SCALING SUPPORT VECTOR MACHINES TOWARDS EXASCALE COMPUTING FOR CLASSIFICATION OF LARGE-SCALE HIGH-RESOLUTION REMOTE SENSING IMAGES

Ernir Erlingsson[1], Gabriele Cavallaro[2], Morris Riedel[1,2], Helmut Neukirchen[1]

[1] School of Engineering and Natural Sciences, University of Iceland, Iceland
[2] Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

ABSTRACT

Progress in sensor technology leads to an ever-increasing amount of remote sensing data which needs to be classified in order to extract information. This big amount of data requires parallel processing by running parallel implementations of classification algorithms, such as Support Vector Machines (SVMs), on High-Performance Computing (HPC) clusters. Tomorrow's supercomputers will be able to provide exascale computing performance by using specialised hardware accelerators. However, existing software processing chains need to be adapted to make use of the best fitting accelerators. To address this problem, a mapping of an SVM remote sensing classification chain to the Dynamical Exascale Entry Platform (DEEP), a European pre-exascale platform, is presented. It will allow to scale SVM-based classifications on tomorrow's hardware towards exascale performance.

Index Terms— Remote Sensing, Support Vector Machines (SVMs), High-Performance Computing (HPC), Exascale Computing, Hardware Accelerators

1. INTRODUCTION

One of the challenges in remote sensing during the last decades is to classify land cover into distinct classes based on high resolution datasets acquired from airborne and satellite sensors [1]. These sensors are continuously improving, resulting in an enormous volume, velocity, and variety of sensed data. For example, Copernicus, the world's largest single Earth observation program, with its fleet of Sentinel satellites alone, creates over 10 Petabyte of new data per year (with just Sentinel-1, -2 and -3 fully operational and more to come in the future). This leads to new challenges within the entire lifecycle of remote sensing data [2].

Among the remote sensing classifiers, Support Vector Machines (SVMs) [3] have often been found to be more effective with respect to accuracy and stability. However, SVMs are computationally very demanding. Where in the past serial implementations of classification algorithms were

sufficient to deal with remote sensing data, today's big data requires parallel processing to deal with its huge amount of data and to get results within an acceptable time frame [4]. These parallel processing implementations are not restricted to multi-core implementations running on a single machine with shared memory, but even run on High-Performance Computing (HPC) clusters with distributed memory.

Todays high-end HPC clusters reach a performance in the order of petaflops (10^{15} floating point operations per second). While this performance may be sufficient for today's applications, higher performance is envisaged for the future and thus, the design of exascale computing systems has begun, yielding a thousandfold performance increase. For remote sensing, this will allow for tackling of larger scale problems or to obtain classification results in near real-time.

However, the road towards exascale is paved with obstacles and requires not only specialised HPC hardware accelerators to clear them, but additionally the software needs to be adapted to the new hardware in order to obtain exascale performance. This paper proposes a novel approach to make the classification process of remote sensing data using SVMs ready for exascale computing. To this aim, a new software chain has been developed that maps the different phases of classification via an SVM to the specialised exascale hardware accelerators that are being developed in the European Union's Horizon 2020 exascale project *Dynamical Exascale Entry Platform – Extreme Scale Technologies* (DEEP-EST). To the best of our knowledge, no other work exists on making SVMs ready for exascale to date.

2. BACKGROUND

In this section we provide foundations on classification using SVMs on the DEEP exascale entry platform.

2.1. Support Vector Machines

Generally, in supervised machine learning, sensor data can be classified with decent accuracy using trained models created with Support Vector Machines (SVMs) [3]: First, the raw data is prepared with feature engineering which usually significantly improves the performance of the algorithm. We use the
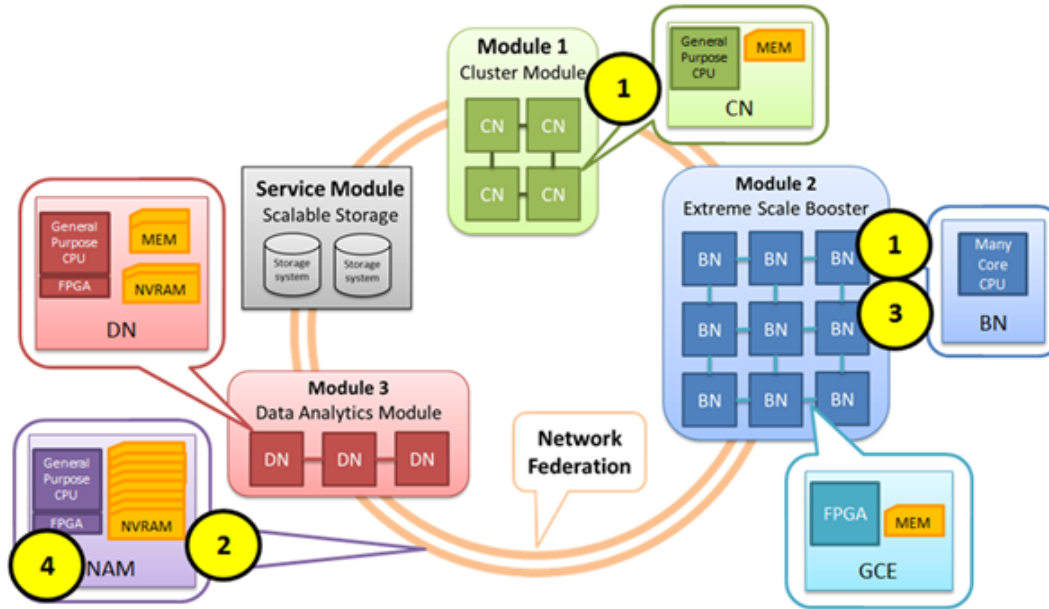
Fig. 1. Utilizing the Modular Supercomputer architecture for cross-validation (consult Sect. 2.2 for module details)

Self-Dual Attribute Profile (SDAP) [5] technique due to its suitability in extracting features from remote sensor data with both high high-spectral and spatial resolutions. The SVM performs non-linear classifications by using kernels which allows it to operate implicitly in a high-dimensional feature space.

When using SVMs in an HPC environment, it is pertinent that its implementation scales well across a large amount of CPUs: our PiSvM [6] implementation is a slightly modified version of the parallel πSvM [7] which has been improved [4] to make more efficient use of the Message Passing Interface (MPI) standard [8] for parallel processing in HPC clusters.

Briefly, PiSvM's classifier generation is divided into three phases: Cross-validation, training, and testing. The cross-validation phase is used to choose the optimal hyperparameters that will result in the highest accuracy of the classifier, to best prevent both underfitting and overfitting of the model. The training phase feeds the preprocessed training data into the SVM algorithm which in turn calculates data-space coefficients using the previously selected hyperparameters. These coefficients can then implicitly be used to classify arbitrary input data (e.g., new acquired hyperspectral images) via a generated model classifier. In order to evaluate the classifier's accuracy, the testing phase aims at comparing its output for a selected set of input data with a correct classifications known beforehand. Note that for the scope of this paper we mostly include the testing within in the training phase.

2.2. The Dynamical Exascale Entry Platform

While the Dynamical Exascale Entry Platform (DEEP) [9] is of pre-exascale performance, it is supposed to provide the blueprints for a future exascale computing system. It is based on the Modular Supercomputer Architecture (MSA) [10] which consists of an arbitrary number of (accelerator) modules of different types (cf. Fig. 1), where each is tailored to fit the needs of a specific set of computation, storage, or communication tasks. These specialised accelerators are needed to reach exascale performance which cannot be reached with the current general purpose CPUs, storage, and interconnects.

The *Cluster Module* has the fastest CPUs, which makes is suitable for tasks that are the most computationally expensive, with limited scalability. The Extreme Scale Booster (ESB) [10] module can be described as putting the emphasis on scaling: it consists of manycore CPUs less powerful than the Cluster module, but makes up for it by the sheer number of cores. The ESB is best suitable for well scaling parallel tasks. Furthermore, the ESB module also includes the Global Collective Engine (GCE) integrated in its fabric that can with its Field-Programmable Gate array (FPGA) speed-up MPI collective operations in hardware, e.g. summing up values transmitted in MPI messages.

Finally, the Network Attached Memory (NAM) module [11] is a special module that can be described as existing in the fast fabric interconnection between the other modules. It has a large amount of fast memory (non-volatile high-performance RAM) and mostly functions as a fast intermediary storage. However, it also has a general purpose CPU and an FPGA which can be used by any application to perform near-data processing for increased performance.

Other modules, such as the Data Analytics Module, are currently not used by our PiSvM classification chain.
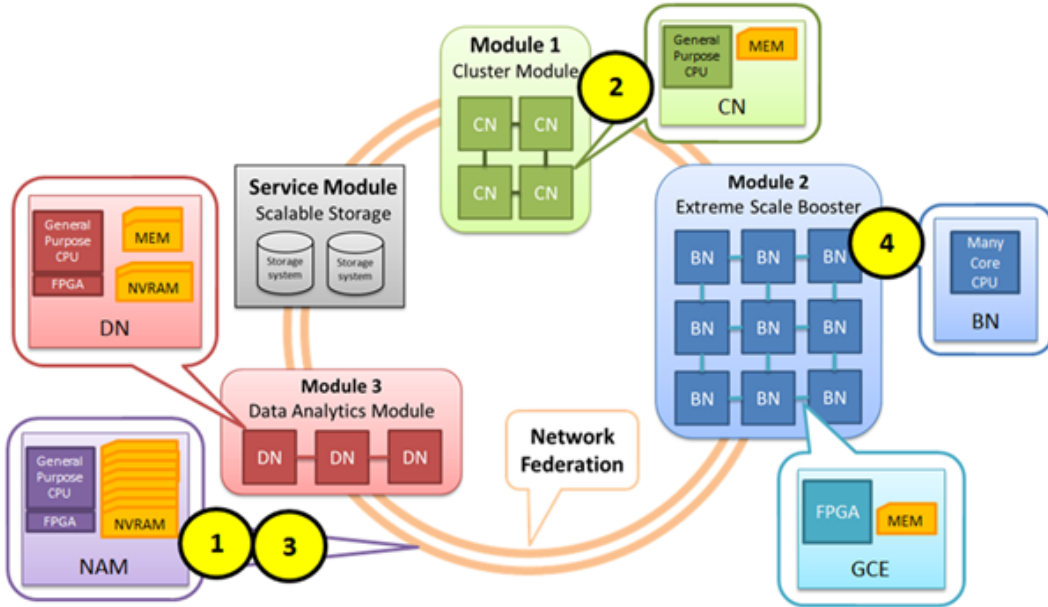
Fig. 2. Utilizing the Modular Supercomputer architecture for training (consult Sect. 2.2 for module details)

3. SUPPORT VECTOR MACHINES ON THE DYNAMICAL EXASCALE ENTRY PLATFORM

In this section, we propose several techniques that utilize the MSA modules of the DEEP system to accelerate the PiSvM workflow for very large datasets, e.g. Sentinel remote sensing datasets. This will lead to a better performance and usability of the PiSvM implementation.

The workflow is split into three parts, where each part is divided into a number of key steps; *cross-validation*, where the hyperparameters are selected; *training*, which also includes testing; *classification*, where the generated classifier is used to classify new data. For each step a corresponding accelerator module has been carefully selected, using novel approaches where necessary, in order to achieve exascale performance with PiSvM.

Furthermore, we automate the workflow of using the different accelerator modules by reducing the number of manual, and often redundant, operations made by the scientist. Hence, a scientist can specify the initial input parameters and the desired level of accuracy and hopefully observe as PiSvM converges on a model with the desired accuracy.

3.1. PiSvM Cross-Validation

The following is a step-by-step description, showing how the MSA is exploited to enhance the model's cross-validation phase with PiSvM. These steps are also depicted in Fig. 1: (1) It is determined with training and testing that a parameter space search is required in order to perform model selection, i.e. validation; (2) The relevant cross-validation datasets are loaded into the NAM module to speed-up data access; (3) An n-fold cross-validation over a grid of parameters (kernel, cost) is performed to produce an estimate of the out-of-sample performance by executing independent training processes on "folded" subsets of the datasets. This step is somewhat computationally expensive but each fold, however, is embarrassingly parallel, i.e. independent parallel computations, and can therefore utilise the ESB module. Each calculated parameter is then stored in the NAM module for quick access; (4) The best parameters with regards to the maximum accuracy in all folds across all parameter spaces can be computed in hardware via near-data processing by the FPGA inside the NAM module. Finally, this parameter set is then given as input to the training/test pipeline.

3.2. PiSvM Training

Similarly to the cross-validation phase, the MSA is also exploited to enhance the training phase of PiSvM, as is depicted in Fig. 2 and described in the following step-by-step: (1) The remote sensing training and testing datasets are loaded into the NAM module, this reduces latency and bandwidth restrictions when data is accessed; (2) Training a model is computationally expensive due to the inherent convergence process executed by PiSvM, it is therefore best suited for the DEEP Cluster Module, which offers powerful CPUs; (3) The trained model is placed in the NAM module to improve the access to it during testing and cross-validation; (4) Model testing in order to evaluate its accuracy is embarrassingly parallel which can therefore take advantage of the ESB module which contains the largest quantity of CPUs. Finally, if the model accuracy is deemed too poor during testing, step (2) is repeated with different kernel and/or cost parameters.

3.3. PiSvM Classification

For the actual classification of new data, the powerful CPUs of the DEEP Cluster Module are best suited. If the model obtained in the previous phases is still stored in the NAM module, it can be even retrieved from there.

4. SUMMARY AND OUTLOOK

We presented an approach to make SVM-based classification of big remote sensing data ready for the exascale computing era. The key to the approach is to exploit the specialised hardware accelerators of an exascale HPC system. We illustrated this by providing a mapping of a full SVM parallel processing chain to the various accelerator modules of the European Dynamical Exascale Entry Platform (DEEP). Wherever possible, the benefits of these hardware accelerators modules are leveraged: massively parallel manycore processing of embarrassingly parallel processes is performed on the Extreme Scale Booster (ESB), whereas tasks that scale not as well are executed on the high-performance Cluster Module, frequently used data is stored in the Network Attached Memory (NAM) including near-data processing via FPGA. This all is facilitated by a fast fabric interconnection and the Global Collective Engine (GCE) for MPI collective operations.

The described accelerator modules are currently being developed as part of the Dynamical Exascale Entry Platform – Extreme Scale Technologies (DEEP-EST) project, hence it is not yet possible to evaluate the suggested SVM exascale chain. Once the hardware is ready, we will deploy our solution to the pre-exascale system and evaluate it. To fully automate our SVM for exascale chain, we intend to use the same workflow techniques as we have successfully applied for an SVM chain in a non-exascale HPC environment [12, 13]. This eases the non-trivial usage of an HPC system.

By tracing our parallel processing PiSvM implementation on preliminary hardware prototypes, we have already noticed that there is further room for parallel processing speed-up. We have thus started to work on further PiSvM improvements.

Finally, first experiments of hyperspectral image classification using deep learning with Convolutional Neural Networks (CNNs) look promising [14]. We will therefore investigate whether this is a viable alternative to SVMs on the DEEP pre-exascale platform.

5. REFERENCES

[1] A. K. Maini and V. Agrawal, *Satellite Technology: Principles and Applications: 3rd Ed.*, John Wiley & Sons, 2014.

[2] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proc. of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[3] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20(3), pp. 273–297, 1995.

[4] G. Cavallaro, M. Riedel, M. Richerzhagen, J.A. Benediktsson, and A. Plaza, "On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 8, no. 10, pp. 4634–4646, 2015.

[5] G. Cavallaro, N. Falco, M. Dalla Mura, and J. A. Benediktsson, "Automatic Attribute Profiles," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1859–1872, 2017.

[6] M. Richerzhagen, "piSvM," website, 2015, https://github.com/mricherzhagen/pisvm.

[7] D. Brugger, "πSvM," website, 2014, http://pisvm.sourceforge.net.

[8] MPI Forum, "MPI: A Message-Passing Interface Standard. Version 3.0," September 2012, http://mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf.

[9] DEEP-EST project consortium, "DEEP-EST project," website, 2018, http://deep-est.eu/.

[10] N. Eicker, "Taming Heterogeneity in HPC," Keynote Presentation at SAI Computing Conference 2016, July 2016, https://youtu.be/aM9AkgG5ud4.

[11] J. Schmidt, "NAM – Network Attached Memory," Doctoral Showcase poster at International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2016, Nov. 2016, http://www.deep-projects.eu/images/nam_poster_SC16.pdf.

[12] S. Memon, G. Cavallaro, M. Riedel, and H. Neukirchen, "Facilitating Efficient Data Analysis of Remotely Sensed Images Using Standards-Based Parameter Sweep Models," in *Proc. IGARSS*. 2017, IEEE.

[13] S. Memon, G. Cavallaro, B. Hagemeier, M. Riedel, and H. Neukirchen, "Automated Analysis of Remotely Sensed Images Using the UNICORE Workflow Management System," in *Proc. IGARSS*. 2018, IEEE.

[14] J. Lange, G. Cavallaro, M. Götz, E. Erlingsson, and M. Riedel, "The Influence of Sampling Methods on Hyperspectral Image Classification with 3D Convolutional Neural Networks," in *Proc. IGARSS*. 2018, IEEE.