Fabian Berns[1], Kjeld Schmidt[1], Ingolf Bracht[1], Christian Beecks[1,2]
[1]University of Münster, [2]Fraunhofer Institute for Applied Information Technology FIT

3CS Algorithm for Efficient Gaussian Process Model Retrieval

Motivation

- Gaussian Process Models (GPM) have been widely applied for various pattern recognition tasks [2]
- **Automatic Retrieval of GPM**, as well as GP Evaluation and Application usually suffers from $\mathcal{O}(n^3)$ complexity [1,2,3]
- 3CS aims to efficiently deliver GPMs for large-scale data, that overcome these complexity constraints

Concatenated Composite Covariance Search (3CS)

- 3CS efficiently retrieves Gaussian Process Models for large-scale data by means of local approximations [1,3]
- Therefore, we use covariance function $\mathcal{K}$, which partitions the given data by means of change points $\mathrm{T} = \{\tau_i\}_{i=1}^{a}$

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

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 [4,5] with regards to runtime and model accuracy (cf. Table II and Figure 4)
- 3CS proves to be scalable to large datasets, while maintaining model quality (cf. Table III)

Dataset	Runtime			MSE		
	CKS	ABCD	3CS	CKS	ABCD	3CS
Airline	0:00:07	0:00:09	**0:00:03**	**0.107**	**0.107**	**0.107**
SolarIr	0:00:08	0:00:09	**0:00:06**	0.087	0.087	**0.076**
MaunaLoa	0:00:49	0:01:28	**0:00:10**	0.153	0.153	**0.017**
SML	2:56:58	3:04:51	**0:00:50**	0.110	0.110	**0.040**
PowerPlant	32:36:42	34:08:01	**0:01:46**	**0.071**	**0.071**	0.102

TABLE II
EFFICIENCY AND ACCURACY OF DIFFERENT GPM RETRIEVAL ALGORITHMS

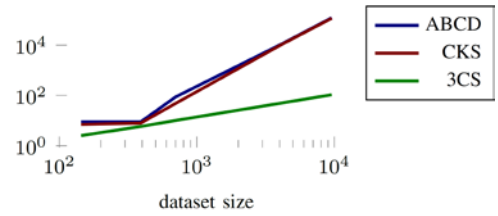

Fig. 4. Runtime for different state-of-the-art algorithms.

Dataset	Runtime		MSE
	Parallel	Non-Parallel	
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

References

[1] Snelson, Edward, and Zoubin Ghahramani. "Local and global sparse Gaussian process approximations." Artificial Intelligence and Statistics. 2007.
[2] Williams, Christopher KI, and Carl Edward Rasmussen. Gaussian processes for machine learning. Vol. 2. No. 3. Cambridge, MA: MIT press, 2006.
[3] Berns, Fabian, and Christian Beecks. "Automatic Gaussian Process Model Retrieval for Big Data." Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.
[4] Duvenaud, David, et al. "Structure discovery in nonparametric regression through compositional kernel search." International Conference on Machine Learning. PMLR, 2013.
[5] Lloyd, James Robert, et al. "Automatic construction and natural-language description of nonparametric regression models." Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.