Finding Lookmarks for Extreme-Scale Simulation and Scientific Data

Lawrence O. Hall
Department of Computer Science & Engineering
University of South Florida
Tampa, Florida 33620-5399
hall@cse.usf.edu

Kevin W. Bowyer
Computer Science & Engineering
384 Fitzpatrick Hall
Notre Dame, IN 46556
kwb@cse.nd.edu

1. Extended Abstract

Petascale simulations hold the promise of extraordinarily detailed evaluation of disease, weather patterns, and other highly complex processes. There are challenges in distributing data among processors of a system capable of petascale computation to effectively carry out the simulation. There are also significant challenges in building and evaluating the simulation. In the process of building the simulation, unexpected anomalies may arise and rapidly tracking down all occurrences of them will enable faster, more efficient debugging of the simulation. After extremely large scale simulations are built, finding interesting regions in them becomes important [10, 7, 4, 12]. The process of browsing the simulation to find anomalies and interesting regions can take weeks to months. Providing tools that speed up this process and allow simulation users to quickly find regions of interest promises to greatly enhance the usability and scientific discovery power of petascale simulations.

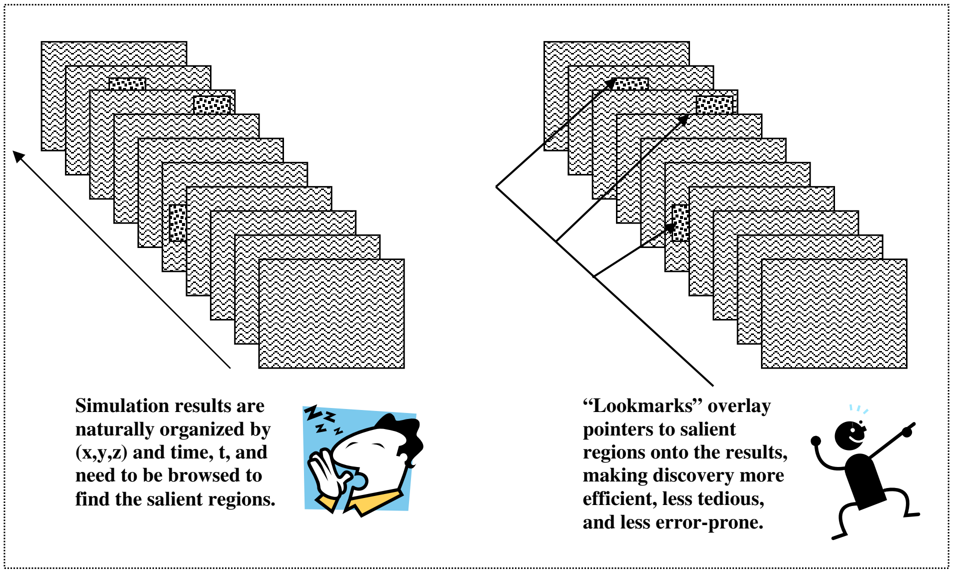
There are also many large-scale scientific data sets being collected. For example, it is estimated that only a small percentage of all astronomical data is viewed by scientists. Automated tools that could recognize interesting events and produce lookmarks to them could be valuable tools that lead to new discoveries. A lookmark is analogous to a bookmark and is a link that points the user to a region of a very large data set. Figure 1 illustrates the envisioned process of learning to produce lookmarks for very large-scale simulations. Methods that can learn to predict lookmarks from a modest amount of labeled data, where the training data can be incrementally acquired for large simulation and scientific data sets that are time varying and have rare interesting regions is a challenging goal for next generation learning.

Challenges are numerous and include the following. This type of data will be highly skewed with only a small percentage of examples that are in the (most) interesting class(es). Both the interesting class(es) and the non-interesting class(es) may be nonhomogeneous, as the class labels correspond to a user view of the role of the examples rather than the process that generates the examples. The large nonhomogeneous class will typically consist of many types of uninteresting data which themselves could be considered classes. Interesting examples may have been only in certain time steps of the simulation, or in certain time periods of the scientific data set; e.g., the recognition of red tides from satellite images when the red tides are intermittent. It is always difficult to get people to manually label a large amount of data for training, making semi-supervised and active learning highly useful. For many of these types of data sets, it is important to find the right region in a simulation, or in a body of water and so the measurement of success is different from the overall accuracy metric used in typical machine learning problems. Predictions may be smoothed spatially in the data set to produce regions rather than simply point predictions. It is important that the lift [] or some similar metric from the predicted regions be positively evaluated.

The idea of regional accuracy requires an effective, accepted measure of overlap. It also suggests that modified learning algorithms which can incorporate feedback from regional error may be necessary. So, they may be predicting from individual examples, but getting evaluated on regions and this circle needs to be closed.

Paraview [5], a collaboration of Kitware and DOE National Labs, is an open source visualization tool which can be applied to very large, distributed data sets. It runs on multiple platforms and supports distributed models. There is a custom plug-in to allow data to be labeled to reflect its salience.

The uninteresting data in very large data sets may be better separated from the interesting data after clustering into more homogeneous groups. Since it is very large data, it requires scalable clustering algorithms [6, 3, 14, 11, 8]. However, there is the question of the right number of clusters, which will require that a cluster validity metric be utilized [13, 9]. One will want to intelligently search for a good number of clusters. Perfection is not required since the goal is to enhance the separability of the interesting regions only.



Simulation results are naturally organized by (x,y,z) and time, t, and need to be browsed to find the salient regions.

“Lookmarks” overlay pointers to salient regions onto the results, making discovery more efficient, less tedious, and less error-prone.

Figure 1. Discovery of interesting regions using lookmarks.

Early work on finding salient regions from simulation data has been done on a storage container being crushed. The can data consists of as many as 443,872 labeled nodes across 25 to 44 time steps and has between six and nine features (depending upon the simulation). This is small enough it can be fit in a single memory, but we want to model the situation of a data set that has to be partitioned across memories. So, it was arbitrarily partitioned by breaking the can into spatial regions and one example is shown in Figure 2. Regions of interest were created by labeling a region being crushed in a particular time step. Then a predictive model using Random Forests [2, 1] was built on a time step(s) from a partition and applied to discover regions in a different partition. As there was access to data from four different simulations, it was possible to train on one simulation and predict regions of interest (lookmarks) in a separate simulation. The simulations differed by the speed of the can crushing, the number of nodes, the thickness of the bar during the crushing, and the number of time steps.

In Figure 3 an example is shown as a predicted region of interest in a simulation compared to the ground truth. What is interesting is that the classifier was trained on a different simulation that was partitioned as shown earlier (so classifiers were built on each partition and all classifiers are used to do predictions here). One can see that after smoothing (using an averaging operator with a chosen radius) the prediction is close to the ground truth. Certainly, the regional accuracy is good here.

We have outlined the needs for next-generation data mining approaches in large, skewed, time varying scientific data sets where region recognition is important. The minimal current work that has been done in the area indicates that next-generation approaches may be able to provide useful tools to scientists and engineers evaluating large-scale simulations and scientific data sets.

References

[1] R. Banfield, L. Hall, K. Bowyer, and W. P. Kegelmeyer. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):173–180, 2007.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] F. Farnstrom, J. Lewis, and C. Elkan. Scalability of Clustering Algorithms Revisited. *SIGKDD Explorations*, pages 51–57, 2000.

[4] L. Hall, D. Bhadoria, and K. Bowyer. Learning a model from spatially disjoint data. In *2004 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 2*, pages 1447–1451, October 2004.

[5] A. Henderson. *The ParaView Guide*. Kitware, Inc., United States, 2004.

[6] P. Hore, L. Hall, and D. Goldgof. A Cluster Ensemble Framework for Large Data sets. *IEEE International Conference on Systems, Man, and Cybernetics*, 2006.

[7] L.Shoemaker, R. Banfield, L. Hall, K. Bowyer, and W. Kegelmeyer. Using classifier ensembles to label spatially disjoint data. *Information Fusion*, 2007. To Appear.

[8] L. H. P. Hore and D. Goldgof. Creating streaming iterative soft clustering algorithms. In *NAFIPS 07*, San Diego, CA, 2007.

[9] N. Pal and J. Bezdek. On cluster validity for fuzzy c-means model. *IEEE Trans. Fuzzy Systems*, 1:370–379, 1995.

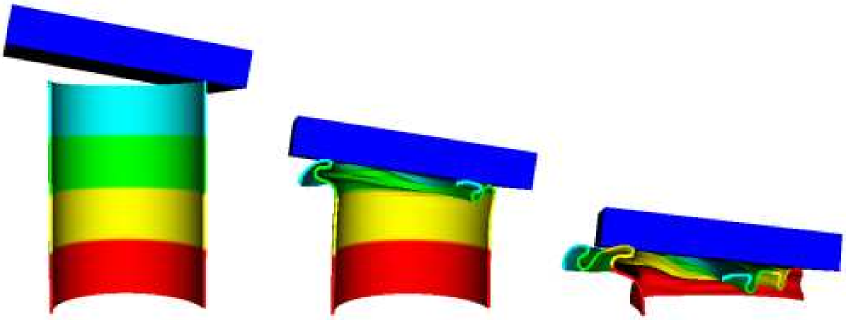


Figure 2. A visualization of the data as distributed across compute nodes for horizontal partitions. Four canister partitions and an impactor bar partition are shown in different gray levels as the storage canister is crushed. Partitions 0 to 3 in numerical order from top to bottom are beneath the impactor bar in the left view.

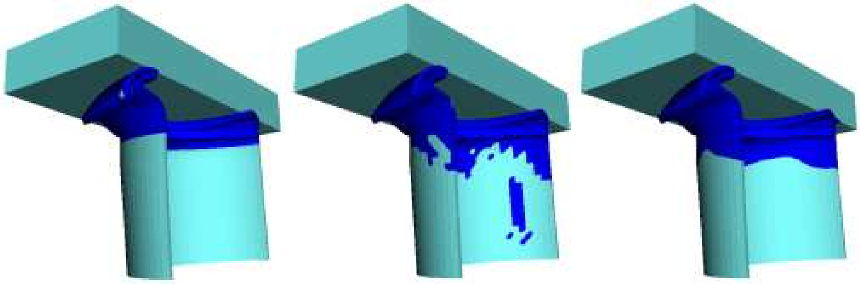


Figure 3. Left: Ground truth as labeled in time step 15 of Simulation 1. Center: Predicted salient regions including false positives (smaller regions) before smoothing. Right: Predicted salient regions after smoothing with no false positives.

[10] L. Shoemaker, R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. Learning to predict salient regions from disjoint and skewed training sets. In *18th IEEE Conference on Tools with Artificial Intelligence (ICTAI 2006), Arlington, Virginia, USA*, pages 116–123, 2006.

[11] A. Strehl and J. Ghosh. Clusters ensembles- a knowledge reuse framework for combining multiple partitions. *Journal of Machine learning Research*, 3:583–617, 2002.

[12] L. Tsap, M. Duchaineaub, D. Goldgof, and M. Shin. Data-driven feature modeling, recognition and analysis in a discovery of supersonic cracks in multimillion-atom simulations. *Pattern Recognition*, 40:2400–2407, 2007.

[13] X. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:841–847, 1991.

[14] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : An Efficient Data Clustering Method for Very Large Databases. *Proc. ACM SIGMOD Int'l. Conf. on Management of Data, ACM Press*, pages 103–114, 1996.