

SCALABLE LEARNING FOR GEOSTATISTICS AND SPEAKER RECOGNITION

BALAJI VASAN SRINIVASAN,
ADVISOR: PROF. RAMANI DURAISWAMI,
RESEARCH SUMMARY

ABSTRACT. *With improved data acquisition methods, the amount of data that is being collected has increased several fold. One of the objectives in data collection is to learn useful underlying patterns. In order to work with data at this scale, the methods not only need to be effective with the underlying data, but also have to be scalable to handle larger data collections. My research focused on developing scalable and effective methods targeted towards different domains, geostatistics and speaker recognition in particular.*

Initially we focused on kernel based learning methods and develop a GPU based parallel framework for this class of problems. An improved numerical algorithm that utilizes the GPU parallelization to further enhance the computational performance of kernel regression was proposed. These methods were then demonstrated on problems arising in geostatistics and speaker recognition.

In geostatistics, data is often collected at scattered locations and factors like instrument malfunctioning lead to missing observations. Applications often require the ability to interpolate this scattered spatiotemporal data on to a regular grid continuously over time. This problem can be formulated as a regression problem, and one of the most popular geostatistical interpolation techniques, kriging is analogous to a standard kernel method: Gaussian process regression. Kriging is computationally expensive and needs major modifications and accelerations in order to be used practically. The GPU framework developed for kernel methods was extended to kriging and further the GPU's texture memory was better utilized for much enhanced computational performance.

Speaker recognition deals with the task of verifying a person's identity based on samples of his/her speech utterances. Text-independent framework was considered here and three new recognition frameworks were developed for this problem. We proposed a kernelized Renyi distance based similarity scoring for speaker recognition. While its performance is promising, it does not generalize well for limited training data and therefore does not compare well to state-of-the-art recognition systems. These systems compensate for the variability in the speech data due to the message, channel variability, noise and reverberation. State-of-the-art systems model each speaker as a mixture of Gaussians (GMM) and compensate for the variability (termed nuisance). We propose a novel discriminative framework using a latent variable technique, partial least squares (PLS), for improved recognition. The kernelized version of this algorithm was used to achieve a state-of-the-art speaker ID system, that shows results competitive with the best systems reported on in NIST's 2010 Speaker Recognition Evaluation.

During the past decade, it has become relatively easy to collect huge amounts of data. Examples include data in astronomy, internet traffic, meteorology and surveillance. A goal of this collection is to mine the data for useful information and thus learn meaningful statistical patterns that allow one to predict/recognize unseen patterns.

1. LEARNING METHODS

Learning is a principled method for distilling predictive and scientific theories from raw data. There are different flavors to learning. In *regression*, a few observations are used to model a continuous target variable, e.g. predicting the temperature/rainfall at some point in the future based on current weather patterns. *Classification* attempts to model the observations to predict a discrete target variable, e.g. classifying a person based on his face image. In information retrieval, the observations are used to *rank* the data in order of certain preferences. Certain methods attempt to capture the general pattern underneath the data, e.g. Parzen window estimate to learn the underlying data distribution. A common theme in all these methods is to look for *special structures* in the unstructured raw data.

Learning methods can be broadly categorized as parametric and non-parametric approaches. Parametric approaches assume a structure to the function to be estimated and uses observations to estimate the parameters of the assumed structure and thus the function. When there is prior information available about the model, the parametric model is a favorable choice.

Non-parametric methods do not assume any such structure for the function and generally “allow the data to speak for itself”. They are very robust in modeling the non-linearities, and can be used when the parametric approaches fail (due to absence of prior knowledge of the model or due to improved robustness requirements).

Both these methods have their own advantages and disadvantages. The use of either of these methods for a particular application is determined by its effectiveness with the underlying data. Effectiveness of a learning method to a particular data depends on the nature of the attributes recorded and the target application.

2. LARGE SCALE DATA

The ease of data collection has led to a surge of the number of observable attributes and the low cost of data storage has resulted in a large number of samples being stored. This results in the availability of *tall fat* data for learning. Internet bigwigs like Google handle several petabytes of data daily and the data at this scale offer sufficient information to aid in better learning models. However, with the data at such scale, the scalability of any chosen learning approach becomes as important as its effectiveness to the underlying data. Non-parametric methods are computationally much more expensive than parametric methods and therefore require special focus.

The scalability of an approach can be addressed either via algorithmic improvements or via parallelization. Algorithmic improvements approximate the underlying problem or cast them in a different framework, and thereby reduce the overall asymptotic complexity. Parallelization techniques make use of modern multi-core architecture (e.g. OpenMP, GPU/CUDA) or distributed systems (e.g. Hadoop, MPI) to enable scalability to large datasets.

3. CURSE OF DIMENSIONALITY

In large scale data, several noisy dimensions are also encountered in most cases necessitating the need for denoising the data before applying any learning method. Further, for most machine learning tasks even though the data is very high dimensional, the true intrinsic dimensionality is typically very small. That is, the number of actual dimensions required for the target modeling/prediction is much lesser than those observed/recorded. All these have led to the study of several dimensionality reduction techniques and subspace modeling.

Dimensionality reduction techniques can be supervised or unsupervised. The popular principal component analysis (PCA) is an unsupervised technique that learns projects where the data variability is maximum. PCA is very used to remove noisy directions from the data. However, it leads to the same projections irrespective of the target application. Often, an application specific subspace is desirable for better learning, Fisher discriminant analysis, canonical correlation analysis, partial least squares are some techniques that aid in such supervised dimensionality reduction.

4. CHOICE OF LEARNING METHOD

The choice of a particular learning method is dictated by the characteristics of the underlying data and the target application. If the data is well-correlated and low-dimensional, any prior knowledge available on the data can be used to build a parametric model. In the absence of prior knowledge, non-parametric methods can be used. If the data is high-dimensional, PCA based dimensionality reduction is often the first step used. Alternately, if the precise target to be modeled is known, supervised dimensionality reduction can be used directly to achieve more correlated projections. The choice of learning methods can be made as before after the appropriate dimensionality reduction. Some dimensionality reduction techniques such as Canonical Correlation Analysis (CCA), Partial Least Squares (PLS) can be directly extended to a regression / classification technique, which is handy in some applications as well.

5. RESEARCH CONTRIBUTIONS

The primary focus of my research was to develop effective and scalable learning methods for geostatistics and speaker recognition. The scalability of the methods that we developed/explored was addressed via graphical processors (GPUs). The key contributions of my research are as follows:

- (1) A GPU-based accelerated framework for common computation bottlenecks in kernel machines [2, 7]
- (2) A fast preconditioner with flexible Krylov solver that works on top of GPU acceleration for an efficient regression framework [3]
- (3) A fast kriging on graphical processors and an efficient technique for kriging parameter estimation [4]
- (4) A non-parametric Kernelized Rényi distance (KRD) to measure distances between distributions based on samples from the distribution [1, 5]
- (5) A KRD-based greedy subset selection technique to select the most informative subset of a large dataset [1]

- (6) KRD-based similarity scoring for speaker recognition [5]
- (7) Partial least squares framework for improved speaker recognition [8, 9]
- (8) Kernelized partial least squares for speaker recognition with nuisance compensation and its extension to a one-shot similarity framework for symmetric speaker scoring [6]

REFERENCES

- [1] B.V. Srinivasan and R. Duraiswami. Efficient subset selection via the kernelized Rényi distance. In *IEEE International Conference on Computer Vision*, pages 1081–1088. IEEE Computer Society, September 2009. 2
- [2] B.V. Srinivasan and R. Duraiswami. Scaling kernel machine learning algorithm via the use of GPUs. In *GPU Technology Conference*. NVIDIA Research Summit, 2009. 2
- [3] B.V. Srinivasan, R. Duraiswami, and N. Gumerov. Fast matrix-vector product based fgmres for kernel machines. In *Copper Mountain Conference on Iterative Methods*, 2010. 2
- [4] B.V. Srinivasan, R. Duraiswami, and R. Murtugudde. Efficient kriging for real time spatio-temporal kriging. In *Conference on Probability and Statistics in Atmospheric sciences.*, 2010. 2
- [5] B.V. Srinivasan, R. Duraiswami, and D.N. Zotkin. Kernelized Rényi distance for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010. 2, 3
- [6] B.V. Srinivasan, D Garcia-Romero, D.N. Zotkin, and R. Duraiswami. Kernel partial least squares framework for speaker recognition. In *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011. 3
- [7] B.V. Srinivasan, Q. Hu, and R. Duraiswami. GPUML: Graphical processors for speeding up kernel machines. In *Workshop on High Performance Analytics - Algorithms, Implementations, and Applications*. Siam International Conference on Data Mining, 2010. 2
- [8] B.V. Srinivasan, W.R. Schwartz, R. Duraiswami, and L.S. Davis. Partial least squares on graphical processor for efficient pattern recognition. Technical report, University of Maryland - CS-TR-4968, 2010. 3
- [9] B.V. Srinivasan, D.N. Zotkin, and R. Duraiswami. A partial least squares framework for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011. 3