BioGeometry NEWS July 2005
biogeometry.cs.duke.edu

Software

Coresets for Shape Fitting and Kinetic Data Structures

by Pankaj K. Agarwal and Hai Yu

Background. Computing various descriptors of the extent of a set P of n points in $\mathbf{R}^d$ has found many useful applications in shape analysis, data mining and other areas. These descriptors, called *extent measures*, either compute statistics of P itself (e.g., diameter, width), or they compute statistics of a (possibly nonconvex) geometric shape (e.g., sphere, box, cylindrical shell) enclosing P. Although traditionally P is assumed to be stationary, some recent applications, including the protein-structure analysis, call for maintaining extent measures of a set of moving points. These points may represent a rigid body in motion or a deformable object. The exact algorithms for computing most of these extent measures are generally expensive, and faster approximation algorithms are more suitable for these problems.

Approximations via Coresets. Using the notion of ε-*kernel*, Agarwal *et al.* [1] developed a unified framework for computing various extent measures of a point set approximately. Roughly speaking, a subset $Q \subseteq P$ is an ε-kernel of P if for every slab (i.e., a region enclosed between two parallel hyperplanes) W containing Q, the expanded slab $(1+\varepsilon)W$ contains P. The notion of ε-kernel yields simple, efficient approximation algorithms for a wide range of problems, including algorithms for minimum-width spherical and cylindrical shells containing P with running time $O(n+1/\varepsilon^{O(1)})$; data structures for maintaining approximate extent measures of moving points, which process $1/\varepsilon^{O(1)}$ events; and maintaining extent measures in the streaming model in polylog(n) time and space.

continued on next page

BioGeometry Meeting

The NCA&T community played a prominent role in the civil rights movement. The February One Monument of the Greensboro Four tells of the students whose early sit-in served as a catalyst: “These four A&T freshmen envisioned and carried out the lunch counter sit-in of February 1, 1960, in downtown Greensboro. Their courageous act against social injustice inspired protests across the nation and is remembered as a defining moment in the struggle for civil rights.”

Plans are being finalized for the BioGeometry meeting at North Carolina A&T State University on August 22-23 in Greensboro. PIs and students from all four institutions will participate. Details are available at http://biogeometry.cs.duke.edu/meetings/ITR/05aug/.

Since NCA&T is unique among the schools on the BioGeometry project, we present here a brief introduction.

NCA&T is one of the nation’s leading Historically Black Colleges and Universities (HBCU). It was established in 1891 as the Agricultural and Mechanical College for the “colored race” and in 1972 became part of the University of North Carolina system.

It now has an enrollment approaching 11,000 students, is one of North Carolina’s three engineering colleges, and is the largest producer of African American doctorates in engineering in the US.

NCA&T emphasizes programs in engineering, the sciences, and technology. It also has a strong program in accounting, and is a top producer of African Americans with bachelor’s degrees in teacher education and master’s degrees in counseling, adult education, instructional technology and leadership.

We refer the reader to the survey [2] for a detailed account and many recent developments on this topic.

The Coreset Software. The coreset technique is not only attractive in theory, but is effective in solving low-dimensional real-world problems. We developed the coreset software, which contains practical implementations of many approximation algorithms mentioned above, as an attempt to bridge the gap between theory and practice in this area. The software, in its preliminary form, will be released on the biogeometry website [6].

The basis of our software is a simple and practical algorithm for computing ε-kernels of a point set in any dimension [5]. In contrast to the original algorithm [1], which requires generalized linear programming as a subroutine, the new algorithm needs only an approximate nearest-neighbor-searching data structure. We use the software package ANN for computing approximate nearest neighbors [4]. We tested our program on a variety of inputs including both synthetic and real-world data. The empirical results show that our algorithm works extremely well in low dimensions (≤ 4) both in terms of size of the kernel and the running time. However, the size of the kernel and the running time increase exponentially with the dimension.

Our software also provides a simple and general incremental algorithm for the shape fitting problem. Using this algorithm, we have implemented approximation algorithms for minimum enclosing cylinders, minimum-volume bounding boxes, and minimum-width annuli. Empirical results show that the incremental algorithm is much faster than the standard algorithms based on coresets. In almost all cases, it returns an ε-approximate solution in a small constant number of iterations.

Another component of the coreset software is the application of the ε–kernel algorithm in the kinetic setting. In particular, we have implemented kinetic data structures for maintaining an approximation to the minimum orthogonal bounding box and the convex hull of a set of moving points in $\mathbf{R}^2$. By first computing a small coreset for the moving points and then maintaining the extent of the coreset, our algorithm reduces the number of events significantly and maintains a good approximation of the extent as the moving points set.

Visualization. A visualization program to demonstrate the performance of coresets in kinetic data structures is also available. This is the easiest and most direct way to view how effectively a small coreset can capture the extent measure of a large input point set, even if the points move. The program visualizes the convex hull of the entire point set as well as that of the coreset. Since the convex hull of the coreset processes much fewer events, it is more stable and visually pleasing.

Future Development. The coreset software will be updated and expanded on a regular basis. Currently, we are implementing a new algorithm for computing robust coresets [3]. A robust coreset can correctly capture the extent measure of a point set even in the presence of a small number of outliers. Therefore such a coreset will be very useful in handling inaccurate input data. In addition to adding new algorithms to the software package, we also plan to improve the user interface and visualization components. Moreover, as most of the algorithms in the software are conceptually simple, we plan to release a more polished and well-commented version of the source code in order to help the user understand the implemented algorithms.

References

[1] P.K. Agarwal, S. Har-Peled, and K. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(2004), 606-635.

[2] P.K. Agarwal, S. Har-Peled, and K. Varadarajan. Geometric approximation via coresets. In *Current Trends in Combinatorial and Computational Geometry* (E. Welzl, ed.), Cambridge University Press, 2005, in press.

[3] P.K. Agarwal, S. Har-Peled, and H.Yu. Coresets for approximating the extent of shallow levels, 2005, submitted for publication.

[4] D. Mount and S. Arya, ANN: Library for approximate nearest neighbor searching, http://www.cs.umd.edu/~mount/ANN/

[5] H. Yu, P.K. Agarwal, R. Poreddy, and K. Varadarajan. Practical methods for shape fitting and kinetic data structures using coresets. In *Proc. 20th Annu. ACM Sympos. Comput. Geom.*, 2004, pp. 263-272.

[6] http://biogeometry.cs.duke.edu/software/coreset.

The BioGeometry project is funded by the National Science Foundation under grant CCR-00-86013. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.