# 2022.4.7 Paper Reading

***Retail sales forecasting with meta-learning*** (Ma & Flides, 2021) EJOR

马少辉 (nau.edu.cn) （物流管理系）

# Retail sales forecasting with meta-learning

核心思路：利用卷积神经网络学习特征，用以进行基于特征单独组合预测。

作者认为的贡献有4：

- 第一个实证评估元学习在零售产品销售预测环境中的表现；
- 提出了一种新颖的元学习器，它可以**自动从原始时间序列数据中学习特征表示。**
- 探讨了基础预测器的成分对元学习器预测性能的影响；
- 研究了**从外部潜在影响（外生变量）中提取特征**的价值。

论文在 FFORMA 的基础上进行改进，具体是：

- 认为 FFORMA 优化预测误差组合（$\sum_{i=1}^{k} w_i l(f_i)$）不如**优化组合预测误差**（$l(\sum_{i=1}^{k} w_i f_i)$）直接——使用后者作为元学习器优化目标；
- 特征提取多主观、无监督——使用卷积神经网络从原始序列进行**特征学习**；（也叫表征学习）
- 除销售数据外，还从**外生因素**的序列中学习特征（如价格、事件、假日、广告等）；简单时序预测方法在无促销期时表现好，但是在促销期包括促销因素可以提高准确率。
- 组合局部预测与全局预测。

对第一点有一些思考：

- （最初认为，如果对组合预测损失优化，那么学习和元学习的区别在哪里？）
- 之前也尝试过在 LSTM 的时变元学习器中考虑优化组合预测误差，但是结果差不多甚至略差；
- 从理论上讲，后者在损失函数为凸且权重和为1的情况下比前者小，而且在允许权重不为1甚至加截距的前提下易于扩展，前者偏差过大；
- 前者的优势在于求导较容易，因为优化目标是权重，其是线性组合形式，只需额外求权重对特征的导数。这一点对 FFORMA 很重要，因为 XGBoost 要求二阶导；但是在基于神经网络/深度学习的元学习器下，一般是用一阶导优化，求导负担下降了，故后者在计算上应该可以接受？

## 方法

**元学习结构**：

与FFORMA相比，多了一步提取特征的过程。假设参考集、测试集中的 SKU 使用相同宽度的滚动窗口进行预测，以便我们可以在元学习和元预测阶段使用相同长度的数据来拟合基础预测。数据先被导入到卷积网络中用于提取特征，特征在代入到网络的下一个模块用于学习权重分配。

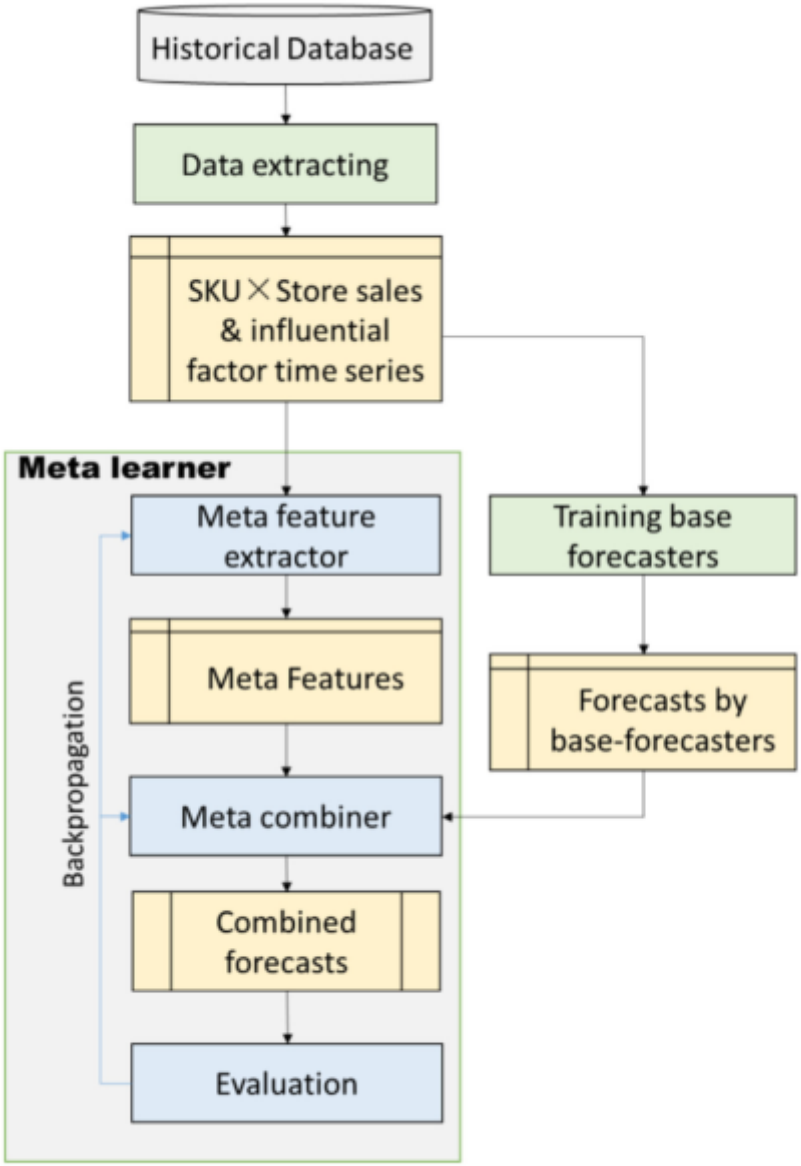



**Fig. 1.** A meta-learning framework for retail sales forecasting.

元学习器：借鉴CNN在提取图片特征的经验。将用于基准模型预测的数据输入至元学习器中。（由下文实验可知，每一块输入长度都是相同的；M5数据可能是长短不一的）

元学习器的网络结构如下：两通道输入，分别是销售数据与外部特征。特征提取部分均包含三个堆叠的卷积块，都包含卷积层和ReLU激活。

前两个层有**挤压**与**激发**层，挤压操作通过使用全局平均池在学习的特征图上生成摘要统计信息，激发操作目标是捕获学习特征之间的依赖关系，使用sigmoid作为门控机制激活。有文献表明使用此层可以提高卷积层产生的表示的质量(Hu et al., 2019)。挤压层与激发层数学表示如下：

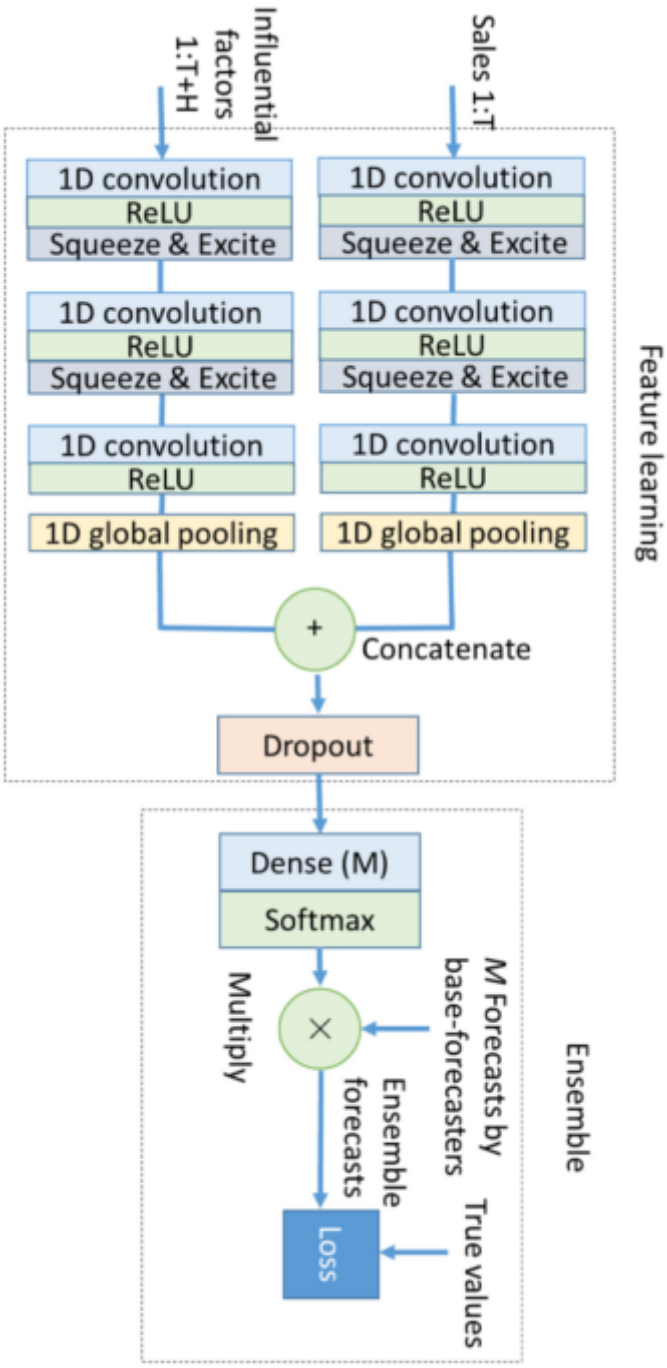
$$\tilde{\mathbf{u}}_k = \frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_{k,t}.$$

$$\mathbf{q} = \sigma\left(\mathbf{w}_2 \mathrm{ReLU}\left(\mathbf{w}_1 \tilde{\mathbf{U}}\right)\right),$$

上式中，$K$ 是 filter （卷积核）个数，$w_1$ 是 $\frac{K}{r} * K$ 维，降维参数；$w_2$ 是 $K * \frac{K}{r}$ 维，升维参数；$r$ 是降维比。最终块的输出缩放为

$$\tilde{\mathbf{u}}_k = q_k \cdot \mathbf{u}_k,$$

之后经全局池化以减少参数维度，利用Dropout减轻过拟合，最后一个Dense层用于学习权重，softmax进行权重分配。




组合时，利用得到的 $h$ 期预测与网络学到的权重进行加权组合。权重和为1。

$$\hat{y}_{i,T+h} = \sum_m w_i^{(m)} \hat{y}_{i,T+h}^{(m)},$$

学习权重时，损失函数的定义如下，其是一种标准化的均方误差。

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{H}\sum_{h=1}^{H}\left(\hat{y}_{i,T+h}(\theta) - y_{i,T+h}\right)^2}{S_i},$$

$$S_i = \frac{1}{MH}\sum_{m=1}^{M}\sum_{h=1}^{H}\left(\hat{y}_{i,T+h}^{(m)} - y_{i,T+h}\right)^2.$$

**基准学习器**

- 局部学习器：ETS、自回归分布滞后、ARIMAX、支持向量回归（SVR）、极值学习器（ELM）；其使用1/3个滞后作为输入
- 全局学习器：Pool 下的自回归分布滞后、Pool下的 ELM、随机森林、梯度增强回归树（GBRT）；其使用3/7个滞后作为输入

# 实验设计

本文的实验希望回答以下问题：

- Q1: 本文的元学习器的预测性能与基础预测器的性能相比如何？和简单组合比如何？元学习器是否在某些情况下特别有效，例如在促销时期？
- Q2：这种新颖的元学习器的性能与 FFORMA 元学习器的性能相比如何？
- Q3：所提出的监督特征学习方法的预测性能与常用的手动选择特征的性能相比如何？
- Q4：除了历史销售时间序列之外，从潜在影响因素中提取特征是否有益？
- Q5：使用局部方法与全局方法基础预测进行组合是否有好处？
- Q6：使用元学习器寻找最佳整体预测而不是寻找最佳个体预测器是否有益？

数据：IRI数据集(Bronnenberg, Kruger, & Mela, 2008)；包括来自 50 个市场和 30 个类别的商店样本的杂货和药品**周度**数据。数据长度统一为153周。

数据划分：使用宽度为 55 周的固定滚动窗口进行估计和预测，每 7 周将窗口向前移动，故在 153 周的数据样本中生成 15 个数据槽。训练与测试都是针对每个数据槽而言的。前10个槽用于元学习器训练，后5个槽用于测试性能。每个数据槽，前48周数据用于训练，而后7周用于评估基础预测。（对间断性的回避：实验都排除了销售中断的 SKU，因为不知道丢失的销售是由于缺货还是间歇需求）

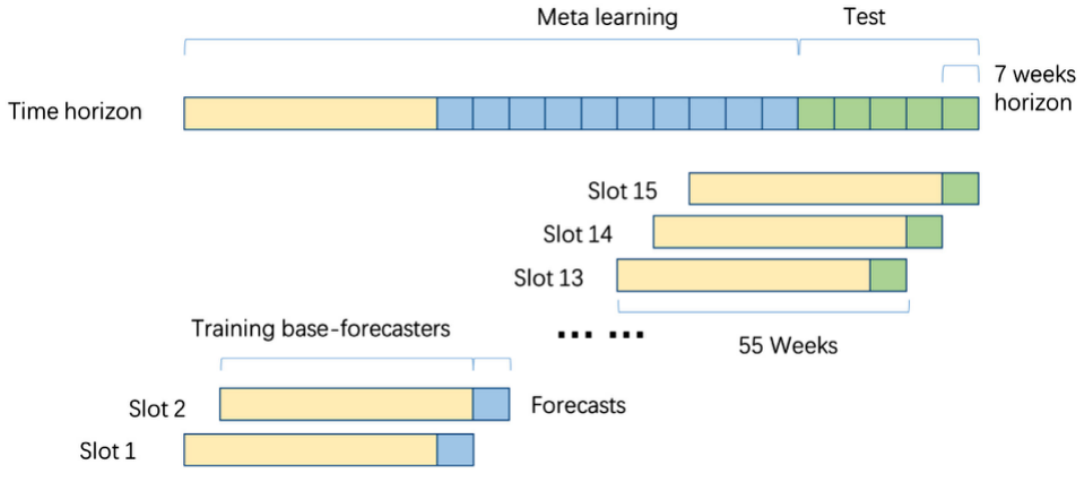



**Fig. 3.** Data manipulation to generate training and test data.

对比实验：除了以上介绍的元学习方法（M0）外，还与以下元学习方法进行比较：

- 无监督选择特征的元学习器（M1）：使用 `tsfeatures` 包的27个特征进行元学习，元学习器是如下所示的全连接网络：

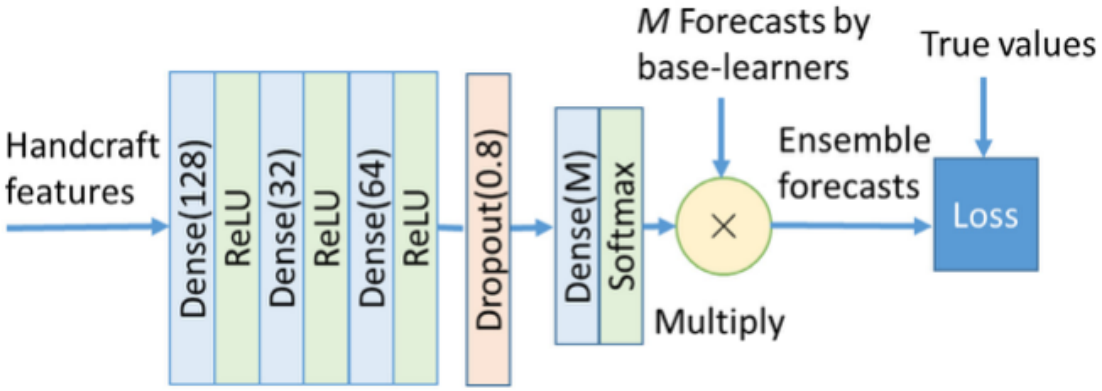



**Fig. 4.** The network structure of meta-learner using hand-selected features.

- 没有从外部因素学习特征的元学习器（M2）：减少了M0的一个外部因素输入通道，其它网络结构不变。
- 仅使用局部基准学习器的元学习器（M3）：只使用局部学习器的组合，其它因素不变。
- 仅使用全局基准学习器的元学习器（M4）：只使用全局学习器的组合，其它因素不变。
- 以选择最优基准模型为目标的元学习器（M5）：损失函数为分类交叉熵，最后采用最优的一个模型，其它因素不变。分类标签的制定是通过基准预测在预测范围（七周内）的平均绝对误差进行评估的。
- 使用预测基准模型绝对误差的元学习器（M6）： Cerqueira et al. (2017) 使用了这个元学习器。其定义损失函数为：

$$L(\theta) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \left( \hat{e}_i^{(m)} - e_i^{(m)} \right)^2,$$

其中 $e_i^{(m)}$ 是第 $m$ 个基准学习器在第 $i$ 个 SKU 的预测误差，$\hat{e}_i^{(m)}$ 是元学习器预测第 $m$ 个基准学习器在第 $i$ 个 SKU 的预测误差。权　重的分配基于 $\hat{e}_i^{(m)}$ 的 softmax 变换进行。

- FFORMA：预测池与损失函数同 M0，尝试了使用仅从销售数据中提取的特征以及从加入外生影响因素提取的特征。

此外，还尝试了三种组合预测的比较基准：

- 简单平均（E1）
- 加权平均，根据训练数据误差表现的 softmax 变换赋权（E2）
- 前4个训练集最优模型的简单平均（E3）

（作者声称 OLS 赋权与受约束回归的赋权他们也尝试了，但是效果不佳，故未展示细节与结果）？

为了比较模型的预测性能，使用三个指标：sMAPE（对称平均绝对百分比误差）、MPE（平均百分比误差）（这个误差会出现正负预测结果相抵的情况，可能是考虑到可以用之前的存货抵消现在的缺货才使用）、平均相对平均绝对误差（AvgRelMAE）定义为：

$$\text{AvgRelMAE} = \left( \prod_{i=1}^{N} \frac{\sum_{h=T+1}^{T+H} \left| \hat{y}_{ih} - y_{ih} \right|}{\sum_{h=T+1}^{T+H} \left| \hat{y}_{ih}^{0} - y_{ih} \right|} \right)^{\frac{1}{N}}$$

这个误差需要定义一个基线预测计算（实际，基准预测用的是ETS，组合预测用的是GBRT-7），反映方法相对基准的改进比例。

具体训练设置：所有元学习器的批量大小设置为 4096。选择训练样本的前 8 个插槽用于训练每个模型，其余 2 个插槽在验证集中用于参数调整。50 个 epoch 。元学习器中三个卷积网络块的过滤器分别设置为 64、128 和 64，dropout 率设置为 0.8。此外，作者将 He、Zhang、Ren 和 Sun (2015) 提出的初始化用于所有卷积层。

# 实验结果

基准模型预测：以 sMAPE 与 AvgRelMAE 计，无论是训练集还是测试集，都是 GBRT 或者 RF 排前两名；局部预测的机器学习方法不如统计方法，而全局预测正好相反；虽然ETS平均预测结果不好，但是战胜其它所有预测模型的次数最多。最终选用17个基准预测（8*2+ETS）中最好的9个模型进行元学习。

**Table 6**
Forecasting performance of base-forecasters in training set.

| Base forecaster | Horizon | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | h=1 | | h=4 | | h=7 | | h=1-7 | | |
| | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | Bias adj. |
| ETS | 19.367 | 1.000 | 20.366 | 1.000 | 20.859 | 1.000 | 20.137 | 1.000 | 1.043 |
| ADL-1 | 16.717 | 0.871 | 17.516 | 0.862 | 17.672 | 0.840 | 17.200 | 0.855 | 1.017 |
| ADL-3 | 16.898 | 0.878 | 17.724 | 0.870 | 17.944 | 0.854 | 17.417 | 0.864 | 1.019 |
| ARX-1 | 17.198 | 0.897 | 17.976 | 0.885 | 18.210 | 0.874 | 17.716 | 0.884 | 0.997 |
| ARX-3 | 18.074 | 0.941 | 18.828 | 0.930 | 18.982 | 0.911 | 18.529 | 0.922 | 0.987 |
| ELM-1 | 18.142 | 0.952 | 19.441 | 0.981 | 19.356 | 0.931 | 18.902 | 0.944 | 0.980 |
| ELM-3 | 19.705 | 1.043 | 20.679 | 1.037 | 20.770 | 0.999 | 20.337 | 1.026 | 1.015 |
| SVM-1 | 17.164 | 0.912 | 17.812 | 0.897 | 17.915 | 0.870 | 17.534 | 0.886 | 1.001 |
| SVM-3 | 17.509 | 0.926 | 18.213 | 0.910 | 18.427 | 0.895 | 17.964 | 0.908 | 1.012 |
| GBRT-3 | **16.231** | 0.844 | 17.097 | 0.845 | **17.304** | 0.831 | 16.785 | 0.841 | 1.029 |
| GBRT-7 | **16.144** | **0.842** | **16.930** | **0.838** | **17.320** | 0.831 | **16.709** | **0.839** | 1.021 |
| ADLP-3 | 16.727 | 0.876 | 17.569 | 0.878 | 17.675 | 0.853 | 17.230 | 0.869 | 1.033 |
| ADLP-7 | 16.593 | 0.869 | 17.414 | 0.869 | 17.667 | 0.853 | 17.139 | 0.865 | 1.031 |
| RF-3 | 16.304 | **0.842** | 17.108 | 0.842 | 17.328 | **0.826** | 16.815 | 0.839 | 1.036 |
| RF-7 | 16.235 | **0.837** | **16.985** | **0.834** | 17.365 | **0.828** | **16.762** | **0.835** | 1.040 |
| ELMP-3 | 16.614 | 0.866 | 17.532 | 0.875 | 17.638 | 0.853 | 17.173 | 0.867 | 1.033 |
| ELMP-7 | 16.496 | 0.858 | 17.345 | 0.863 | 17.670 | 0.851 | 17.090 | 0.862 | 1.035 |

The top two performed models are shown in bold: The ETS forecasts are used as the baseline for calculating AvgRelMAE.

**Table 7**
Forecasting performance of base-forecasters in test set.

| Base forecaster | Horizon | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | h=1 | | h=4 | | h=7 | | h=1-7 | | |
| | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | MPE |
| ETS | 19.305 | 1.000 | 20.105 | 1.000 | 21.152 | 1.000 | 20.219 | 1.000 | 2.163 |
| ADL-1 | 16.842 | 0.885 | 17.691 | 0.871 | 18.743 | 0.867 | 17.756 | 0.869 | -0.339 |
| ADL-3 | 16.999 | 0.893 | 17.929 | 0.883 | 19.053 | 0.886 | 17.957 | 0.882 | **0.247** |
| ARX-1 | 17.246 | 0.907 | 18.059 | 0.904 | 18.996 | 0.894 | 18.190 | 0.902 | 1.045 |
| ARX-3 | 18.101 | 0.952 | 18.924 | 0.945 | 19.900 | 0.942 | 19.055 | 0.947 | 1.154 |
| ELM-1 | 17.959 | 0.932 | 19.280 | 0.956 | 20.629 | 0.984 | 19.448 | 0.970 | **0.254** |
| ELM-3 | 19.744 | 1.048 | 20.679 | 1.043 | 21.654 | 1.038 | 20.739 | 1.041 | 1.602 |
| SVM-1 | 17.175 | 0.915 | 17.950 | 0.909 | 18.833 | 0.894 | 18.058 | 0.907 | 1.283 |
| SVM-3 | 17.531 | 0.925 | 18.380 | 0.922 | 19.347 | 0.920 | 18.467 | 0.923 | 1.693 |
| GBRT-3 | 16.372 | 0.846 | 17.353 | 0.853 | 18.601 | 0.861 | 17.379 | 0.847 | -1.653 |
| GBRT-7 | **16.201** | **0.844** | **17.137** | **0.842** | 18.597 | 0.870 | **17.301** | **0.847** | -1.558 |
| ADLP-3 | 16.788 | 0.869 | 17.815 | 0.880 | 18.902 | 0.879 | 17.805 | 0.872 | -1.927 |
| ADLP-7 | 16.673 | 0.863 | 17.608 | 0.868 | 18.868 | 0.881 | 17.692 | 0.867 | -1.514 |
| RF-3 | 16.454 | 0.848 | 17.351 | 0.846 | **18.580** | **0.853** | 17.400 | 0.843 | -1.328 |
| RF-7 | **16.318** | **0.836** | **17.186** | **0.840** | **18.554** | **0.856** | **17.293** | **0.837** | -0.798 |
| ELMP-3 | 16.683 | 0.866 | 17.708 | 0.873 | 18.855 | 0.877 | 17.725 | 0.867 | -1.570 |
| ELMP-7 | 16.525 | 0.856 | 17.464 | 0.861 | 18.856 | 0.879 | 17.581 | 0.860 | -0.700 |

The top two performed models shown in bold: The ETS forecasts are used as the baseline for calculating AvgRelMAE.


**Fig. 8.** Proportions of the sales time series for which a particular base-forecaster performs as the best.

元学习与组合预测结果的对比如下：

**Table 8**
Forecasting performance of nine meta-learners and three ensemble benchmarks in the test data.

| Meta-learner | Horizon | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | h=1 | | h=4 | | h=7 | | h=1-7 | | |
| | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | sMAPE | AvgRelMAE | MPE |
| M0 | **15.953** | **0.987** | **16.747** | **0.980** | **17.965** | **0.960** | **16.849** | **0.968** | **-0.170** |
| M1 | 15.980 | 0.989 | 16.765 | 0.981 | 17.972 | 0.964 | 16.865 | 0.970 | -0.046 |
| M2 | 15.980 | 0.989 | 16.771 | 0.983 | 17.981 | 0.963 | 16.870 | 0.970 | -0.034 |
| M3 | 16.183 | 0.997 | 17.114 | 1.001 | 18.514 | 0.993 | 17.231 | 0.995 | -1.117 |
| M4 | 16.140 | 0.999 | 16.950 | 0.992 | 18.125 | 0.971 | 17.053 | 0.982 | 0.394 |
| M5 | 17.067 | 1.058 | 18.073 | 1.069 | 19.203 | 1.035 | 18.135 | 1.050 | 3.886 |
| M6 | 16.128 | 1.002 | 16.939 | 0.994 | 18.050 | 0.965 | 17.006 | 0.979 | -0.077 |
| E1 | 16.171 | 1.004 | 16.985 | 0.997 | 18.161 | 0.977 | 17.078 | 0.986 | -0.064 |
| E2 | 16.103 | 1.001 | 16.900 | 0.990 | 18.079 | 0.968 | 16.994 | 0.978 | -0.292 |
| E3 | 16.237 | 1.001 | 17.169 | 1.006 | 18.501 | 0.992 | 17.247 | 0.996 | -1.334 |
| FFORMA1 | 16.060 | 0.992 | 16.868 | 0.985 | 18.110 | 0.969 | 16.975 | 0.977 | -0.060 |
| FFORMA2 | 16.023 | 0.992 | 16.824 | 0.983 | 18.049 | 0.965 | 16.928 | 0.974 | 0.254 |

The best performing methods are shown in bold: The GBRT-7 forecasts are used as the baseline for calculating AvgRelMAE

以基准模型中表现最好的GBRT-7为基准，除M5意外的组合都战胜了基准，显示了模型平均好于模型选择的意义；本文提出的方法M0战胜了所有组合；M3与M4仅包含局部/全局预测之一，效果不佳；由M0-M2与FFORMA1-FFORMA2的对比可以看出，从外部因素学习特征有意义，但改进不是特别大。

下图是方法对比的假设检验结果。M0-M2的差距不显著，而相比其它方法的差距是显著的。

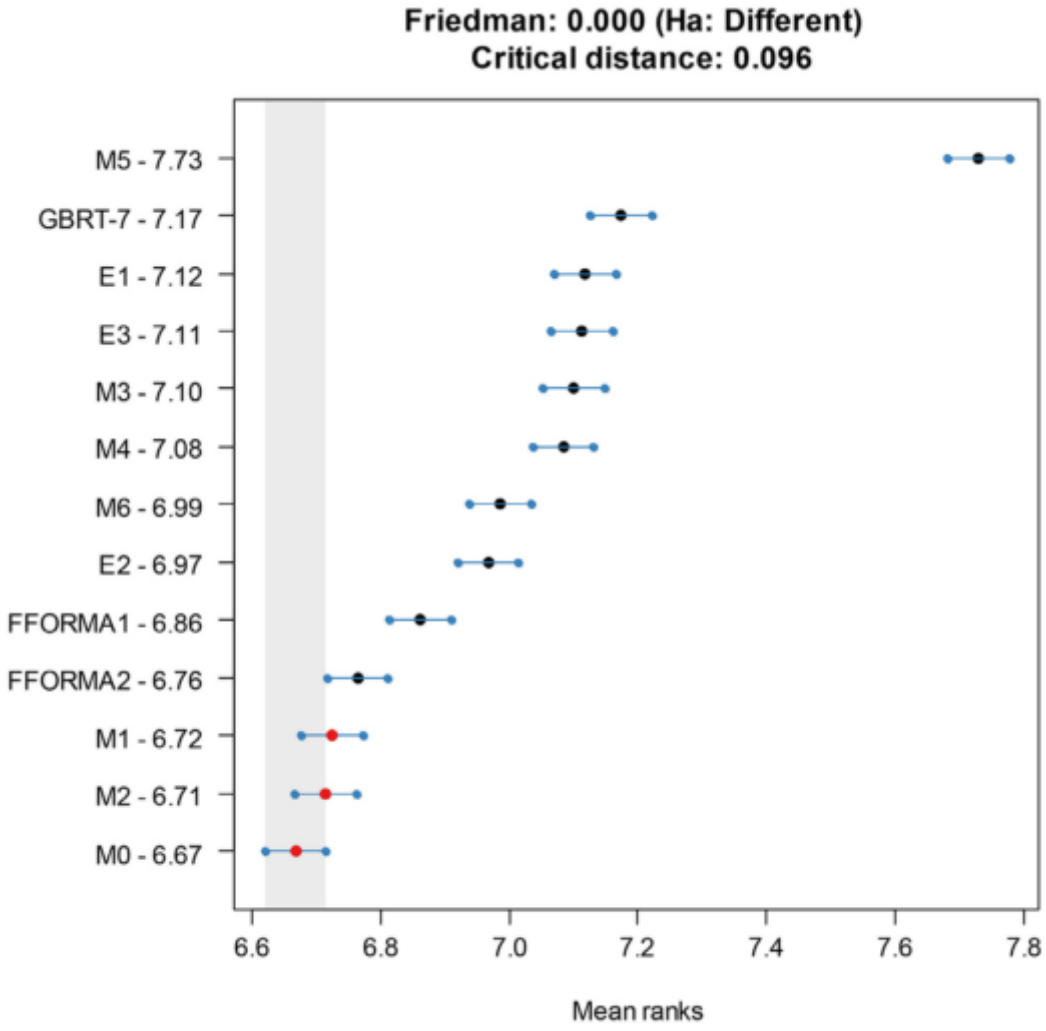



**Fig. 9.** Nemenyi test at 5% significance level on nine meta-learners, three simple combination methods and GBRT-7.

不同基准的权重分配箱线图如下，GBRT-7获得了最多权重。

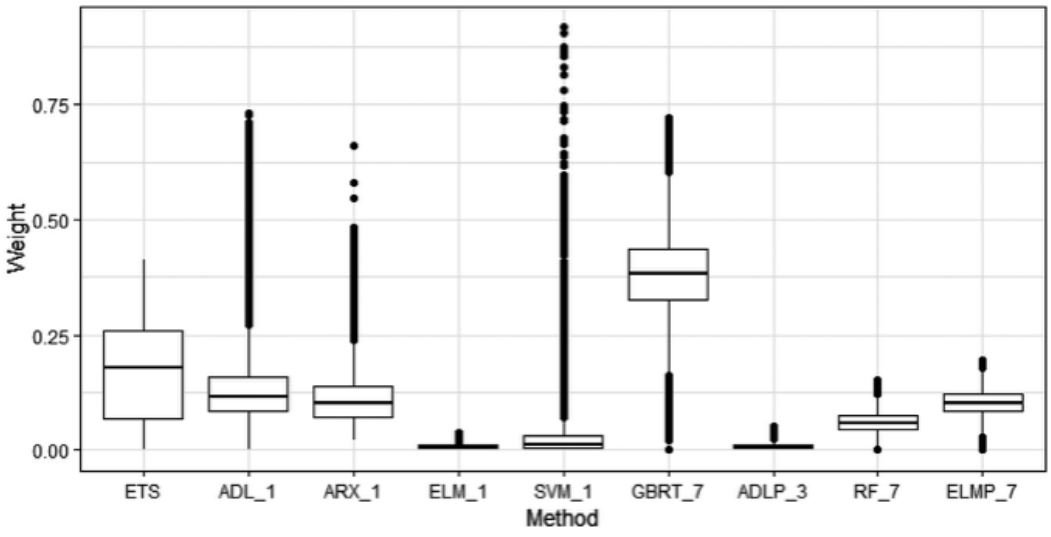



**Fig. 10.** The boxplot of the weights of nine base-forecasters used by M0 when forecasting test periods.

对于促销与非促销时期的表现，文中也做了细分对比，发现M0在两个时期都占优，促销改进5%，非促销改进1.6%；对样本外数据也做了验证，M0依然是最好的组合方法，不过相对GBRT-7的提升受到商品类别的影响。

**Table 9**
Forecasting performance of eight meta-learners and three ensemble benchmarks in promotion and non-promotion periods.

| | Promotion | | Non-promotion | |
|---|---|---|---|---|
| | AvgRelMAE (to ETS) | AvgRelMAE (to GBRT -7) | AvgRelMAE (to ETS) | AvgRelMAE (to GBRT -7) |
| M0 | **0.760** | **0.950** | **0.830** | **0.984** |
| M1 | 0.762 | 0.952 | 0.831 | 0.985 |
| M2 | 0.762 | 0.951 | 0.832 | 0.986 |
| M3 | 0.794 | 0.992 | 0.842 | 0.998 |
| M4 | 0.770 | 0.962 | 0.843 | 0.999 |
| M5 | 0.836 | 1.044 | 0.890 | 1.055 |
| M6 | 0.772 | 0.964 | 0.838 | 0.993 |
| E1 | 0.780 | 0.974 | 0.845 | 1.001 |
| E2 | 0.770 | 0.962 | 0.838 | 0.993 |
| E3 | 0.795 | 0.994 | 0.842 | 0.999 |
| FFORMA1 | 0.771 | 0.963 | 0.834 | 0.989 |
| FFORMA2 | 0.770 | 0.962 | 0.833 | 0.987 |

The best performing method is shown in bold.

**Table 10**
Forecasting performance of eight meta-learners and three ensemble benchmarks for existing and new SKUs.

| | Existing SKUs | | New SKUs | |
|---|---|---|---|---|
| | AvgRelMAE (to ETS) | AvgRelMAE (to GBRT -7) | AvgRelMAE (to ETS) | AvgRelMAE (to GBRT -7) |
| M0 | **0.814** | **0.966** | **0.843** | **0.975** |
| M1 | 0.816 | 0.968 | 0.845 | 0.977 |
| M2 | 0.816 | 0.968 | 0.845 | 0.977 |
| M3 | 0.839 | 0.995 | 0.861 | 0.995 |
| M4 | 0.826 | 0.980 | 0.856 | 0.990 |
| M5 | 0.886 | 1.051 | 0.904 | 1.045 |
| M6 | 0.824 | 0.978 | 0.853 | 0.986 |
| E1 | 0.829 | 0.984 | 0.859 | 0.992 |
| E2 | 0.823 | 0.976 | 0.853 | 0.986 |
| E3 | 0.839 | 0.996 | 0.861 | 0.995 |
| FFORMA1 | 0.822 | 0.975 | 0.851 | 0.983 |
| FFORMA2 | 0.820 | 0.973 | 0.848 | 0.980 |

New SKUs here refer to the SKUs that are sold in a store in the test periods but are not sold in the same store in the training periods.

**Table 11**
Forecsting performance of the eight meta-learners and three ensemble benchmarks over six categories (evaluated with AvgRelMAE, the GBRT-7 forecasts are used as the baseline).

| | Milk | Beer | Mayo | Coffee | Yogurt | Laundet |
|---|---|---|---|---|---|---|
| M0 | 0.947 | 0.986 | 0.970 | 0.975 | 0.967 | 0.964 |
| M1 | 0.949 | 0.986 | 0.973 | 0.976 | 0.969 | 0.967 |
| M2 | 0.947 | 0.988 | 0.972 | 0.976 | 0.969 | 0.967 |
| M3 | 0.994 | 0.998 | 0.985 | 0.999 | 0.995 | 0.997 |
| M4 | 0.954 | 0.995 | 0.983 | 0.990 | 0.986 | 0.980 |
| M5 | 1.017 | 1.062 | 1.051 | 1.034 | 1.063 | 1.038 |
| M6 | 0.956 | 0.991 | 0.979 | 0.984 | 0.983 | 0.978 |
| E1 | 0.958 | 0.996 | 0.986 | 1.007 | 0.987 | 1.003 |
| E2 | 0.959 | 0.991 | 0.978 | 0.988 | 0.977 | 0.979 |
| E3 | 0.995 | 0.997 | 0.987 | 1.001 | 0.996 | 1.000 |
| FFORMA1 | 0.957 | 0.989 | 0.977 | 0.981 | 0.979 | 0.973 |
| FFORMA2 | 0.953 | 0.988 | 0.970 | 0.977 | 0.977 | 0.971 |

关于网络所学习特征的可解释性，作者试图做所提取的特征“活跃的部分”以及卷积提取特征与 `tsfeatures` 特征的相关关系图，没有明显的结论。作者认为仅过一层卷积层还可以解释一些过滤器，而三层之后不易解释。

作者的结论：

- M0具有更优越的性能，特别是在促销周的效果更好，因为此期间销量更不稳定，不易预测；
- M0比FFORMA要好；
- 使用有监督特征学习 (M0) 的元学习器始终比使用无监督手选特征 (M1) 的元学习器表现更好，虽然检验不显著，但是免去了选择特征的麻烦；
- 从销售时间序列和影响因素中学习特征的元学习器可以潜在地提高仅使用销售时间序列作为输入的元学习器的预测性能，尽管效果不显著；解释是销售量蕴含了外生变量的部分信息；
- 使用两种建模策略（局部/全局）的混合模型可以提高预测性能，与M4获胜者结论一致；
- 建议使用元学习器的组合而非选择。

# 个人感想

- 特征：通过卷积网络的方式提取特征，使用有监督的方法进行适应问题的特征训练，有利于元学习，同时牺牲了特征的可解释性；但时间序列特征（即使是人工挑选的）解释也需要专业知识；如果采用高度比较的特征提取/选择（类似于catch22），其实解释起来也很困难。——机器识别特征与人识别特征思路是不同的。
- 研究回避了间断数据的问题；对于 OLS 组合，研究认为其结果比较糟糕——和目前我的研究有些结论冲突；可能是简单使用OLS，未考虑面板数据、多重共线性等问题。（但是对于间断数据，目前我的工作仅在基准方法选择上体现间断数据特性，并未在元学习上体现间断需求的特点）
- 时变权重问题：将如上的网络改成LSTM等类似的允许时变结果输出的网络，或许可以利用时变特征得到时变权重；此外，还可以考虑输入局部数据获得局部特征，如输入滞后28期的数据来进行局部特征的学习——短期数据通道？。
- 局部特征：除了利用网络学习，还有一些别的方法，如 ***Optimal Forecast Combination Based on Neural Networks for Time Series Forecasting*** (Wang et al., 2018)，其利用滞后四期的值作为特征进行聚类，构造局部特征。
- 预测的损失函数：如果设置为组合预测的损失，可能更准确，以及支持加入截距的扩展？