On Advancement of Information Spaces to Improve Prediction-Based Compression

Presentation of work originally published in the Proc. of the 2018 IEEE International Conference on Big Data

Ugur Cayoglu[1], Frank Tristram[2], Jörg Meyer[1], Tobias Kerzenmacher[3], Peter Braesicke[3], Achim Streit[1]

Abstract: One of the scientific communities that generate the largest amounts of data today are the climate sciences. New climate models enable model integrations at unprecedented resolution, simulating timescales from decades to centuries of climate change. Nowadays, limited storage space and ever increasing model output is a big challenge. For this reason, we look at lossless compression using prediction-based data compression. We show that there is a significant dependence of the compression rate on the chosen traversal method and the underlying data model. We examine the influence of this structural dependency on prediction-based compression algorithms and explore possibilities to improve compression rates. We introduce the concept of Information Spaces (IS), which help to improve the accuracy of predictions by nearly 10% and decrease the standard deviation of the compression results by 20% on average.

Keywords: compression algorithms; encoding; meteorology; prediction-based compression; information spaces

Introduction

New climate models such as ICON-ART [Sc18] make it possible to run high-resolution simulations of the atmosphere and its composition at an unprecedented scale and detail while making full use of the available capacity of high-performance computers. But with these improvements, the storage space required to save the output of the simulations also increases. In such situations an efficient compression method can help reduce the required storage space. In prediction-based compression all data points are processed in a predefined traversal sequence. As the sequence is processed, a prediction is given for each data point based on the values prior in the sequence. The difference between the actual value and its prediction will then be saved on disk. A good prediction leads to a better compression rate. For an extended version of this work please refer to our original publication [Ca18].

[1] Karlsruhe Institute of Technology, Steinbuch Centre for Computing (SCC), Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany, Ugur.Cayoglu@kit.edu

[2] Karlsruhe Institute of Technology, Institute of Applied Physics (APH)

[3] Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research (IMK-ASF)

doi:10.18420/inf2019_39

Method

Our method calculates position and neighbourhood information of each data point s_i during its traversal. We call this the Information Space (IS) of the data point.

Let S_i be all the data points in the data ordered by traversal path. The IS of a data point s_i is the set of data points $s_j \in S_i$ with $j < i$ and each element of the coordinate tuple within a certain range r of s_i.

$$IS(s_i) = \{s_k | \forall s_k \in S_i : a^i_j - r \leq a^k_j \leq a^i_j + r\} \tag{1}$$

with a^n_m defining the coordinate position at dimension m of element n of sequence S. The IS is then divided into its components to isolate the information contained in the various dimensions. We call these components Information Context (IC).

Each IC contains information along several dimensions. ICs can contain overlapping data points, but none is a subset of another. This allows predictions on the basis of information from different dimensions and later merge them into a consolidated prediction.

Evaluation

We analysed the performance of different prediction-based compression algorithms on climate data. The results of our experiments show that changing the starting point of the compression algorithm has only negligible effects on the compression rate, while changing the traversal path can influence the compression rate significantly. Further experiments show that with the help of IS it is possible to improve the predictions of each predictor. More importantly, the stability of the predictions can be increased. This results in higher quality forecasts with less fluctuations than with established methods.

Our current configuration achieved a 10% improvement in prediction accuracy and decreased the standard deviation of the compression results by over 20% on average.

Bibliography

[Ca18] Cayoglu, U.; Tristram, F.; Meyer, J.; Kerzenmacher, T.; Braesicke, P.; Streit, A.: Concept and Analysis of Information Spaces to improve Prediction-Based Compression. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 3392–3401, Dec 2018.

[Sc18] Schröter, J.; Rieger, D.; Stassen, C.; Vogel, H.; Weimer, M.; Werchner, S.; Förstner, J.; Prill, F.; Reinert, D.; Zängl, G.; Giorgetta, M.; Ruhnke, R.; Vogel, B.; Braesicke, P.: ICON-ART 2.1 – A flexible tracer framework and its application for composition studies in numerical weather forecasting and climate simulations. Geoscientific Model Development Discussions, 2018:1–37, 2018.