总结报告15

论文4篇

（2020.2.16—2020.2.17）

一、contents

Wang, W., Wang, J., Kolar, M., & Srebro, N. (2018). **Distributed Stochastic Multi-Task Learning with Graph Regularization**.

Xu, J., Tan, P. N., Luo, L., & Zhou, J. (2016). **GSpartan: A geospatio-temporal multi-task learning framework for multi-location prediction**. 16th SIAM International Conference on Data Mining 2016, SDM 2016, 657–665.

Li, C., Huang, S., Liu, Y., & Zhang, Z. (2018). **Distributed jointly sparse multitask learning over networks**. IEEE Transactions on Cybernetics, 48(1), 151–164.

Verma, V. K., & B, P. R. (2017). **Distributed Multi-task Learning for Sensor Network**. 1, 792–808.

二、总结

Wang, W., Wang, J., Kolar, M., & Srebro, N. (2018). **Distributed Stochastic Multi-Task Learning with Graph Regularization**.

主旨说明：

1. 对W的限制条件之一：两个近的点间w相差不大，利用图矩阵L，变成正则项进入目标函数中
2. 使用随机梯度SGD，减少计算次数

论文脉络：

1. Introduction
2. Graph-based multi-task learning
3. Distributed algorithm for ERM

3.1 Directly solving the regularizer

3.2 Directly optimizing the loss

4. Stochastic algorithms

4.1 Directly solving the regularizer

4.2 Directly optimizing the loss

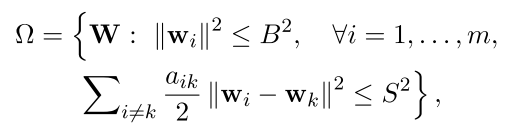
5. Connection to consensus learning

6. Experiments

Details:

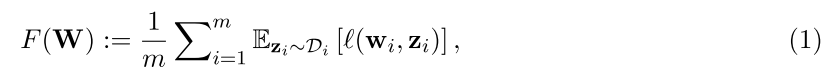
1. Introduction
2. Graph-based multi-task learning：提出了基于图矩阵的目标函数

权重矩阵要满足下面的要求：



这个Ω的意义是：范数有界，且related的predictors不相似度小

全局的优化目标为：



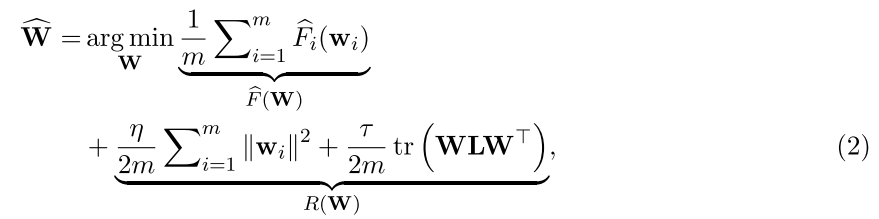
其中，为每个node的优化目标。

注意，对集Ω的应用：

对集Ω其中一个条件利用Laplacian 图矩阵，即L，则该条件可以转化为



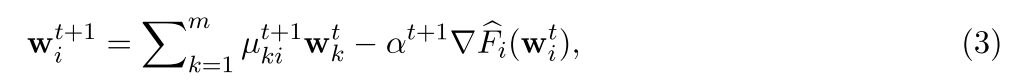
Regularized ERM (Empirical Risk Minimization) Objective：



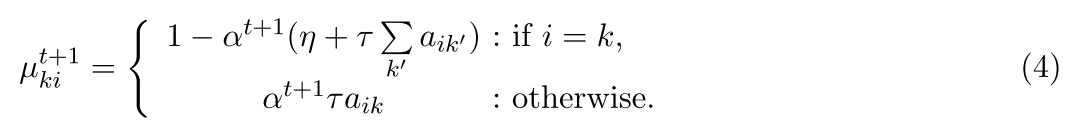
1. Distributed algorithm for ERM

解决目标函数（2）：

首先对使用gradient descent



其中，为weights for combining neighboring predictors



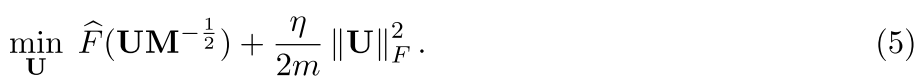
正常要对和都线性化，但是作者在这里仅对其中一个线性化

3.1 Directly solving the regularizer

首先，利用变换



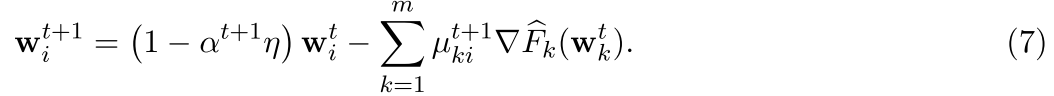
将目标函数改为，



关于U使用GD，降回W-space:



分布到每一个node：

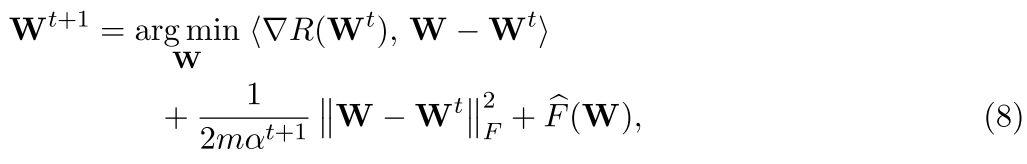


Note：通过通信一圈得到每个node的梯度，M-1提前计算

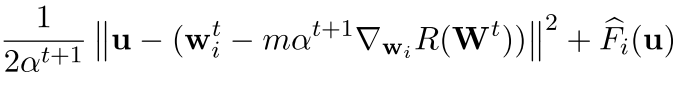
3.2 Directly optimizing the loss

由于上面的算法需要通信一圈，不太符合分布式的设置，于是得到下面的算法：

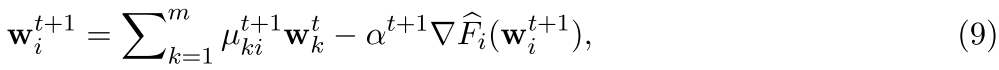
linearize the graph regularizer but fully optimize over the loss:



分配到node：||W-Wt||2F线性化后可能有不好的性质，强凸保证



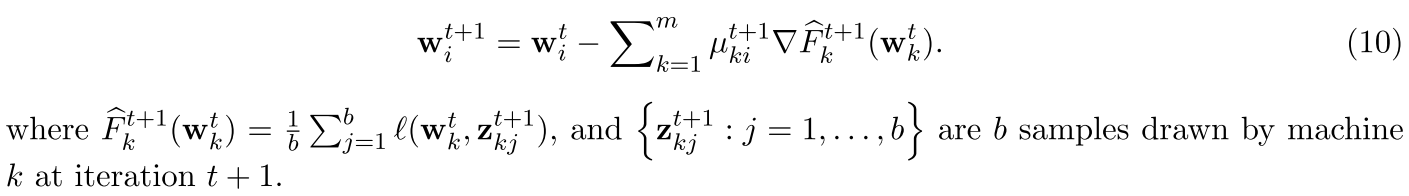
即，



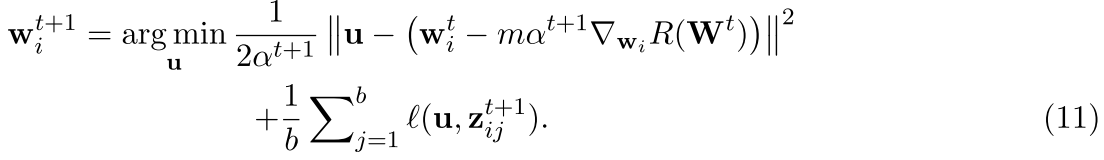
4. Stochastic algorithms

4.1 Directly solving the regularizer

minibatch SGD with **b samples** per machine



4.2 Directly optimizing the loss



Xu, J., Tan, P. N., Luo, L., & Zhou, J. (2016). **GSpartan: A geospatio-temporal multi-task learning framework for multi-location prediction**. 16th SIAM International Conference on Data Mining 2016, SDM 2016, 657–665.

主旨说明：提出GSpartan模型

1）Local models共享一些公共的、低秩representations

2）每个局部模型（local models）作为基模型（base models）的线性组合

3）有两种变量：response variable，即我们想要预测的变量；

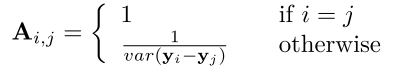
predictor variables，是由全局或地区模型产生的其他输出

比如，预测每月降雨量是response variable ;mean temperature at 2 meters, mean sea level pressure, 500 hPa geopotential height, and near surface relative humidity是predictor variables。

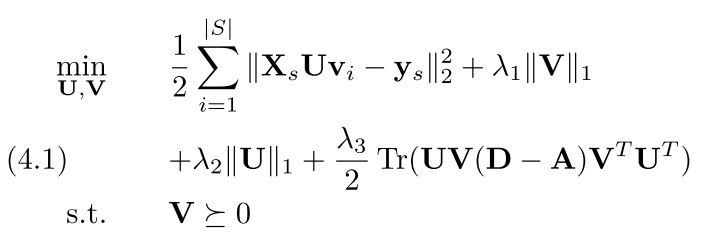
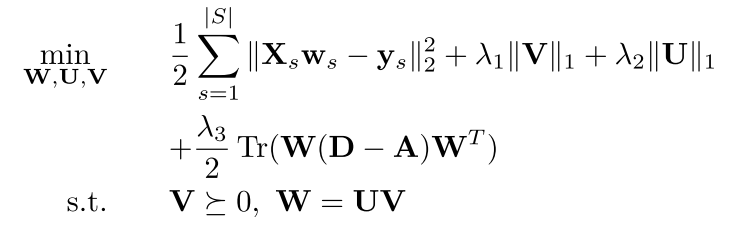
4）实验部分通过去掉拉普拉斯图矩阵的正则化进行比较，还和一些其他方法进行了比较

5）拉普拉斯图矩阵：通过修正变差函数的逆（inverse of a modified variogram）给出

Variogram：方差图是一种空间统计中发展起来的一种度量方法，用来确定一对位置之间的空间依赖性。见引用文献[10]



目标函数：

Note：W=UV

U：a feature representation of the base models

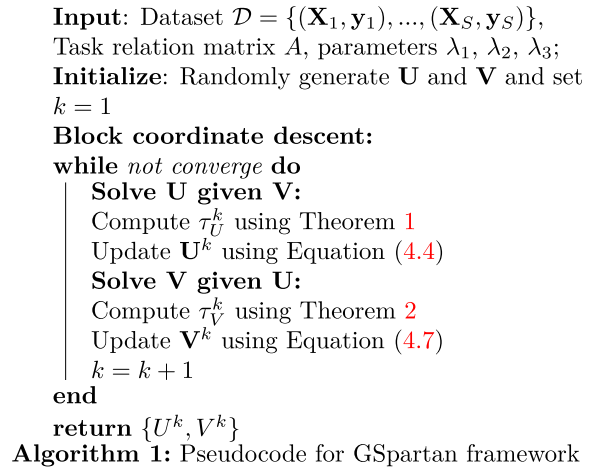
V：the weighted combination of the base models that form the local model at each location.

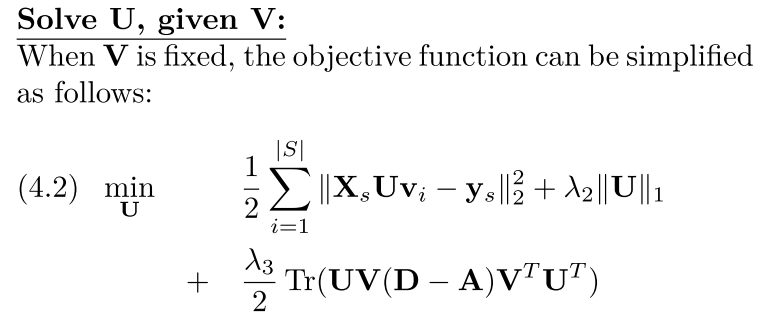
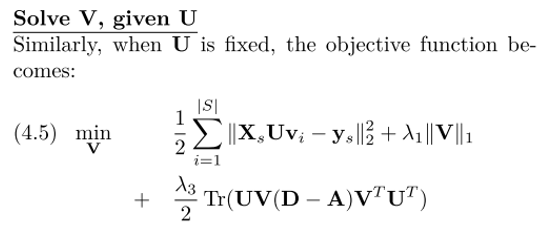
矩阵U是基本模型的特征表示，V表示在一点形成局部模型的基本模型的加权组合。

A：拉普拉斯图矩阵，引入空间自相关

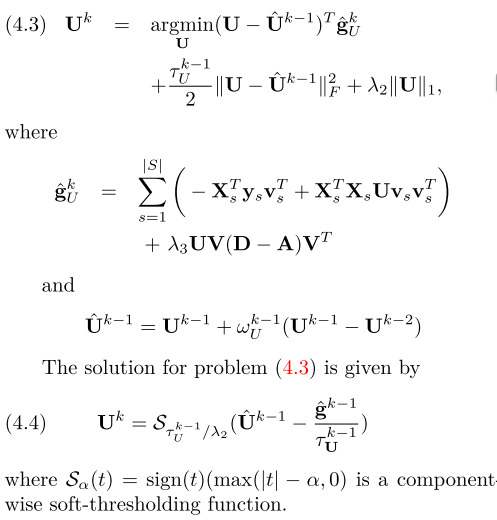
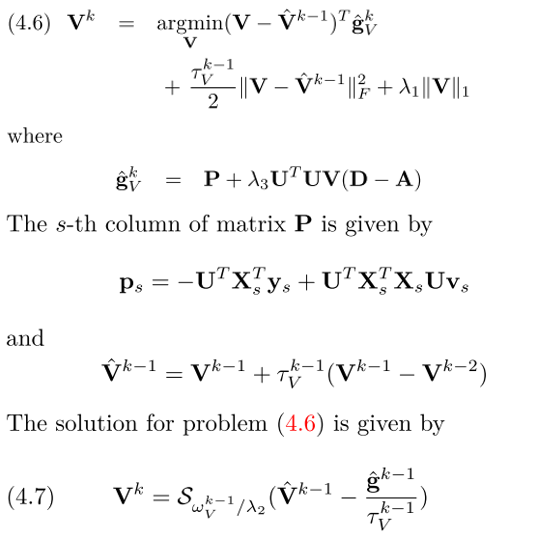
优化算法：

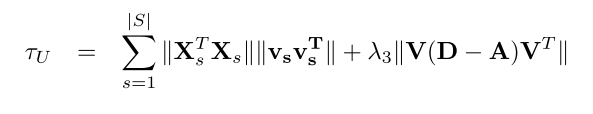
通过BCD，交替优化U和V





通过proximal gradient descent进一步计算， 同样通过proximal gradient descent，





Li, C., Huang, S., Liu, Y., & Zhang, Z. (2018). **Distributed jointly sparse multitask learning over networks**. IEEE Transactions on Cybernetics, 48(1), 151–164.

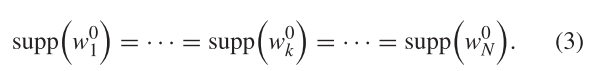
主旨说明：AC-dJSMT algorithm

1. 文章考虑了多任务的两点：

的一项为惩罚intertask similarities；的一项用来惩罚joint sparsity。

如果只考虑其中一种情况，只须让对应的项的系数为0.

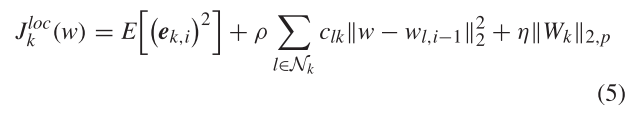
1. 为了分布式将W分配给每个节点Wk：a local parameter matrix（consisting of the parameter vectors of all its neighbors）
2. 作者认为joint sparsity意味着每一个node的参数向量哪个位置是0，哪个位置不是0，对于所有的节点都是一样的，比如第一个分量all nodes全是0，第二个分量全不是0…这样的结构，即

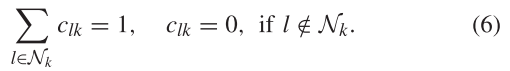


问题：认为所有nodes都一样的形式，这样的假设合理吗？现实中有这样的例子吗？

4) intertask similarities的部分intertask combiner，见下优化算法所述。

目标函数：

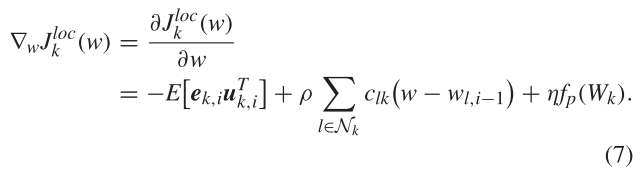


Note：.另外有：intertask combiners

用来惩罚稀疏

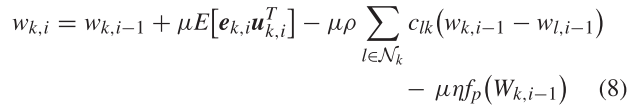
解决方法：

steepest-descent method

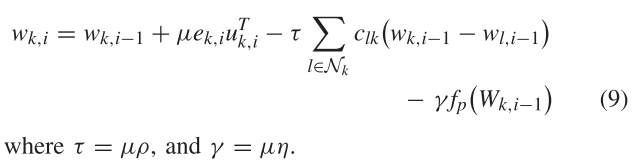


Note:是关于w的L2，p范数的次梯度

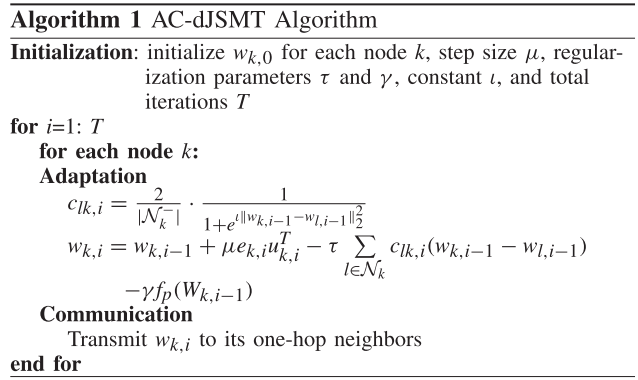
迭代更新公式：



用来代替得到，



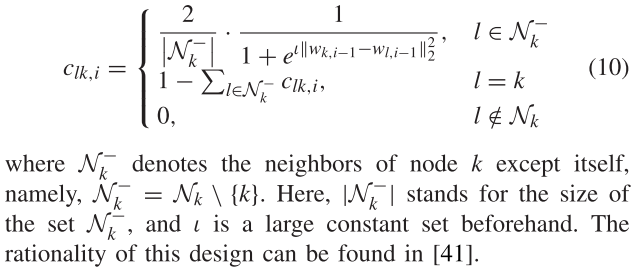
优化算法：



·当is fixed，得到了FC-dJSMT（fixed  distributed jointly sparse multitask）

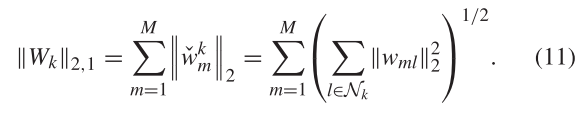
·当is adaptive，得到了AC-dJSMT（adaptive  distributed jointly sparse multitask）

由下面的公式给出的取值，

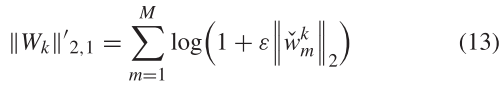


·对于稀疏惩罚项||W||2，p的范数的讨论

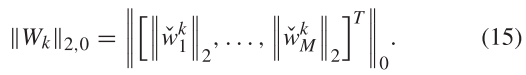
1. L2，1-norm



1. Reweighted L2，1-norm ：RW范数



1. L2，0-norm



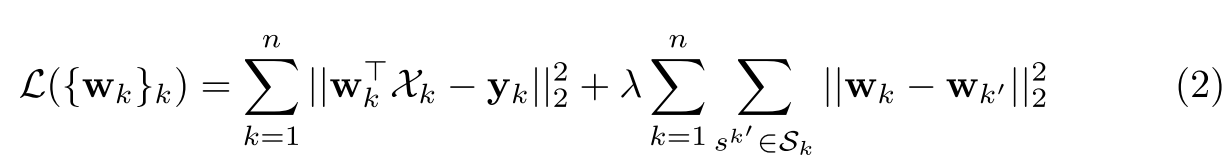
Verma, V. K., & B, P. R. (2017). **Distributed Multi-task Learning for Sensor Network**. 1, 792–808.

主旨说明：

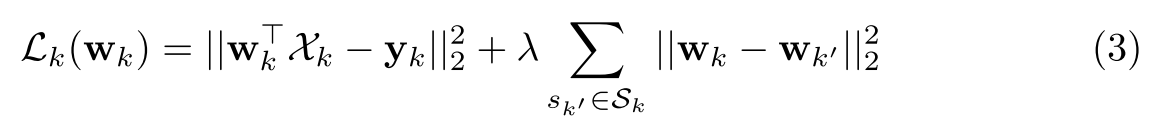
比较中规中矩的一篇分布式+多任务学习，在neighbors间传递**参数parameter**

目标函数：

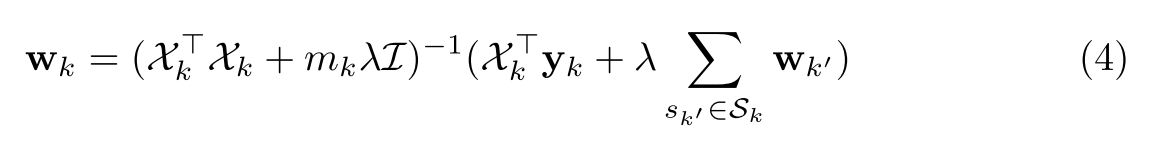
全局的：



分配到每一个节点的：

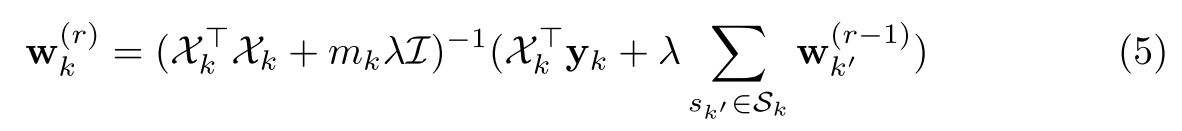


得到：



优化算法：

第r次迭代：



DMTL算法:

