Data Summarization and Distributed Computation

Graham Cormode

ABSTRACT

The notion of summarization is to provide a compact representation of data which approximately captures its essential characteristics. If such summaries can be created, they can lead to efficient distributed algorithms which exchange summaries in order to compute a desired function. In this talk, I'll describe recent efforts in this direction for problems inspired by machine learning: building graphical models over evolving, distributed training examples, and solving robust regression problems over large, distributed data sets.

ACM Reference Format:
Graham Cormode. 2018. Data Summarization and Distributed Computation. In *PODC '18: ACM Symposium on Principles of Distributed Computing, July 23–27, 2018, Egham, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3212734.3212795

Although we may have to deal with a very high volume of data, we can often work with only an approximate representation of it to perform a desired computation. For example, in building a predictive model, we need to extract a modest number of parameters from a large number of examples in order to define the model. This leads us to consider the notion of *data summarization*: trying to capture aspects of the inherent behaviour of the data in a compact form, without necessarily having full access to the entirety of the data or complete information about what questions will be asked of it.

A powerful and generic approach to building a data summary is to randomly sample parts of the dataset. This can be used to answer a large number of different questions, and in many cases the guarantees on the resulting accuracy can be expressed as a function of the size of the sample. Other approaches to summarization emerge from using transformations of the data, random projections, and rounding, amongst many others. In some cases, these techniques yield better accuracy/size tradeoffs than random sampling, or allow summarization when random sampling fails.

Data summarization is closely linked with distributed computation. A central question about a distributed computation surrounds the amount of communication and coordination needed between the different computational entities. Algorithms for data summarization can contribute to distributed computation, and yield improved bounds or new approaches to solving problems, particularly when an approximate solution can be acceptable. The structure of distributed computation can impose additional constraints: we would typically like summaries that can be applied to fragments of a data set, and be combined so that they summarize the union of their inputs.

PODC '18, July 23–27, 2018, Egham, United Kingdom

ACM ISBN 978-1-4503-5795-1/18/07.
https://doi.org/10.1145/3212734.3212795

This talk will give an overview of some recent work in data summarization with a distributed flavour, and their applications in data analysis.

Learning Graphical Models from a Distributed Stream [3]. A current challenge for distributed data management is to support the construction and maintenance of machine learning models over data that is large, multi-dimensional, and evolving. The need to scale to distributed, streaming data requires new models and algorithms. In this setting, as well as computational scalability and model accuracy, we also need to minimize the amount of communication between distributed processors, which is the chief component of latency. We consider Bayesian networks, the workhorse of graphical models, and describe a communication-efficient method for continuously learning and maintaining a Bayesian network model over data that is arriving as a distributed stream partitioned across multiple processors. This builds on prior work that has addressed the abstract problem of maintaining the count of a number of events in a distributed setting [2]. Adopting this primitive yields a strategy for maintaining model parameters that leads to an exponential reduction in communication when compared with baseline approaches to maintain the exact MLE (maximum likelihood estimation). Optimizing the parameter settings leads to further improvements in the communication cost, while providing similar prediction errors for the target distribution and for classification tasks as the baseline.

Deterministic Distributed Summaries for High-dimensional Regression [1]. Work on approximate linear algebra has led to efficient distributed and streaming algorithms for problems such as approximate matrix multiplication, low rank approximation, and regression, primarily for the Euclidean norm ℓ_2. In recent work, we study other ℓ_p norms, which are more robust for $p < 2$, and can be used to find outliers for $p > 2$. Unlike previous algorithms for such norms, we give algorithms that are (1) deterministic, (2) work simultaneously for every $p \geq 1$, including $p = \infty$, and (3) can be implemented in both distributed and streaming environments. We focus on ℓ_p-regression, and show the resulting distributed algorithms are accurate and communication-efficient.

SHORT BIOGRAPHY

Graham Cormode is a Professor in Computer Science at the University of Warwick in the UK, where he works on research topics in data management, privacy and big data analysis. Previously, he was a principal member of technical staff at AT&T Labs-Research. His work has attracted over 11,000 citations in the literature and has appeared in over 90 conference papers, 40 journal papers, and been awarded 30 US Patents. Cormode is the co-recipient of the 2017 Adams Prize for Mathematics for his work on Statistical Analysis of Big Data. He has also edited two books on applications of algorithms to different areas, and co-authored a third.

ACKNOWLEDGMENTS

The work of GC presented in this talk is supported in part by European Research Council grant ERC-2014-CoG 647557 and a Royal Society Wolfson Research Merit Award.

REFERENCES

[1] Graham Cormode, Charlie Dickens, and David Woodruff. Leveraging well-conditioned bases: Streaming and distributed summaries in minkowski p-norms. In *International Conference on Machine Learning*, 2018.

[2] Zengfeng Huang, Ke Yi, and Qin Zhang. Randomized algorithms for tracking distributed count, frequencies, and ranks. In *ACM Principles of Database Systems*, pages 295–306, 2012.

[3] Yu Zhang, Srikanta Tirthapura, and Graham Cormode. Learning graphical models from a distributed stream. In *IEEE International Conference on Data Engineering*, 2018.