Forecasting Gang Homicides with Multi-level Multi-task Learning

Nasrin Akhter[1(✉)], Liang Zhao[1], Desmond Arias[1], Huzefa Rangwala[1], and Naren Ramakrishnan[2]

[1] George Mason University, Fairfax, VA, USA
{nakhter3,lzhao9,earias2}@gmu.edu, rangwala@cs.gmu.edu
[2] Virginia Tech, Arlington, VA, USA
naren@cs.vt.edu

Abstract. Gang-related homicides account for a significant proportion of criminal activity across the world, especially in countries of Latin America. They often arise from territorial fights and, distinct from other types of homicides, are characterized by area-specific risk indicators. Current crime modeling and prediction research has largely ignored gang-related homicides owing to: (i) latent dependencies between gangs and spatial areas, (ii) area-specific crime patterns, and (iii) insufficiency of spatially fine-grained predictive signals. To address these challenges, we propose a novel context-aware multi-task multi-level learning framework to jointly learn area-specific crime prediction models and the potential operating territories of gangs. Specifically, to sufficiently learn the finer-grained area-specific tasks, the abundant knowledge from coarse-grained tasks is exploited through multi-task learning. Experimental results using online news articles from Bogotá, Colombia demonstrate the effectiveness of our proposed method.

Keywords: Multi-task learning · Gang homicide · Crime forecasting

1 Introduction

Homicidal violence is concentrated in the Americas [1], especially in Latin American and Caribbean countries [2]. Gang wars, involving narco businesses, are the key contributors to Latin America's homicidal violence problem. Similarly, 90% of gun violence can be attributed to gangs in the United States [3]. Existing homicide prediction research mostly ignores gang involvement and spatial heterogeneity of crime indicators within cities. Often a country or a city has crime pockets dominated by local gangs which are likely to influence the crime scene of neighboring areas. In this study, we aim to identify patterns of activities in multiple locations as indicators for future events. For instance, the arrest of a gang leader could incite aggression by rivals gangs in the neighborhoods, leading to homicides.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 28–37, 2018.
https://doi.org/10.1007/978-3-319-93372-6_3

Forecasting gang-related homicides from online news demands several challenges to be solved: (1) **Scarcity of fine-grained location information**. City-level news articles report local crimes as well as crimes with nationwide impact. Although there may be a reasonable amount of city-level data available from city-level newspapers, dividing them into even finer levels, i.e., suburbs within a city, often suffers from data scarcity. (2) **Heterogeneity of geographical locations**. Although nearby locations may be influenced by the same regional phenomena, each region has its own exclusive set of characteristics and principle actors affecting that particular region. Accounting for this location heterogeneity is crucial in predicting future area-specific homicidal violence. (3) **Multi-resolution feature structure**. The set of keywords in the homicide-reporting news articles often exhibit a subtle hierarchical structure. On top is the common homicide and violence related keywords, area-specific entity names and keywords lie at the bottom level. A model, trained on a global set of keywords, is unlikely to learn this two-level feature structure. In order to address these challenges, we propose a novel multi-task learning framework that learns area-specific patterns for predicting area-specific violence intensity attributed by gang-homicides. The study was carried out on Bogotá, a Colombian city with a high level of violent crime [1].

2 Related Work

Hotspot mapping is one of the most popular approaches for mapping crime-prone locations [4,5]. Crime has also been predicted using time series model ARIMA [6]. Twitter has been effective in identifying risk indicators [7,8]. The crime prediction literature has featured a variety of methods such as regression models [9], Bayesian approaches [10] and neural networks [11].

Of all the crime prediction models, only a handful of them focus solely on homicide. The use of hotspot maps to predict homicide and gun-crime can be found in [12]. Berk et al. [13] studied murder rates among probationers and parolees. Nineteen years of data were utilized to forecast homicide, robbery, burglary, and motor vehicle theft in [14].

Multi-task learning (MTL) is concerned with learning multiple related tasks simultaneously to achieve a better generalization performance [15]. Different assumptions on task relatedness result in different MTL strategies. For instance, assumptions can be made that the task parameters share a common subspace [16], or that they use a tree-structured model to share a common underlying structure [17]. MTL has been successfully employed in various applications including text classification, natural language processing, and computer vision.

3 Problem Formulation

We learn two different sets of keywords for two different levels of features: area-agnostic common keywords such as ‘cocaine’ (cocaine), ‘levantón’ (kidnapping),

'sicarios' (hitmen), and area-specific keywords focusing on gangs based on specific locations. Given a geographical region, the set of newspaper articles published in the past h days covering the news focused on the *i-th* area in that region is denoted by $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,h}\}$, where $x_{i,j}$ denotes a collection of news articles published on day j based on the *i-th* area. Violence intensity over a window of w days is denoted by $Y^{(1)}$. The number of homicides per day is denoted by $Y^{(2)}$. The prediction problems can be formulated as two different mappings. Firstly, we seek to learn $f : X_i \rightarrow {Y^{(1)}}_{i,h+k}$, where the right hand side denotes the intensity of violence in the *i-th* area, predicted k days before the target time window by reading past h days of news articles (classification problem). Another function $f : X_i \rightarrow {Y^{(2)}}_{i,h+k}$ maps the sets of news articles from past h days to the number of homicides that may be committed in future, again k days in advance (regression problem). From the ground truth data, we compute homicide statistics such as the average and median of actual number of homicides over our study period in each area. If the average number of homicides committed over a given time window exceeds the median, we identify a 'large scale' violence for that time period. Otherwise, it is 'small scale'. *History days*, denoted by h, refers to the number of days' articles in the past that the model would use to make a prediction. *Lead time*, denoted by k, refers to how many days in advance the model would make a forecast.

4 Models

We model S different tasks for S different locations. Our proposed strategy simultaneously learns models for all S locations in a multi-task feature learning framework. We divide our feature matrix W row-wise into G and R such that R rows follow the top G rows as shown in Fig. 1. G denotes the general features and R denotes the area-specific features. Our model minimizes the following:

$$\min_{W} \quad f(W) + \lambda_1 g_1(G) + \lambda_2 g_2(R), \tag{1}$$

where $f(W)$ is the empirical loss. We use the least squares loss which is smooth and convex. g_i represents the regularization function. The tunable parameter λ_i controls the model sparsity and balances the emphasis between the loss and the penalty. We propose: (I) multi-level multi-task (MLMT) model, and (II) constrained multi-level multi-task (cMLMT) model that simultaneously learn features for all areas in a multi-task, multi-level feature learning framework.

4.1 MLMT Model

We apply regularization at two levels to capture the multi-level feature representation. The models need to be able to take advantage of shared common features across locations and learn location-specific features. We apply $\ell_{2,1}$-norm to jointly learn a set of across-task features. Area-specific features for each task are selected in the next level when we directly apply ℓ_1-norm regularization on gang-related feature set R. The objective function for our proposed model is:

$$\min_{W=[G;R]} \sum_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda_1 \|G\|_{2,1} + \lambda_2 \|R\|_1. \tag{2}$$

S denotes the total number of tasks. $\mathcal{L}$ denotes the loss function. $\lambda_1 \|G\|_{2,1}$ denotes the $\ell_{2,1}$-norm on G. $\lambda_2 \|R\|_1$ is the ℓ_1-norm regularization which enforces individual sparsity on each task. λ_1 controls the group sparsity, and λ_2 controls sparsity in area-specific features. The $\ell_{2,1}$-norm on G makes the model select a common set of features for all the tasks while ℓ_1-norm on R learns features exclusively associated with each area.

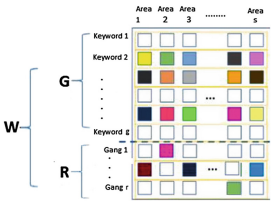

Fig. 1. Illustration of MLMT. Each column represents a model for an area. The rows represent the feature vectors. The general features are specified by first G rows. The area-specific features are represented by the rest (R rows).

4.2 cMLMT Model

This model offers a way to constrain the feature learning process. It is often desirable to identify the level of correlation between gangs and locations. In this formulation, we prohibit the area-specific features to take on negative scores. As a result, they become either zero or take on positive weights. This affects the overall scores distribution across tasks. As the models try to minimize the empirical loss, the gang-area correlation, modeled by area-specific features, are rearranged in such a way that the empirical loss do not increase significantly. The resulting area-specific weights offer insight into each gang's contribution in the ongoing violence in each location. Below is the constrained form of MLMT:

$$\begin{aligned} \min_{W=[G;R]} \quad & \sum_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda_1 \|G\|_{2,1} + \lambda_2 \|R\|_1. \\ \text{s.t.} \quad & R \geq 0 \end{aligned} \tag{3}$$

5 Algorithm

Both of our optimization problems have two non-smooth terms. Equation (3) is a constrained form of Eq. (2) such that $R \geq 0$. To solve these problems, we develop an algorithm based on proximal gradient descent. The basic idea is to

first use the gradient at the current search point and apply proximal operator on $W^i - \frac{1}{L}\nabla F(W^i)$ to find an approximate solution point. In other words, we find an approximate solution point by applying $prox_{\lambda g}(W^i - \frac{1}{L}\nabla F(W^i))$. This is a gradient step towards the optimal solution point. Line 5 of our algorithm can also be viewed as proximal operator of first order approximation. The approximate solution point found in the current iteration would be used as the *current* search point in the next iteration. The step size is $\frac{1}{L}$, and L is determined by a line search method. The details are given in Algorithm 1 where,

$$\nabla F(W) = X(X^T W - Y). \tag{4}$$

Algorithm 1. The Proposed Algorithm

Require: : $\mathbf{X}$, $\mathbf{Y}$, $\rho, \eta > 1$
Ensure: : solution $\mathbf{W}$
1: Initialize $W^0, \eta = 0.5$
2: **for** $i \leftarrow 1, 2, 3, \ldots$ **do**
3: Initialize L = 1
4: **repeat**
5: $\hat{W}^i \leftarrow W^i - \frac{1}{L}\nabla F(W^i)$ ▷ $\hat{W}^i = [\hat{G}^i; \hat{R}^i]$
6: $G^i \leftarrow prox_{2,1}(\hat{G}^i)$
7: $R^i \leftarrow prox_1(\hat{R}^i)$
8: $L \leftarrow \eta L$
9: **until** line search criterion is satisfied
10: **if** the objective stop criterion is satisfied **then**
11: **Return** W^i
12: **end if**
13: **end for**

Note that we divide W into G and R such that $W = [G; R]$. We have two sub-problems to solve: proximal $\ell_{2,1}$ regularized problem in line 6 and proximal ℓ_1 regularized problem in line 7, both of which have closed form solutions. We solve the proximal operator with $\ell_{2,1}$-norm on G by,

$$Prox_{2,1}(G) = (max(\|G\|_2 - \lambda, 0)/\|G\|_2)G. \tag{5}$$

Juxtaposition of two quantities implies matrix multiplication. Recall that G is the set of general features that occupies the top G rows in our feature matrix. For proximal of ℓ_1 on R, the closed form solution is given by,

$$Prox_1(R) = sign(R).max(abs(R) - \lambda, 0). \tag{6}$$

The dot denotes element-wise multiplication. For the constrained optimization problem given in (3), the solution to proximal operator with ℓ_1-norm is given by,

$$Prox_1(R) = max(abs(R) - \lambda, 0). \tag{7}$$

In every iteration, the algorithm finds an approximate solution point that gets closer to the optimal solution point. The algorithm iterates until the optimal point is found, or the maximum number of iteration is reached.

6 Experiments

6.1 Dataset

The experiments were carried out on 10,672 newspaper articles collected from several news agencies such as El Colombiano, El Universal, RCN Radio, El Tiempo, El Confidencial, NTN24, and El Nuevo Herald between April 2015 and May, 2016. The articles were in Spanish. We used police records on homicides in Bogotá for evaluating our model's performance.

6.2 Data Preprocessing

We worked on three regions in Bogotá: far Northwest, center and center south, and far south. Often the articles refer to multiple locations. We use the geometric median of the GPS coordinates of the localities appearing in an article to determine the finer-grained location information. Each news article is assigned a location based on the geometric median, m, of the GPS locations, L, of the areas mentioned in that news article.

$$m = \underset{x \in L}{\operatorname{argmin}} \sum_{y \in L} distance(x, y), \tag{8}$$

where $distance(x, y)$ is the orthonormic distance calculated using Vincentry's formula [18]. There is a possibility that some news articles are not assigned to any of our three target areas even though it may belong to one. To compensate this situation, we accommodate a fourth task that contains all the news articles that are based on Bogotá, but are not assigned to any of our three pre-defined regions.

6.3 Experimental Setup

We denote 'large scale' violence by 1, and 'small scale' violence by 0. Our model outputs either 0 or 1 in the violence intensity setting. Examples of general keywords can be found in Sect. 3. For area-specific features, we use names of gangs, armed groups, and members of those groups such as 'Los Rastrojos', 'Clan Úsuga', 'Pastor Alape', which are either drug-trafficking paramilitary groups or members of those groups.

The input for each task is an $n \times m$ matrix where n denotes the number of input samples, and m denotes number of features. Each input sample (i.e., row) is constructed by counting the frequencies of the features occurring in the news articles published in past h days starting from a particular date. Each cell in that row, therefore, represents the frequency of a specific feature (i.e.,

general keyword or area-specific keywords) in the news articles over the same time period. Imagine a sliding window that starts from the starting date of the training period with a window width of h. We slide this window over time until the right end of that window touches the end date of the training period. While the window slides, the input matrix gets constructed.

We have three tunable parameters in our model: lead time k, history days h, and time window w are the tunable parameters. As an example, if k is 3, h is 5, and w is 4, then the model will read past 5 days of news articles to predict the violence intensity over 4 days; the prediction will be made 3 days in advance. Changing the values of these variables would yield different models, each having their own specification of lead time, time window, and history days. In this article, we show the results when $k = 1$, $h = 5$, and $w = 2$. The regularization parameters λ_1 and λ_2 were selected via a 5-fold cross-validation.

6.4 Comparison Methods

For the classification task, we compare our proposed models with Support Vector Machine (SVM), Logistic Regression, regularized LASSO and the baseline approach **monotonic multi-task** (MMT) given by:

$$\min_{W} \sum_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda \|W\|_{2,1}. \tag{9}$$

The baseline method does not distinguish between general and area-specific features. The regularization parameter λ was selected via a 5-fold cross-validation. These models are area-ignorant in the sense that they do not capture the multi-level feature structure. We use an L_2-penalized logistic regression and the *liblinear* solver. For the SVM, we use the radial basis function (*rbf*) kernel with co-efficient gamma set to 0.7. The regularization parameter λ for LASSO were determined via 5-fold cross-validation. For the regression task, we compare our model with seasonal ARIMA and Support Vector Regression (SVR). We use the police record data to build the seasonal ARIMA model. For SVR, we use the *rbf* kernel with the penalty parameter set to 0.8. Note that the parameter values for the comparison methods were selected by a trial and error method. We select the values that give the best performance for each comparison model.

6.5 Results and Discussion

For the homicidal violence prediction task, we consider four performance metrics: precision, recall, F1-score, and ROC AUC (Area Under the Receiver Operating Characteristic Curve). Table 1 shows the performance comparisons between our proposed model and the comparison methods for the classification task. The results show that our proposed model MLMT performs better, on average, than other methods. MLMT outperforms the baseline method by 5% to 10% in precision, recall, and F1-score in Area 1. In Area 3, the baseline is outperformed by MLMT by 10.3% to 20.8% in precision, recall, F1-score, and AUC. This implies

Table 1. Violence intensity prediction performance comparison (precision, recall, F1-score, AUC).

Methods	Area 1 p, r, f1, auc	Area 2 p, r, f1, auc	Area 3 p, r, f1, auc	Area 4 p, r, f1, auc
Logistic regression	0.44, 0.5, 0.47, 0.23	0.40, 0.41, 0.40, 0.46	0.46, 0.46, 0.46, 0.49	0.49, 0.49, 0.49, 0.57
SVM	0.44, 0.5, 0.47, 0.35	0.22, 0.5, 0.31, 0.53	0.4, 0.5, 0.44, 0.59	**0.79**, 0.51, 0.62, 0.57
LASSO	0.47, 0.44, 0.46, 0.45	0.50, 0.50, 0.50, 0.47	0.59, 0.58, 0.59, 0.59	0.51, 0.51, 0.51, 0.51
MMT	0.64, 0.83, 0.72, **0.97**	**0.69, 0.72, 0.71**, 0.77	0.69, 0.81, 0.75, 0.79	0.66, **0.68, 0.67**, 0.69
MLMT	**0.74, 0.88, 0.81, 0.97**	**0.69, 0.72**, 0.70, **0.78**	**0.71, 0.83, 0.76, 0.80**	0.65, 0.67, 0.66, **0.7**
cMLMT	0.64, 0.83, 0.72, 0.94	0.67, 0.71, 0.69, 0.72	0.68, 0.79, 0.73, 0.79	0.64, 0.64, 0.64, 0.69

Table 2. Homicide count prediction performance comparison (RMSE, MAE).

Methods	Area 1 rmse, mae	Area 2 rmse, mae	Area 3 rmse, mae	Area 4 rmse, mae
SVR	0.87, 0.75	1.58, 1.25	1.65, 1.34	1.65, 1.34
SARIMA	**0.11, 0.01**	0.88, 0.62	0.81, 0.60	0.79, 0.56
LASSO	0.37, 0.14	0.92, 0.64	0.52, 0.27	1.37, 0.86
MMT	0.28, 0.08	**0.53, 0.29**	0.37, 0.14	**0.60, 0.36**
MLMT	**0.26, 0.07**	0.54, **0.29**	**0.36, 0.13**	**0.60**, 0.37
cMLMT	0.28, 0.08	0.55, 0.30	0.38, 0.14	0.61, 0.38

that each location does have its own specific factors that affect the intensity of homicide-induced violence in that area.

Table 2 shows a performance comparison for the regression task. We use RMSE and MAE as the performance metrics. Note that the seasonal ARIMA model was not constrained with *history days* and *lead time*. It enjoyed as much data as we had for the training period with no restriction on history days, which may have attributed to better RMSE and MAE scores for Area 1. However, if we compare MLMT with only the baseline MMT, it outperforms the baseline. We present a comparison of the performances by varying lead time in Fig. 2. **History days** was fixed to 5. Figure 2 (left panel) shows that the MLMT model achieves better F1 score than the others, especially with increased lead time. This is explained by the fact that a precursor incident such as an arrest or a murder committed by a rival group will not necessarily generate an immediate reaction. Often, the rival group's attempt to take control of a local business controlled by another group, or a retaliatory murder may take some time to happen. This gap between an event and the reactive violence may be a reason for why the models generally perform better with an increasing lead time.

Figure 2 (right panel) also shows a performance comparison in AUC when the number of history days varies. *Lead time* was fixed to 1 day with varying number of history days. We compare only the baseline method and MLMT since other models perform worse. Figure 2 shows that MLMT mostly performs better, or no less than the baseline method. This consistency in better F1 score and AUC when the lead time and history-days vary shows the necessity of capturing the multi-level feature structure for predicting gang-related homicides.

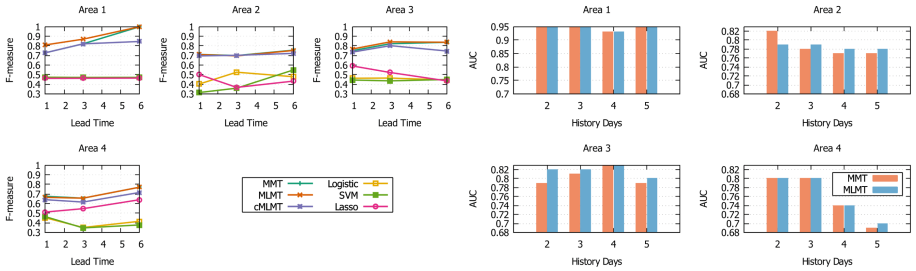

Fig. 2. Visualization of performance comparison. F-measure comparison when the lead time varies (left) and AUC comparison when the number of history days varies (right).

Table 3. Model-selected top 4 area-specific features

Area 1	Area 2	Area 3	Area 4
Mao	Roman Ruiz	El Médico	ELN
Clan Úsuga	AUC	EPL	Otoniel
Comba	FARC	El Coronel	FARC
Omar	clan Úsuga	Roman Ruiz	El Coronel

Table 3 shows the model-selected gang-related features. While the general keywords present an area-agnostic global view of the feature space, the gang-related features demonstrate a subtle dependency on the spatial areas. For instance, violence in Area 2 connects highly with three armed groups: FARC, AUC (Autodefensas Unidas de Colombia), and Clan Úsuga. AUC was a rival of FARC, and Clan Úsuga emerged when AUC was being demobilized. While Clan Úsuga is also a top contributor to violence in Area 1 with its leaders Omar and Mao, another group Rastrojos also contributes via its leader Comba in the same area. Rastrojos is a rival of Clan Úsuga. Rivalry leads to more violence in general. We find a different group EPL (Ejército Popular de Liberación) affecting Area 3. FARC is also present in Area 4 together with its another former rival ELN (National Liberation Army). The rest in Table 3 are members of the aforementioned groups. The area-agnostic features together with these gang-related area-specific features indicate a multi-level feature structure. Note that the sets of top contributors for each task are mostly different from each other.

7 Conclusion

We present a novel approach that learns features at two different resolutions to predict gang-homicide and violence intensity. Existing homicide prediction works do not distinguish between shared common information across locations and location-specific information. Our proposed method addresses these issues

by simultaneously learning models for multiple tasks while capturing the multi-level structure of the features. Empirical results show that our proposed model can effectively predict gang-homicides and homicidal-violence intensity.

Acknowledgments. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

1. Igarapé Institute: Homicide monitor (2017)
2. Muggah, R.: Interactive Map Tracks Murder Rate Worldwide (2015)
3. Saul, J.: Why 2016 Has Been Chicago's Bloodiest Year in Almost Two Decades (2016)
4. Gorr, W.L., Lee, Y.: Early warning system for temporary crime hot spots. J. Quant. Criminol. **31**(1), 25–47 (2015)
5. Weisburd, D., Braga, A.A., Groff, E.R., Wooditch, A.: Can hot spots policing reduce crime in urban areas? An agent-based simulation. Criminology **55**(1), 137–173 (2017)
6. Chen, P., Yuan, H., Shu, X.: Forecasting crime using the ARIMA model. In: Fifth International Conference on FSKD 2008, vol. 5, pp. 627–630. IEEE (2008)
7. Gerber, M.S.: Predicting crime using Twitter and kernel density estimation. Decis. Support Syst. **61**, 115–125 (2014)
8. Wang, X., Brown, D.E., Gerber, M.S.: Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In: 2012 IEEE International Conference on ISI, pp. 36–41. IEEE (2012)
9. Shingleton, J.S.: Crime trend prediction using regression models for Salinas, California. Ph.D. thesis. Naval Postgraduate School, Monterey, California (2012)
10. Liao, R., Wang, X., Li, L., Qin, Z.: A novel serial crime prediction model based on Bayesian learning theory. In: 2010 International Conference on ICMLC, vol. 4, pp. 1757–1762. IEEE (2010)
11. Kang, H.W., Kang, H.B.: Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE **12**(4), e0176244 (2017)
12. Mohler, G.: Marked point process hotspot maps for homicide and gun crime prediction in chicago. Int. J. Forecast. **30**(3), 491–497 (2014)
13. Berk, R., Sherman, L., Barnes, G., Kurtz, E., Ahlman, L.: Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. J. R. Stat. Soc.: Ser. A **172**(1), 191–211 (2009)
14. Pepper, J.V.: Forecasting crime: a city-level analysis. In: Understanding Crime Trends: Workshop Report, National Research Council, pp. 177–210 (2008)
15. Zhao, L., Sun, Q., Ye, J., Chen, F., Lu, C.T., Ramakrishnan, N.: Multi-task learning for spatio-temporal event forecasting. In: Proceedings of the 21th ACM SIGKDD, pp. 1503–1512. ACM (2015)
16. Acharya, A., Mooney, R.J., Ghosh, J.: Active multitask learning using supervised and shared latent topics. In: Pattern Recognition and Big Data, p. 75 (2016)
17. Kim, S., Xing, E.P.: Tree-guided group lasso for multi-task regression with structured sparsity (2010)
18. Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Surv. Rev. **23**(176), 88–93 (1975)