3CS Algorithm for Efficient Gaussian Process Model Retrieval

Fabian Berns[1], Kjeld Schmidt[1], Ingolf Bracht[1], Christian Beecks[1,2]

[1]University of Münster
Department of Computer Science -- Data Management and Analytics Group

[2]Fraunhofer Institute for Applied Information Technology FIT

Introduction (I)

- Gaussian Process Models (GPM) have been widely applied for various pattern recognition tasks
- We identified three kinds of GPMs

  i. **Default instantiation**

  ii. **Domain-specific instantiations**

  iii. **Automatically retrieved GPM given arbitrary data**

- GPM Evaluation and Application usually suffers from $\mathcal{O}(n^3)$ complexity

Automatic GPM Retrieval – Formalization

- Gaussian Process: $f \sim GP(m, k)$
- Marginalizing GP_θ for certain data D yields:

$$\mathcal{L}(m, k, \theta | D) = \frac{1}{2} \cdot [(y - \mu)^T K^{-1} (y - \mu) + \log|K| + n2\pi],$$

- Goal: find best $GP_\theta^* \in \mathbb{G}$ for D

$$\mathbb{G} = \{GP_\theta(m, k) | m \in \mathbb{R}^{\mathcal{X}}, k \in \mathbb{R}^{\mathcal{X} \times \mathcal{X}} \in \mathbb{R}, \theta \in \Theta\}$$

- Automatic GPM Retrieval is defined as follows:

$$GP_\theta^* = \mathrm{argmax}_{GP_\theta^*(m,k) \in \mathcal{G}} \mathcal{L}(m, k, \theta | D), \mathcal{G} \subset \mathbb{G}$$

Concatenated Composite Covariance Search (3CS)

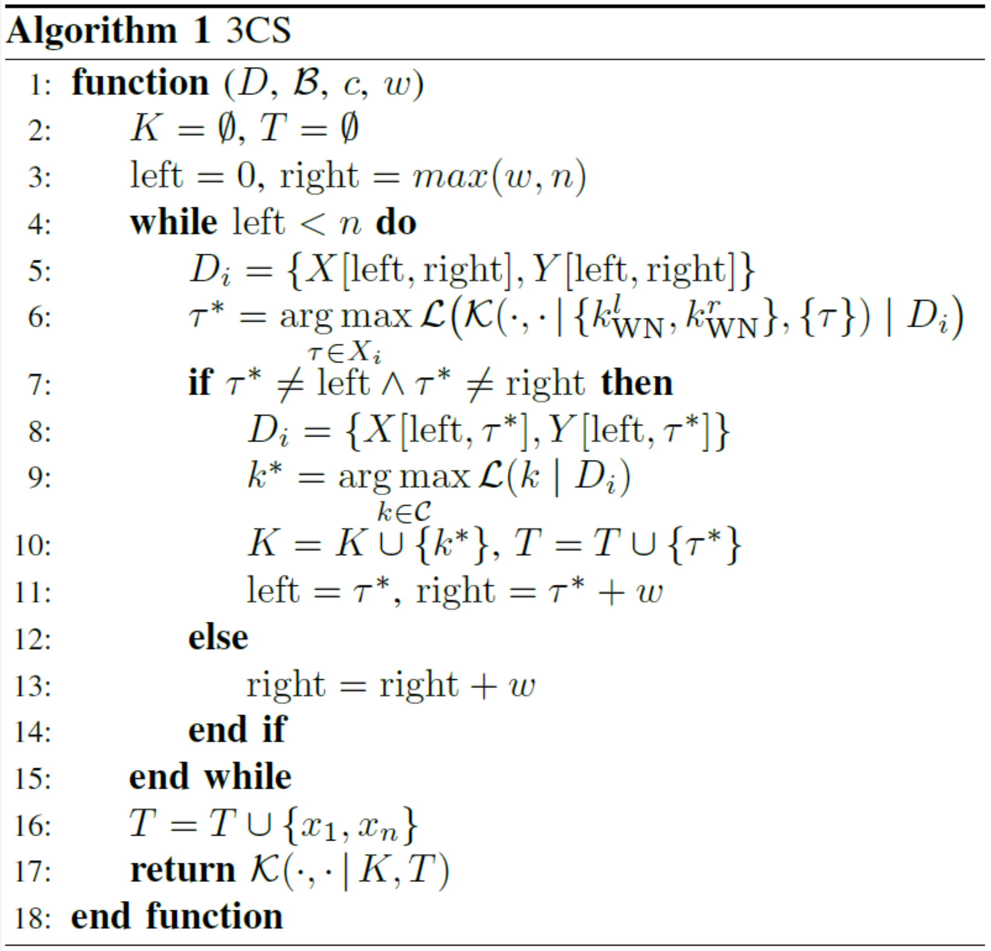
Covariance function is partitioned by means of change points $\mathrm{T} = \{\tau_i\}_{i=1}^{a}$:

$$\mathcal{K}(x, x' | \{k_i\}_{i=1}^{a}, T) = \sum_{i=1}^{a} k_i(x, x') \cdot 1_{\tau_{i-1} < x \leq \tau_i}(x) \cdot 1_{\tau_{i-1} < x' \leq \tau_i}(x')$$

We define the virtual search space as follows:

$$\mathcal{G}_T = \{GP_\theta(m, k) | m \in 0^{\mathcal{X}}, k = \mathcal{K}(\cdot,\cdot\, | \{k_i | k_i \in \mathcal{S}\}_{i=1}^{a}, T), \theta \in \Theta\}$$

$$\mathcal{S} = \left\{ \sum \prod b \,| b \in \mathcal{B} \right\}$$



Algorithm 1 3CS

```
function ($D$, $\mathcal{B}$, $c$, $w$)
    $K = \emptyset$, $T = \emptyset$
    left $= 0$, right $= max(w, n)$
    while left $< n$ do
        $D_i = \{X[\text{left}, \text{right}], Y[\text{left}, \text{right}]\}$
        $\tau^* = \arg\max_{\tau \in X_i} \mathcal{L}(\mathcal{K}(\cdot, \cdot \mid \{k^l_{\text{WN}}, k^r_{\text{WN}}\}, \{\tau\}) \mid D_i)$
        if $\tau^* \neq$ left $\wedge\ \tau^* \neq$ right then
            $D_i = \{X[\text{left}, \tau^*], Y[\text{left}, \tau^*]\}$
            $k^* = \arg\max_{k \in \mathcal{C}} \mathcal{L}(k \mid D_i)$
            $K = K \cup \{k^*\}$, $T = T \cup \{\tau^*\}$
            left $= \tau^*$, right $= \tau^* + w$
        else
            right $=$ right $+ w$
        end if
    end while
    $T = T \cup \{x_1, x_n\}$
    return $\mathcal{K}(\cdot, \cdot \mid K, T)$
end function
```

Evaluation

- Eight Benchmark Datasets were used
  - **144 – 2M data records**
- 3CS **outperforms state-of-the-art algorithms** with regards to runtime and model accuracy
- 3CS proves to be **scalable** to large datasets, while maintaining model quality

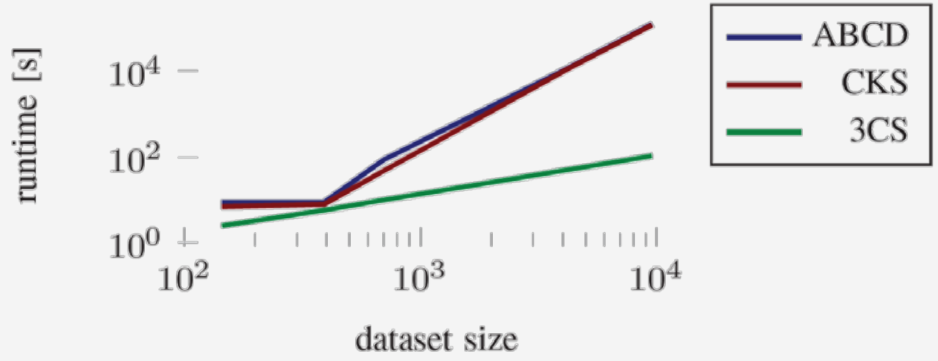
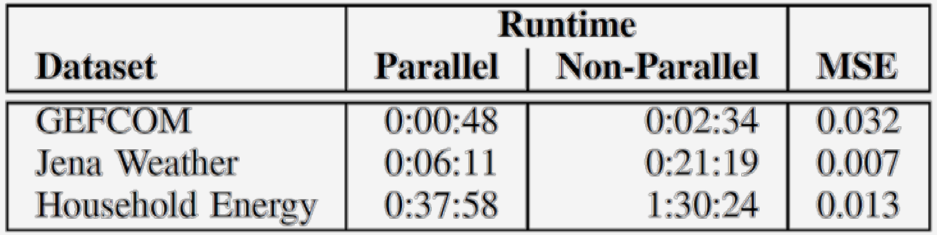


Fig. 4. Runtime for different state-of-the-art algorithms.



Dataset	Runtime Parallel	 Non-Parallel	MSE
GEFCOM	0:00:48	0:02:34	0.032
Jena Weather	0:06:11	0:21:19	0.007
Household Energy	0:37:58	1:30:24	0.013

TABLE III
EFFICIENCY AND ACCURACY OF THE 3CS ALGORITHM ON LARGE-SCALE DATASETS

Conclusions and Future Work

- In this paper:
  - **We proposed the *Concatenated Composite Covariance Search (3CS)* algorithm**
  - **We evaluated that algorithm‘s capabilities by means of eight benchmark datasets and compared its performance to given state-of-the-art methods**
- As future work, we plan to incorporate global approximations into the procedure, develop domain-specific adaptations and include prior knowledge into the retrieval process

Thanks for your attention!

Fabian Berns
Einsteinstraße 62
48149 Münster

Email: fabian.berns@uni-muenster.de
Website: www.uni-muenster.de/dma