Building Serious Big Data Algorithms for Massive Datasets

Big Data is gaining much attention these days. With the master you have chosen, you are already on the right track to play an important part in this exciting new area of science and business. You have developed technical skills that are in very high demand in the Big Data sector.

In this thesis project, we will further hone these skills, by building Big Data algorithms. For our purposes, we will define as a Big Data criterion those datasets that have at least of order $O(10^9)$ datapoints. The algorithms we will use are mainly based on Fixed-Size Least-Squares Support Vector Machines, and we will extend them to Big Data scale. The main goal of the thesis will then be to achieve a remarkable time gain in analysing data sets, and accessing scales of data volume (both in memory and processing) currently inaccessible.

We will work on real-life datasets with economic relevance. The applications we have in mind are what we call serious Big Data; they are not your typical shopping cart analysis applications as professed by Amazon and the likes, but they are high-value applications in important areas like finance, economics and health, bearing not only value for the company or organization employing them, but also enhancing the well-being of the customer. Examples here are credit card theft recognition, risk assessment, fraud detection, illness diagnostics, customer relation management and smart city applications. Serious Big Data solutions can imply paradigm shifts in industries.

Another important aspect of Big Data is the use of appropriate software platforms. Especially as we try to parallelise data analytics algorithms, this becomes important. We will thus try to implement the algorithms discussed above in Big Data software tools like Hadoop, and experiment with a novel language that is rapidly becoming the Big Data standard, called Julia. Upon completion of this thesis, the student should be able to implement Big Data algorithms, analyse massive datasets and apply this analysis to real-life serious Big Data problems. This results in a comparative essay on F-SLS-SVM’s implemented with different levels of parallelisation.
Profile

The project will incorporate a theoretical part, in the sense that we look at the fundamental aspects of the algorithms and how to improve on them, but the main part will probably be implementing them. It is thus advised that you are interested in both. Knowledge of and interest in LS SVM's is a bonus, as is programming experience in distributed environments like Hadoop and the Julia language. Other languages used are common statistical and mathematical languages like R.

Promotor

Professor Bart De Moor

Daily Supervisor

Oliver Lauwers, olauwers@esat.kuleuven.be

Research Unit

STADIUS

Number of students

One or two. If two students are interested, we can add a component that compares the Fixed-Size LS-SVM algorithm with other algorithms, like the KSC-NET, and looks at how improvements in one of them can result in improvements in the other.