_{i=1}^{\mathcal{S}}\sum_{j=2}^{k-1}\left|\left|(V_{\cdot,j} - V_{\cdot,j-1})^T R^T\right|\right|_2^2 + trace\left(y^{(2)}(\Theta - V)^T\right) + \frac{\rho}{2}||\Theta - V||_2^2 + \sum_{i=2}^{k-1} y_{\cdot,i}^{(3)}(V_{\cdot,i-1} - V_{\cdot,i})^T + \frac{\rho}{2}\sum_{i=2}^{k-1}||max(V_{\cdot,i-1} - V_{\cdot,i}, 0)||_2^2$$

Where $R_{i,i} = 1$ and $R_{i,adj(i)} = -\frac{1}{N_i}$, for $i = 1...\mathcal{S}$. N_i is the total number of neighbors of task i.

Theorem 1. *The optimal solution for matrix V can be obtained by computing its column vectors in order as follows:*

$$V_{\cdot,1} = y_{\cdot,1}^{(2)}/\rho + \Theta_{\cdot,1}$$
$$V_{\cdot,i} = \begin{cases} (\beta V_{\cdot,i-1}L + y_{\cdot,i}^{(2)} + \rho(\Theta_{\cdot,i} + V_{\cdot,i-1}) + y_{\cdot,i-1}^{(3)})(\beta L + 2\rho I)^{-1} & V_{\cdot,i} < V_{\cdot,i-1} \\ (\beta V_{\cdot,i-1}L + y_{\cdot,i}^{(2)} + \rho\Theta_{\cdot,i} + y_{\cdot,i-1}^{(3)})(\beta L + \rho I)^{-1} & V_{\cdot,i} \geq V_{\cdot,i-1} \end{cases}$$

Where $L = R^T R$ *and* $i = 2...k-1$.

Proof. Please see the supplemental material[1] . □

Update y

Finally, update $y^{(1)}, y^{(2)}, y^{(3)}$ as follows:

$$y^{(1)} = y^{(1)} + \rho(W - U),\ y^{(2)} = y^{(2)} + \rho(\Theta - V) \quad (11)$$
$$y_i^{(3)} = max(y_i^{(3)} + \rho(V_{i-1} - V_i), 0), \text{ for } i = 2...k-1$$

Experiments

In this section, the performance of the proposed new model, MITOR, is evaluated using 10 real datasets. First, the experimental setup is introduced. The effectiveness and efficiency of MITOR is then evaluated against several existing baseline methods. All the experiments were conducted on a 64-bit machine with Intel(R) core(TM) quad-core processor (i7CPU 2.5 GHz) and 16GB memory.

[1] https://github.com/zhaoliangvaio/homepage/blob/master/materials/aaai_supplementary_multi_ordinal.pdf

Dataset and Experiment Setup

In this study, 8 datasets from civil unrest forecasting and 2 datasets from influenza outbreak forecasting are used for the experimental evaluations.

Civil unrest: each dataset is from each of 8 different countries in Latin America, namely Argentina, Brazil, Chile, Colombia, Mexico, Paraguay, Uruguay, and Venezuela. For these datasets, data sources from Twitter are adopted as the model inputs. In each case the data for the period from July 1, 2013 to February 9, 2014 is used for training, while the data from February 10, 2014 to December 31, 2014, is used for the performance evaluation. The event forecasting results are validated against a labeled events set, known as the gold standard report (GSR) (GSR Dataset). GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin America (O'Connor et al. 2010). The event scale for the civil unrest dataset is the relative crowd size ('none' $\prec$ 'small' $\prec$ 'large'). An example of a labeled GSR event is given by the tuple: (CITY="Hermosillo", STATE ="Sonora", COUNTRY = "Mexico", DATE = "2013-01-20", EVENT SCALE="large").

Influenza outbreaks: The 2 datasets for influenza outbreaks in the U.S. use Twitter data as the data source. A total number of 1,266,301 tweets that contain the flu-related keywords like 'flu' and 'influenza' are included in the datasets. In the first dataset, data of year 2011 is used for training and data of year 2012 is used for performance evaluation. In the second dataset, data of year 2013 is used for training and data of year 2014 is used for performance evaluation. The forecasting results for the flu outbreaks are validated against the corresponding influenza statistics reported by the Centers for Disease Control and Prevention (CDC). CDC publishes the weekly influenza-like illness (ILI) activity level for each state in the U.S. based on the proportional level of outpatient visits to healthcare providers for ILI. The event scale is the relative ILI activity level ('insufficient data' $\prec$ 'minimal' $\prec$ 'low' $\prec$ 'moderate' $\prec$ 'high'). An example of a CDC flu outbreak event is: (STATE = "Virginia", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013", ACTIVITY LEVEL = "low").

Parameter Setting: The hyper-parameters in the proposed model have been chosen based on the performance for the validation set. The validation set consists of a randomly chosen 15% of the training data. Moreover, we have illustrated and discussed the parameter sensitivity in Section "Parameter Sensitivity Study".

Performance Evaluation: To evaluate the prediction performance for ordinal variables, Mean Zero-one Error (MZE) and Mean Absolute Error (MAE) are commonly used.

MZE is the error rate of the classifier: $MZE = \frac{1}{N}\sum_{i=1}^{N}[\![y_i^* \neq y_i]\!] = 1 - Acc$, where y_i is the true label, y_i^* is the predicted label and Acc is the accuracy of the classifier. MZE values range from 0 to 1; they are related to global performance, but do not consider the order.

MAE is the average deviation in absolute value of the predicted rank y_i^* from the true one y_i (Baccianella, Esuli, and Sebastiani 2009): $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i|$. MAE values range from 0 to $k-1$ (maximum deviation in number of scales).

Baselines for comparison: The performance of the proposed models is compared with the baseline as well as the state of the art methods, namely: *SVC1V1* (Support Vector Classifier with OneVsOne), *SVC1VA* (Support Vector Classifier with OneVsAll) (Hsu and Lin 2002), *SVMOP* (Support Vector Machines with OrderedPartitions) (Waegeman and Boullart 2009), and *POM* (Proportional Odds Model) (McCullagh and Nelder 1989). The detail introduction and hyper-parameter setting is included in supplemental material.

Performance

Tables 1, 2, 3 show the performance for all the methods on all the datasets based on both MZE and MAE. The runtime of training is shown on flu dataset. The runtime for civil unrest follows the similar trends and is not included due to space limitations. For the test times, all the methods consume negligible testing times (less than 1 sec).

These results indicate that the methods that utilize multi-task frameworks perform better than most baseline methods overall. Moreover, when the group sparcity $\ell_{2,1}$ constraint and adjacency location based threshold constraint are included, the performance of the MITOR-II model is superior.

Table 1 shows that MITOR-II consistently performs well across different countries, being the best in Argentina, Brazil, Colombia, Mexico, Uruguay, and Venezuela and competitive in Paraguay and Chile and outperforming the baseline models by 10% - 50% both in MZE and MAE. MITOR-I also achieves good scores, but is not as competitive as MITOR-II. This is largely because MITOR-II utilizes geo-information by including the proposed adjacency location based threshold constraint. Interestingly, MITOR-II largely outperformed the baselines by around 50% on the Argentina dataset, but only by 10% on the Mexico dataset. Examining the dataset, 16 of the 23 states have incomplete scale labels in the Argentina dataset, covering nearly 70% of the entire country, while only 17 out of 32 states have incomplete scale labels in the Mexico dataset, around 50% of the country. This may suggest that the threshold regularization term in MITOR-II improves the performance substantially when there is more serious incompleteness of labels.

Tables 2, 3 also demonstrate the effectiveness of the proposed methods. MITOR-II outperformed the baseline models consistently by 20% - 40% both in MZE and MAE. The tables also show the training times for all the methods. We can see that SVM models tend to have short training times, with SVM with OneVsOne binary decomposition scheme (SVC1V1) having the shortest training time. The training time for the proposed methods outperform the baseline POM model, but as MITOR is constructed based on POM, this still demonstrates the proposed optimization method is an efficient way to solve the proposed models.

The table that shows top 10 features selected by MITOR is included in supplemental material.

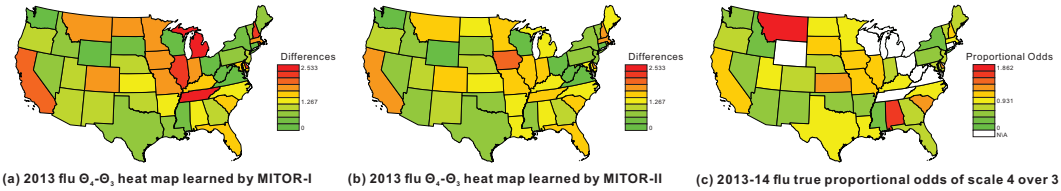

Figure 3: The heat map for the US flu dataset for $\Theta_{\cdot,4} - \Theta_{\cdot,3}$ and ground truth proportional odds of class 4 over class 3

Table 1: Event forecasting performance comparison on civil unrest datasets (MZE, MAE)

Method	Argentina	Brazil	Chile	Colombia	Mexico	Paraguay	Uruguay	Venezuela
SVC1VA	0.0368, 0.0708	0.0440, 0.0857	0.0657, 0.1129	0.0552, 0.0916	0.1284, 0,2284	0.0353, 0.0674	0.0223, 0.0390	0.0615, 0.1127
SVC1V1	0.0339, 0.0670	0.0441, 0.0860	0.0610, 0.1098	0.0506, 0.0884	0.1187, 0.2184	0.0339, 0.0610	0.0227, 0.0403	0.0690, 0.1201
SVMOP	0.0392, 0.0709	0.0482, 0.0854	0.0740, 0.1189	0.0542, 0.0889	0.1187, 0.2184	0.0337, 0.0608	0.0239, 0.0398	0.0690, 0.1201
POM	0.0287, 0.0572	0.0626, 0.1230	0.0524, 0.0989	0.0376, 0.0724	0.0982, 0.1906	0.0367, 0.0717	0.0340, 0.0667	0.0374, 0.0722
MITOR-I	0.0161, 0.0306	0.0344, 0.0665	**0.0436**, **0.0812**	0.0280, 0.0534	0.0967, 0.1875	**0.0284**, **0.0551**	0.0132, 0.0250	0.0289, 0.0551
MITOR-II	**0.0158**, **0.0305**	**0.0339**, **0.0657**	**0.0436**, **0.0812**	**0.0274**, **0.0521**	**0.0875**, **0.1690**	0.0286, 0.0555	**0.0122**, **0.0231**	**0.0286**, **0.0545**

Table 2: Experimental results for 2011-2012 U.S. flu dataset

Model	Training time	MZE	MAE
SVC1VA	144	0.2246	0.3167
SVC1V1	**68**	0.2220	0.3096
POM	1216	0.2250	0.3117
SVMOP	96	0.2269	0.3118
MITOR-I	394	0.1148	0.1900
MITOR-II	395	**0.1145**	**0.1895**

Table 3: Experimental results for 2013-2014 U.S. flu dataset

Model	Training time	MZE	MAE
SVC1VA	187	0.2861	0.4367
SVC1V1	**77**	0.2869	0.4368
POM	800	0.3036	0.4822
SVMOP	114	0.2921	0.4310
MITOR-I	425	0.1796	0.3473
MITOR-II	532	**0.1794**	**0.3466**

Parameter Sensitivity Study

There are two hyper-parameters in the proposed MITOR-II model, as shown in Equation (8), where α controls group sparsity $\ell_{2,1}$ norm and β controls the proposed regularization term on Θ. α is also introduced in MITOR-I model and follows similar trends as it performs in MITOR-II.

Figure 4 show the MZE and MAE of the model versus α and β respectively. Only the results for Mexico within civil unrest datasets and 2011-12 influenza outbreak dataset are shown due to space limitations. The top 2 bar charts in Figure 4 show the MZE and MAE of the model versus α. By varying α across the range from 0.0001 to 10, the performance of the influenza outbreak dataset is stable, with the fluctuation ranges less than 0.01. For the civil unrest dataset, the fluctuation range is 0.015 in MZE and 0.03 in MAE. The best performance is obtained when $\alpha = 0.5$. We can also see a clear trend where both MZE and MAE increase when

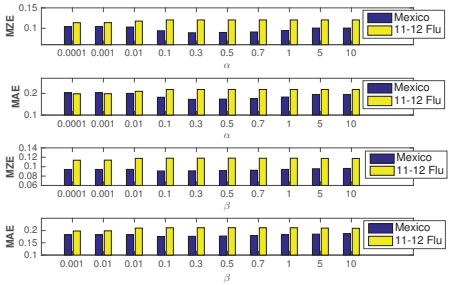

Figure 4: Sensitivity analysis for hyper-parameters

α is too large or too small. The bottom 2 bar charts illustrate the MZE and MAE of the model versus β, which is varied across the same range as α. The fluctuation ranges around 0.01 for both MZE and MAE. In general, the performance is good when β is small, but deteriorates once β becomes too large. This is because a large β will force the model to pay too much attention to being similar to its adjacent tasks and may thus lead to the loss of its own characteristic and a consequent decrease in overall performance.

The Effect of Scale Pattern Regularization

This section validates the effectiveness of the threshold regularization term on scale patterns in MITOR-II. On the flu dataset, Figure 3 compares the scale patterns in terms of Θ learned by MITOR-I (i.e., without threshold regularization) and MITOR-II (i.e., with threshold regularization). Each of Figure 3(a) and (b) shows the difference between 3rd and 4th thresholds $\Theta_{i,4} - \Theta_{i,3}$ for each ith task (state) in the U.S. Figure 3(c) shows the ground truth proportional odds of class 4 over class 3 for each of the states for two years, 2013 and 2014. Figure 3(a) shows a dramatic divergence among the values for different states while in Figure 3(b) the patterns among nearby states is spatially smoothed, which is

more similar to the patterns in ground truth shown in Figure 3(c). This is because MITOR-II can utilize threshold regularization to encourage the nearby states to share their knowledge with each other under the "first law of geography", which will largely alleviate each state's incompleteness of label set. For example, the pattern of the relatively small state "Colorado" suffered from data incompleteness and deviated from the neighbor states, but MITOR-II corrected this, as compared with the ground truth in Figure 3(c).

Conclusions

Effective forecasting of future event scales can be used to qualitatively inform reasonable re-source allocation and enable more accurate proactive actions by practitioners. To address this issue, this paper proposes a novel MultI-Task Ordinal Regression (MITOR) framework that characterizes the feature sparsity, task scale incompleteness, and scale pattern correlation. An efficient algorithm for parameter optimization is proposed to handle this non-convex and non-smooth problem with isotonicity constraints. Extensive experiments on 10 real-world datasets demonstrate that the proposed model outperforms other comparison methods in multiple application domains.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; and Liu, B. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 702–707. IEEE.

Agresti, A., and Kateri, M. 2011. Categorical data analysis. In *International encyclopedia of statistical science*. Springer. 206–208.

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6(Nov):1817–1853.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In *Advances in neural information processing systems*, 41–48.

Arias, M.; Arratia, A.; and Xuriguera, R. 2013. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(1):8.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2009. Evaluation measures for ordinal regression. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, 283–287. IEEE.

Bach, F.; Jenatton, R.; Mairal, J.; Obozinski, G.; et al. 2012. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning* 4(1):1–106.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

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.

CDC. https://www.cdc.gov/.

Cressie, N. 2015. *Statistics for spatial data*. John Wiley & Sons.

Evgeniou, T., and Pontil, M. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 109–117. ACM.

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

GSR Dataset. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EN8FUW. accessed Sep 2017.

Gutiérrez, P. A.; Pérez-Ortiz, M.; Sanchez-Monedero, J.; Fernández-Navarro, F.; and Hervas-Martinez, C. 2016. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering* 28(1):127–146.

Hsu, C.-W., and Lin, C.-J. 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks* 13(2):415–425.

McCullagh, P., and Nelder, J. A. 1989. Generalized linear models, no. 37 in monograph on statistics and applied probability.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)* 109–142.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11(122-129):1–2.

Peterson, B., and Harrell Jr, F. E. 1990. Partial proportional odds models for ordinal response variables. *Applied statistics* 205–217.

Ramakrishnan, N.; Butler, P.; Muthiah, S.; Self, N.; Khandpur, R.; Saraf, P.; Wang, W.; Cadena, J.; Vullikanti, A.; Korkmaz, G.; et al. 2014. 'beating the news' with embers: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1799–1808. ACM.

Thrun, S., and OSullivan, J. 1998. Clustering learning tasks and the selective cross-task transfer of knowledge. In *Learning to learn*. Springer. 235–257.

Tutz, G. 2003. Generalized semiparametrically structured ordinal models. *Biometrics* 59(2):263–273.

Verwaeren, J.; Waegeman, W.; and De Baets, B. 2012. Learning partial ordinal class memberships with kernel-based proportional odds models. *Computational Statistics & Data Analysis* 56(4):928–942.

Waegeman, W., and Boullart, L. 2009. An ensemble of weighted support vector machines for ordinal regression. *International Journal of Computer Systems Science and Engineering* 3(1):47–51.

Williams, R., et al. 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal* 6(1):58.

Zhao, L.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2015a. Spatiotemporal event forecasting in social media. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, 963–971. SIAM.

Zhao, L.; Sun, Q.; Ye, J.; Chen, F.; Lu, C.-T.; and Ramakrishnan, N. 2015b. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1503–1512. ACM.