# 时序特征说明

## 符号说明

(1) $t_1, t_2, \cdots, t_K$ 表示时间序列，设共有 $T(30 \le T \le 31)$ 个时刻，则 $K = T$

(2) $\{S_{t_1}, S_{t_2}, \cdots, S_{t_K}\}$ 表示业务量集合，其中，$S_{t_k}$ 表示 $t_k$ 时间的业务量

(3) $\{F_1, F_2, \cdots, F_N\}$ 表示时序特征集合，其中，$F_n$ 表示第 $n$ 个特征

## 特征说明

| 编号 | 数学表达 | 注释 |
|---|---|---|
| 对集合 $\{S_{t_1}, S_{t_2}, \cdots, S_{t_K}\}$, 设 $K = 30$, $epsilon = 1e - 10$ |||
| 1 | $F_1 = S_{t_1}$ | firstValue(第一天的值) |
| 2 | $F_2 = S_{t_K}$ | lastValue(最后一天的值) |
| 3 | $F_3 = 0.5 * (S_{t_{15}} + S_{t_{16}})$ | medianValue(所有天的业务量的中位数) |
| 4 | $F_4 = S_{t_{30*0.75}} - S_{t_{30*0.25}}$ | interquartile(排在0.75 位置上的业务量减去排在0.25位置上的业务量的差) |
| 5 | if $\|F_1\| > epsilon$<br>　　then $F_5 = S_{t_k}/S_{t_1} = F_2/F_1$<br>　else<br>　　　if $\|F_2\| > epsilon$ then $F_5 = 2$<br>　　　else $F_5 = 1$ | tailHeadRatio(最后一天的业务量/第一天的业务量) |
| 6 | $F_6 = (S_{t_K} - S_{t_1})/30 = (F_2 - F_1)/30$ | avgIncrease(平均涨幅) |
| 7 | $F_7 = max(\{S_{t_k}\|k \in [1, 30]\})$ | maxValue(最大值) |
| 8 | $F_8 = min(\{S_{t_k}\|k \in [1, 30]\})$ | minValue(最小值) |
| 9 | $F_9 = \sum_{k=1}^{K} S_{t_k}/30$ | average(平均值) |
| 10 | $F_{10} = \sum_{k=1}^{K} S_{t_k}^2/30 - F_9 * F_9$ | variance(方差) |
| 11 | if $\|F_9\| < epsilon$ then $F_{11} = F_{10}$<br>else $F_{11} = F_{10}/F_9$ | varianceRatio(方差/平均值) |
| 12 | if $\sum_{k=1}^{K} S_{t_k} * w_k > epsilon$<br>　　then $F_{12} = \frac{\sum_{k=K}^{1} S_{t_k} * w_k}{\sum_{k=1}^{K} S_{t_k} * w_k}$<br>　else $F_{12} = 2$<br>if $\sum_{k=1}^{K} S_{t_k} * w_k \le epsilon$<br>　and<br>　　$\sum_{k=K}^{1} S_{t_k} * w_k \le epsilon$<br>　then $F_{12} = 1$<br><br>其中,$w_k = 0.85 * w_{k-1}, w_1 = 1$ | averageSkew(nlp特征，从最近的一天往前对元素做weighted average，距离越近权重越大) |
| 13 | $F_{13} = \sum_{k=1}^{7} S_{t_k}$ | first7Sum(前面7天的业务量总和) |
| 14 | $F_{14} = \sum_{k=24}^{30} S_{t_k}$ | last7Sum(最后7天的业务量总和) |

| 编号 | 数学表达 | 注释 |
|---|---|---|
| 15 | if $\|\sum_{k=1}^{7} S_{t_k}$ (即 $F_{13}$)$\| > epsilon$<br>　then $F_{15} = \frac{\sum_{k=24}^{30} S_{t_k}}{\sum_{k=1}^{7} S_{t_k}}$<br>　else<br>　　if $\|\sum_{k=24}^{30} S_{t_k}\| > epsilon$<br>　　　then $F_{15} = 2$<br>　　　else $F_{15} = 1$ | tailHead7SumRatio(last7Sum/first7Sum) |
| 16 | $F_{16} = (F_{14} - F_{13})/30$ | tailHead7SumAvgIncrease((last7Sum-first7Sum)/30) |
| 17 | if $\|F_9\| > epsilon$<br>　then $F_{17} = F_{14}/7/F_9$<br>　else<br>　　if $\|F_{14}\| > epsilon$<br>　　　then $F_{17} = 2$ | last7AvgToOverallAvg(last7Sum/7/average) |
| 18 | $F_{18} = \sum(\|S_{t_k}\| < epsilon)$ | zeroCount(为0的天数) |
| 19 | $F_{19} = \{first\|\{\|S_{t_k}\| \geq epsilon\}_{index}\}$ | firstNonzeroIndex(业务量中第一个不为0的下标) |
| 20 | $F_{20} = \{last\|\{\|S_{t_k}\| \geq epsilon\}_{index}\}$ | lastNonzeroIndex(业务量中最后一个不为0的下标) |
| 21 | $F_{21} = \{S_{t_k}\|\{\|S_{t_k}\| \geq epsilon\}_{first}\}$ | firstNonzeroValue(第一个不为0的业务量) |
| 22 | $F_{22} = \{S_{t_k}\|\{\|S_{t_k}\| \geq epsilon\}_{last}\}$ | lastNonzeroValue(最后一个不为0的业务量) |
| 23 | $F_{23} = F_{20} - F_{19} + 1$ | lifeSpan(lastNonzeroIndex - firstNonzeroIndex + 1) |
| 24 | $F_{24} = (F_{22} - F_{21})/F_{23}$ | lifeAvgIncrease(在整个生命周期的平均增长量) |
| 25 | $F_{25} = q - p + 1$<br>其中,$\{S_{(t_k)_p}, \cdots, S_{(t_k)_q}\}$,<br>$[p, q]$为连续区间,$S_{t_k} < epsilon$ | maxZeroLength(连续值为0的天数) |
| 26 | $F_{26} = q - p + 1$<br>其中,$\{S_{(t_k)_p}, \cdots, S_{(t_k)_q}\}$,<br>$[p, q]$为连续区间,$S_{t_k} > epsilon$ | maxNonzeroLength(连续值不为0的天数) |
| 27 | $F_{27} = \frac{\sum_{k=1}^{30}(S_{t_k} - \frac{\sum_{k=1}^{30} S_{t_k}}{30})*(k-(0.5*(1+30)))}{\sum_{k=1}^{30}(k-(0.5*(1+30)))*(k-(0.5*(1+30)))}$ | slope(斜率，最小二乘法可以算拟合的直线的斜率) |
| 28 | $F_{28} = argmax(S_{t_k} - S_{t_{(k-1)}}, t \in [2, 30])$ | maxJump(相邻两天的上升最大值) |
| 29 | $F_{29} = argmax(S_{t_{(k-1)}-S_{t_k}}, t \in [2, 30])$ | maxFall(相邻两天的下降最大值) |
| 30 | $F_{30} = \sum_{k=2}^{30}\|S_{t_k} - S_{t_{(k-1)}}\|/29$ | diffAvg(所有相邻的差异的平均) |
| 31 | $F_{31} = \sum(S_{t_{k+1}} - S_{t_k})/m$<br>其中,$S_{t_{k+1}} > S_{t_k}$,<br>$m$为满足$S_{t_{k+1}} > S_{t_k}$的天数总和 | avgJump(平均上升值) |
| 32 | $F_{32} = \sum(S_{t_k} - S_{t_{k+1}})/n$<br>其中,$S_{t_k} > S_{t_{k+1}}$,<br>$n$为满足$S_{t_k} > S_{t_{k+1}}$的天数总和 | avgFall(平均下降值) |

| 编号 | 数学表达 | 注释 |
|------|---------|------|
| | 对集合$\{S_{t_{10}}, S_{t_{11}}, \cdots, S_{t_{30}}\}$，即取出时间序列中的最后21个时刻，$epsilon = 1e-10$ | |
| 33 | $F_{33} = \frac{\sum_{k=1}^{21}(S_{t_k} - \frac{\sum_{k=1}^{21} S_{t_k}}{21})*(k-(0.5*(1+21)))}{\sum_{k=1}^{21}(k-(0.5*(1+21)))*(k-(0.5*(1+21)))}$ | slopeLast21(最后21天的斜率) |
| 34 | $F_{34} = \frac{\sum_{k=1}^{21}(S'_{t_k} - \frac{\sum_{k=1}^{21} S'_{t_k}}{21})*(k-(0.5*(1+21)))}{\sum_{k=1}^{21}(k-(0.5*(1+21)))*(k-(0.5*(1+21)))}$ <br> 其中$S' = \{\frac{S_{t_k}}{\max \text{ of } \{S_{t_{10}}, S_{t_{11}}, \cdots, S_{t_{30}}\}}\}$ | slopeLast21Norm(slopeLast21/slope) |
| 35 | $F_{35} = \sum_{k=1}^{21} S_{t_k}/21$ | averageLast21(最后21天的平均值) |
| 36 | $F_{36} = \sum_{k=1}^{7} S_{t_k}$ | first7SumLast21(最后21天中的前面7天的业务量总和) |
| 37 | if $\lvert \sum_{k=1}^{7} S_{t_k}(即 F_{36})\rvert > epsilon$ <br>    then $F_{37} = \frac{\sum_{k=15}^{21} S_{t_k}}{\sum_{k=1}^{7} S_{t_k}}$ <br> else <br>      if $\lvert \sum_{k=15}^{21} S_{t_k}\rvert > epsilon$ <br>        then $F_{37} = 2$ <br>        else $F_{37} = 1$ | tailHead7SumRatioLast21 |
| 38 | $F_{38} = (\sum_{k=15}^{21} S_{t_k} - \sum_{k=1}^{7} S_{t_k})/21$ | tailHead7SumAvgIncreaseLast21(last7SumLast21-first7SumLast21)/21 |
| 39 | if $\lvert F_{35}\rvert > epsilon$ <br>    then $F_{39} = \sum_{k=15}^{21} S_{t_k}/7/F_{35}$ <br> else <br>      if $\lvert \sum_{k=15}^{21} S_{t_k}\rvert > epsilon$ <br>        then $F_{39} = 2$ | last7AvgToOverallAvgLast21(last7Sum/7/average) |
| 40 | $F_{40} = argmax(S_{t_k} - S_{t_{(k-1)}}, t \in [2,21])$ | last21MaxJump(在最后21天算maxJump) |
| 41 | $F_{41} = argmax(S_{t_{(k-1)}} - S_{t_k}, t \in [2,21])$ | last21MaxFall(在最后21天算maxFall) |
| 42 | $F_{42} = \sum_{k=2}^{21} \lvert S_{t_k} - S_{t_{(k-1)}}\rvert/20$ | last21diffAvg(在最后21天算差异的平均值) |
| 43 | $F_{43} = \sum(S_{t_{k+1}} - S_{t_k})/m$ <br> 其中，$S_{t_{k+1}} > S_{t_k}$， <br> $m$为满足$S_{t_{k+1}} > S_{t_k}$的天数总和 | last21avgJump(最后21天的平均上升值) |
| 44 | $F_{44} = \sum(S_{t_k} - S_{t_{k+1}})/n$ <br> 其中，$S_{t_k} > S_{t_{k+1}}$， <br> $n$为满足$S_{t_k} > S_{t_{k+1}}$的天数总和 | last21avgFall(最后21天的平均下降值) |
| | 对集合$S' = \{S'_1, S'_2, S'_3, S'_4\} = \{\sum_{t=1}^{7} S_{t_k}, \sum_{t=8}^{14} S_{t_k}, \sum_{t=15}^{21} S_{t_k}, \sum_{t=16}^{28} S_{t_k}\}$ | |
| 45 | $F_{45} = \frac{\sum_{a=1}^{4}(S'_a - \frac{\sum_{a=1}^{4} S'_a}{4})*(a-(0.5*(1+4)))}{\sum_{a=1}^{4}(a-(0.5*(1+4)))*(k-(0.5*(1+4)))}$ | slopeWeekly(每周的斜率的平均值) |
| 46 | $F_{46} = argmax(S'_x - S'_{x-1}, x \in [2,4])$ | weeklyMaxJump(每周算一个maxJump,然后算平均) |
| 47 | $F_{47} = argmax(S'_{x-1} - S'_x, x \in [2,4])$ | weeklyMaxFall(每周算一个maxFall,然后算平均) |
| 48 | $F_{48} = \sum_{x=2}^{4} \lvert S'_x - S_{x-1}\rvert/3$ | weeklydiffAvg(每周算一个diffAvg, 然后算平均) |
| 49 | $F_{49} = \sum(S'_{x+1} - S_{\cdot x})/m$ <br> 其中，$S'_{x+1} > S_{\cdot x}$， <br> $m$为满足$S'_{x+1} > S_{\cdot x}$的天数总和 | weeklyavgJump(每周的平均上升值) |

| 编号 | 数学表达 | 注释 |
|---|---|---|
| 50 | $F_{50} = \sum(S'_x - S'_{x+1})/n$ <br> 其中,$S'_x > S'_{x+1}$, <br> $n$为满足$S'_x > S'_{x+1}$的天数总和 | weeklyavgFall(每周的平均下降值) |
| 熵特征($\{S_{t_1}, S_{t_2}, \cdots, S_{t_K}\}$) | | |
| 51 | $F_{51} = -\sum_{k=1}^{30} \frac{S_{t_k}}{\sum_{k=1}^{30} S_{t_k}} \log \frac{S_{t_k}}{\sum_{k=1}^{30} S_{t_k}}$ | entropy(熵) |