Description of Research

Matthew Piekenbrock

Summary

My graduate research focus is in Machine Learning (ML) and Artificial Intelligence (AI), and my interests are in exploring the intersections between unsupervised learning, statistical learning theory, and empirical analysis. I also enjoy building software in the realm of scientific computing and for reproducible research. Topic areas that interest me include e.g. clustering, dimensionality reduction, topology theory, density estimation, etc. I have supplemental research interests, background knowledge, or experience in random graph modeling, bayesian statistics, computational geometry, reinforcement Learning (such as adversarial learning!) and high performance computing.

Select Project Abstracts

Below are the abstract of some of the research projects I'm involved with. A more comprehensive list, complete with visualization, source code, etc. is available on my online CV. [1]

0.1 Efficient Multiscale Simplicial Complex Generation for Mapper

The primary result of the Mapper framework is the geometric realization of a simplicial complex, depicting topological relationships and structures suitable for visualizing, analyzing, and comparing high dimensional data. As an unsupervised tool that may be used for exploring or modeling heterogeneous types of data, Mapper naturally relies on a number of parameters which explicitly control the quality of the resulting construction; one such critical parameter controls the entire relational component of the output complex. In practice, there is little guidance on what values may provide "better" or more "stable" sets of simplices. In this effort, we provide a new

[1] See: `https://peekxc.github.io/`

algorithm that enables efficient computation of successive mapper realizations with respect to this crucial parameter. Our results not only enhance the exploratory/confirmatory aspect of Mapper, but also give tractability to recent theoretical extensions to Mapper related as persistence and stability.

0.2 Automating Point of Interest Discovery in Geospatial Contexts

With the rapid development and widespread deployment sensor dedicated to location-acquisition, new types of models have emerged to predict macroscopic patterns that manifest in large data sets representing "significant" group behavior. Partially due to the immense scale of geospatial data, current approaches to discover these macroscopic patterns are primarily driven inherently heuristic detection methods. Although useful in practice, the inductive bias adopted by the detection scheme is generally unstated or simply unknown. In this research effort, we describe a semi-supervised framework for automated point of interest discovery inspired by recent theoretical advances in efficient non-parametric density level set estimation techniques. We outline the flexibility and utility of the approach through numerous examples, and give a systematic framework for incorporating semisupervised information while retaining finite-sample guarantees.

0.3 Massive Parallel Iterative Closest Point

The Iterative Closest Point (ICP) problem is now a well-studied problem that seeks to align a given query point cloud to a fixed, reference point cloud a pairwise distance minimization. Intuitively, the "brute-force approach" approach is to calculate the pairwise distance from every point in the query set to every point in the reference set, resulting in quadratic runtime complexity. Alternatively, many spatial indexing data structures utilizing branch-and-bound (B&B) properties have been proposed as a means of reducing the algorithmic complexity of the ICP problem. While these structures are certainly useful, many were primarily developed for serial applications: is well known that direct conversion to their parallel equivalents often results in slower runtime performance than GPU-employed brute-force approaches due to the frequent suboptimal memory access patterns and conditional computations these spatial indexing structures often produce. In this application-motivated effort, we propose a novel two-step method which exposes the intrinsic parallelism of the ICP problem. Our solution involves an $O(\log n)$ approximate search, followed by fast vectorized search

we call the Delaunay Traversal, which we show empirically finishes in $O(k)$ time on average, where $k << n$. We demonstrate the superiority of our method compared to the traditional B&B and brute-force implementations using a variety of heterogeneous, benchmark data sets. We also show the usefulness of our method in the context of Automated Aerial Refueling by improving the runtime of the well known ICP algorithm.