Two-stage transfer surrogate model for automatic hyperparameter optimization.

- 作者：Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme
- 机构：University of Hildesheim
- 会议：ECML2016
- 地址：http://arxiv.org/abs/1802.02219
- 代码：https://github.com/wistuba/TST

论文主要内容

摘要

通过在其他data sets上获得的knowledge来提升在新data sets上的贝叶斯优化。文章通过两个stage来提升transfer knowledge。第一个stage拟合hyperparameter response function（理解成一个回归模型，超参=>score）。第二个stage利用第一个stage的所有model组装成一个surrogate model.

研究内容

定义：

$$\mathcal{A}_{\boldsymbol{\lambda}}\left(D^{\text{train}}\right)=\arg\min_{M_{\boldsymbol{\lambda}}\in\mathcal{M}}\mathcal{L}\left(M_{\boldsymbol{\lambda}},D^{\text{train}}\right)+\mathcal{R}\left(M_{\boldsymbol{\lambda}}\right)$$

hyperparameter response function: $f_D(\boldsymbol{\lambda})=\mathcal{L}\left(\mathcal{A}_{\boldsymbol{\lambda}}\left(D^{\text{train}}\right),D^{\text{valid}}\right)$

Algorithm 1 Sequential Model-based Optimization

Input: Hyperparameter space Λ, observation history $\mathcal{H}$, number of trials T, acquisition function a, surrogate model Ψ.
Output: Best hyperparameter configuration found.

for $t=1$ to T **do**
  Fit Ψ_{t+1} to $\mathcal{H}_t$
  $\boldsymbol{\lambda}\leftarrow\arg\max_{\boldsymbol{\lambda}\in\Lambda}a\left(\mu\left(\Psi_{t+1}\left(\boldsymbol{\lambda}\right)\right),\sigma\left(\Psi_{t+1}\left(\boldsymbol{\lambda}\right)\right),f^{\min}\right)$
  Evaluate $f\left(\boldsymbol{\lambda}\right)$
  $\mathcal{H}_{t+1}\leftarrow\mathcal{H}_t\cup\{(\boldsymbol{\lambda},f\left(\boldsymbol{\lambda}\right))\}$
  if $f\left(\boldsymbol{\lambda}\right)<f^{\min}$ **then**
    $\boldsymbol{\lambda}^{\min},f^{\min}\leftarrow\boldsymbol{\lambda},f\left(\boldsymbol{\lambda}\right)$
return $\boldsymbol{\lambda}^{\min}$

方法

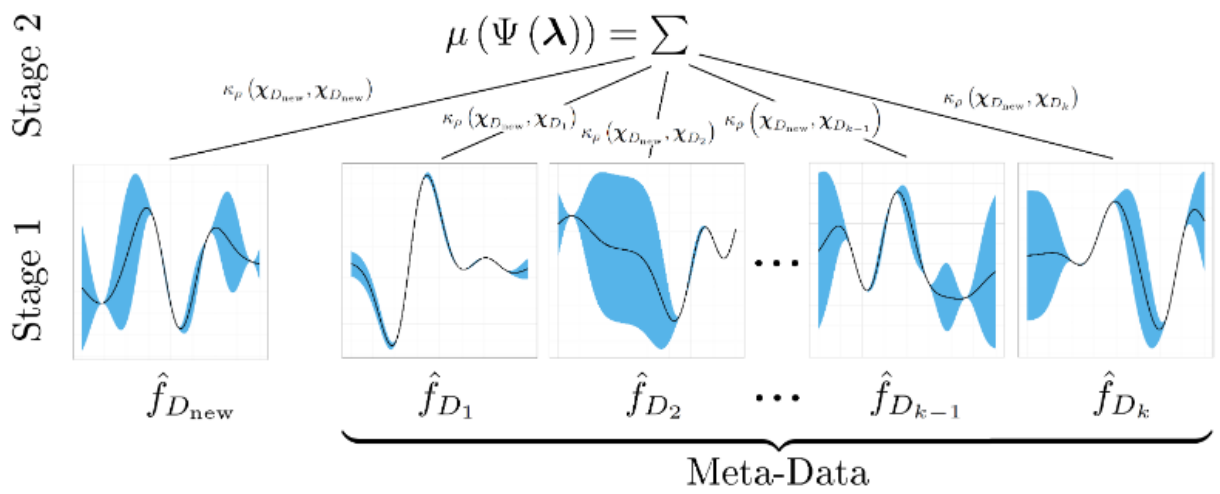

Fig. 1. At the first stage the hyperparameter response functions of the new data set D_{new} as well as data sets $\mathcal{D} = \{D_1, \ldots, D_k\}$ used for previous experiments are approximated using known evaluations. At the second stage the predictions of each individual model $\hat{f}_D$ are taken into account weighted by the similarity between D and D_{new} to determine the final predicted score.

第一个stage

meta-dataset

每个data set有288个Meta-instances （超参，score）、归一化label，平衡不同数据任务的scale

拟合meta-instance，使用高斯过程模型（可以使用任何其他ML model）

model： $\hat{f}_D$

第二个stage

组合第一个stage的所有model：Nadaraya-Watsonkernel-weighted average

代理模型（surrogate model） 的均值函数、核函数：取candidate最大的

$$\mu(\Psi(\boldsymbol{\lambda})) = \frac{\sum_{D \in \mathcal{D} \cup \{D_{\text{new}}\}} \kappa_\rho \left(\boldsymbol{\chi}_{D_{\text{new}}}, \boldsymbol{\chi}_D\right) \hat{f}_D(\boldsymbol{\lambda})}{\sum_{D \in \mathcal{D} \cup \{D_{\text{new}}\}} \kappa_\rho \left(\chi_{D_{\text{new}}}, \boldsymbol{\chi}_D\right)}$$

$$\sigma(\Psi(\boldsymbol{\lambda})) = \sigma\left(\hat{f}_{D_{\text{new}}}(\boldsymbol{\lambda})\right)$$

其中，

$$\kappa_\rho(\chi_D, \chi_{D'}) = \delta\left(\frac{\|\chi_D - \chi_{D'}\|_2}{\rho}\right)$$

$$\delta(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

χ_D 是用来描述data的向量

后面就是使用EI（Expected Improvement）acquisition function过程

Data Set Description

1. 使用Meta-Features：(TST-M)

Table 1. The list of all meta-features used by us.

Meta-Features	
Number of Classes	Class Probability Max
Number of Instances	Class Probability Mean
Log Number of Instances	Class Probability Standard Deviation
Number of Features	Kurtosis Min
Log Number of Features	Kurtosis Max
Data Set Dimensionality	Kurtosis Mean
Log Data Set Dimensionality	Kurtosis Standard Deviation
Inverse Data Set Dimensionality	Skewness Min
Log Inverse Data Set Dimensionality	Skewness Max
Class Cross Entropy	Skewness Mean
Class Probability Min	Skewness Standard Deviation

2. Pairwise Hyperparameter Performance Rankings: (TST-R)

$$\left(\boldsymbol{\chi}_D\right)_{j+(i-1)t} = \begin{cases} 1 & \text{if } \hat{f}_D\left(\boldsymbol{\lambda}_i\right) > \hat{f}_D\left(\boldsymbol{\lambda}_j\right) \\ 0 & \text{otherwise} \end{cases}$$

t个obsversion内的超参的rank来描述一个data

实验结果

metric：

1. average rank
2. average distance to the global minimum： $\operatorname{ADTM}\left(\Lambda_t, \mathcal{D}\right) = \sum_{D \in \mathcal{D}} \min_{\boldsymbol{\lambda} \in \Lambda_t} \frac{f_D(\boldsymbol{\lambda}) - f_D^{\min}}{f_D^{\max} - f_D^{\min}}$

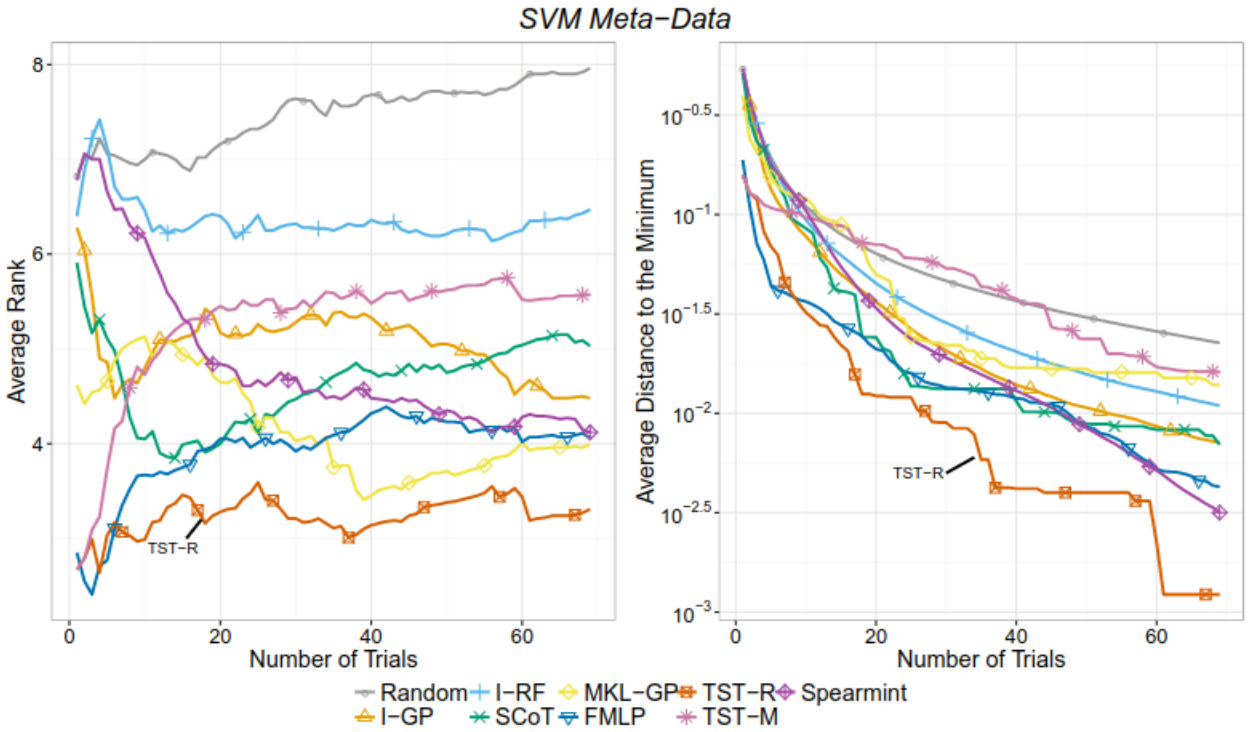

Fig. 2. Our proposed transfer surrogate model TST-R provides the best performance with respect to both evaluation measures for the task of hyperparameter tuning. For both metrics, the smaller the better.

less descriptive? NO

TST-R的distance会在每次增加observe之后更新，只关注在knowledge在新数据的某时间点上

meta-feature后期没有把新数据的信息全考虑进去。

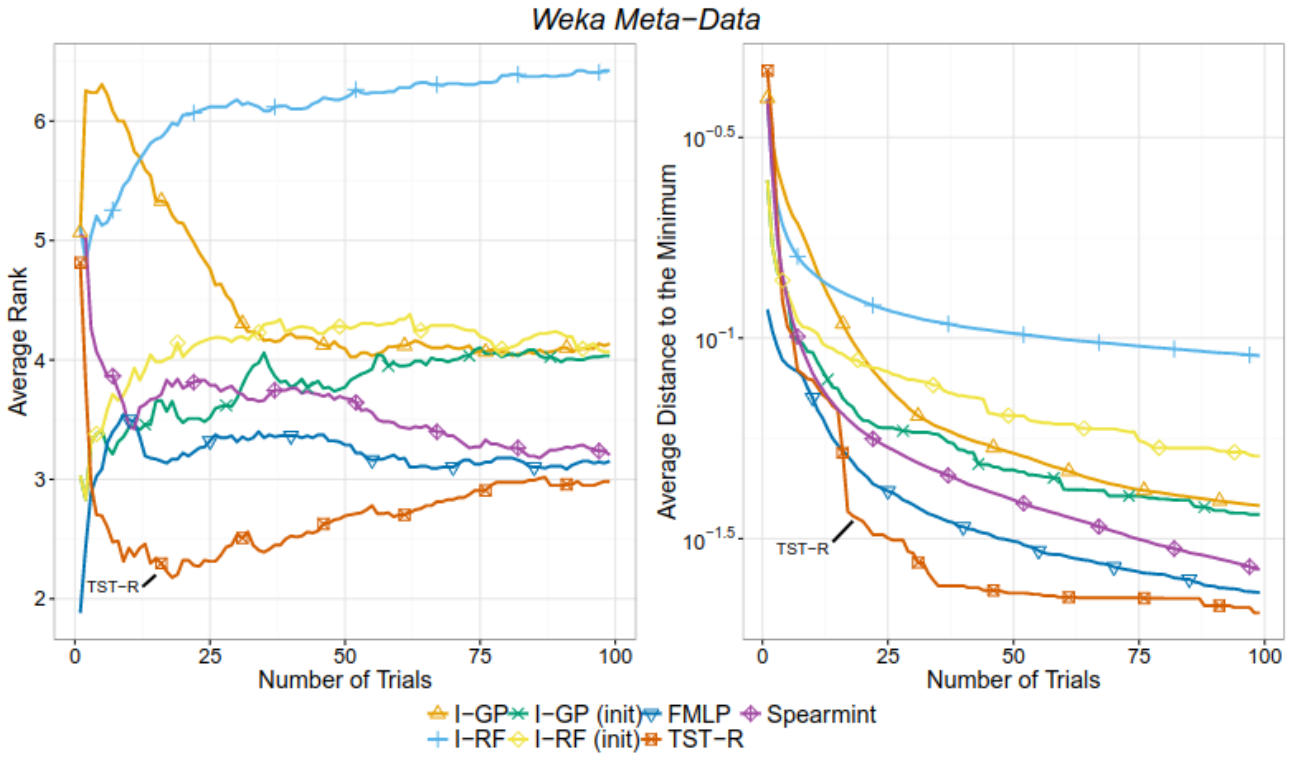

Fig. 3. Our approach TST-R also outperforms the competitor methods for the task of combined algorithm selection and hyperparameter tuning. Surrogate models that use Gaussian processes that train over the whole meta-data are not feasible for this data set [31]. Therefore, we consider I-GP and I-RF with meta-learning initialization.