

BEADS: High Dimensional Data Cluster Visualization

Soujanya Vadapalli

Kamalakar Karlapalem

Centre for Data Engineering, International Institute of Information Technology-Hyderabad, INDIA.

email: {soujanya, kamal}@iiit.ac.in

Visualization of high dimensional data clusters is non-trivial. The data points within a cluster will have some shape (structure of the cluster) and size (spread of the cluster); though, exactly determining the shape is feasible for two or three dimensional data sets. In this poster presentation, we demonstrate BEADS System that uses the standard two dimensional shapes like, square and circle as *metaphors* to build and describe the actual shape and size of higher dimensional clusters. In this poster, the features of BEADS are emphasized by showing visualizations of various high dimensional data sets. Users can also test the utility of BEADS with their custom data sets to understand the data and evaluate it.

1 Introduction

The BEADS visualization system focusses on representing the shape and the spread of a high dimensional cluster on a 2-D plot. Shape of the cluster is defined by structure of relative placement of points in the original d -dimensional space. The spread of a cluster is defined by the region covering the points of that cluster in the d -dimensional space. BEADS has a new visualization scheme that is based on *break and build* approach. Each cluster is broken into small pieces. From a set of standard shapes like circle, rhombus, and square, the shapes that best fit the pieces of the cluster are identified. These standard shapes are then used as building blocks to construct the overall cluster structure.

We use the notion of metaphors to develop a 2-D image representation of high dimensional cluster. The metaphors are shapes in 2-D which are meaningful in higher dimensionality. For example, by referring to shapes like circle, square and rhombus, one can get a conceptual visualization of such regular structures in higher dimensions. Thus, a random shaped cluster is presented as a composition of primitive shapes. Our technique relies on already computed clustering results to display the data cluster-wise. This key idea is developed in this system with formalism for defining shapes, shape matching and shape composition.

Heidi [2] is a data visualization system that also focusses on displaying information cluster-wise; bringing out the inter-cluster and intra-cluster interactions across various subspaces. Heidi and BEADS complement each other: Heidi gives a global view and interactions, while BEADS gives the next level of detail in each cluster.

The highlights of BEADS system are:

1. Visualize high-dimensional data cluster focussing on shape and spread.
2. Intuitive correlation between the d -dimensional standard shapes (depicted by \mathcal{L}_p -norm envelopes) and the 2-D signature images used as metaphors.
3. Different levels of granularity (different number of beads) of visualization for hierarchical exploration of high dimensional clusters.
4. Ability to handle clusters of random (complex) shapes (due to building the cluster shape from pieces of cluster).

2 BEADS System

The input to the system is a set of already computed clusters to be visualized and the output is a set of 2-D visualizations, one for each cluster. The approach of the system is shown in Figure 1.

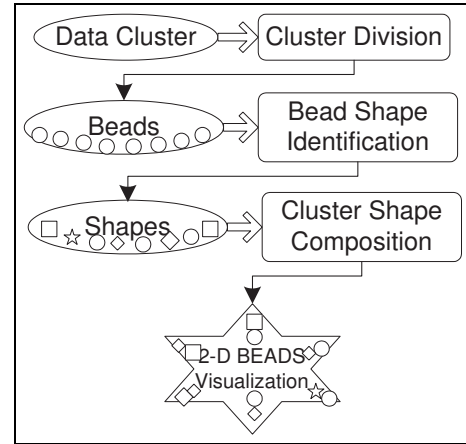


Figure 1: Beads Approach

Cluster Division: A subset of cluster points close to each other, such that the shape-fitting for this group of points could be easily obtained through a mathematical framework, is referred to as a *bead*. Such a group of points close to each other is obtained by using a partitioning clustering algorithm. Each partition is a bead.

Bead Shape Identification: After forming beads, a shape that best-fits each bead is identified from a set of standard shapes. A set of mathematical propositions and conjectures on space occupied by a mathematical function is used in developing the \mathcal{L}_p -norm Identification Algorithm. Due to space limitations, these propositions, conjectures and the algorithm details are beyond the scope of the paper.

Shape Composition: Representing cluster as a set of spatially connected beads on a 2-D plane, by using 2-D shapes as metaphors for the standard shapes identified in the original d -dimensional space is the key idea. Since the corresponding \mathcal{L}_p -norms that best-fit the beads are known, the idea is to use corresponding \mathcal{L}_p -norm envelope for 2-D space, as a *metaphor* to represent the d -dimensional envelope. For instance, a 3-D sphere's shape in 2-D is a circle and a 10-D hyper-cube's shape in 2-D is a square.

All the 2-D shapes of the beads are placed with respect to each other on a 2-D plot, called Beads Plot, based on:

1. Orientation of the beads along various dimensions with respect to the center of the cluster, by dividing the Beads Plot into sectors.
2. Relative placement of the beads in the original space
3. Distance of the bead's center to the center of the cluster
4. Size of bead (defined by the radius of the \mathcal{L}_p -norm identified)

3 Interpreting Beads Visualization

The bead closest to the origin is the part of cluster closest to the center of the cluster in its original space. The position of beads

around the origin and in various sectors indicate the position and orientation of the beads in various axes divisions. By axes division it denotes the division of d -dimensional space by the orthogonal d -axes.

A current limitation in the BEADS approach is the result obtained from the partitioning algorithm. Though, we used k -means algorithm to obtain beads, any another appropriate partitional algorithm could be used at the discretion of the user.

Technical details: Beads is implemented in C++ and heavily uses *gnuplot* software (Linux distribution) to obtain the Beads visualizations. All experiments are conducted on a system with 1.73 GHz CPU and 2 GB RAM.

Computational Complexity: The complexity of BEADS approach for each cluster \mathcal{C}_i is $\mathcal{O}(|\mathcal{C}_i|^2)$, $|\mathcal{C}|$ being the number of points in the cluster. The BEADS visualization system took not more than 1-2 minutes in generating the BEADS 2-D plots.

4 BEADS Visualization Snapshots

Beads has been tested on a number of synthetic data sets generated with various cluster shapes and sizes [4], and a few real data sets. We present a sample BEADS visualization for a 2-D data set for a visual check on the nature of the beads and their orientation for 2-D data cluster. The Bead plots are meant more for higher dimensional data clusters which could not be visualized. A textual representation of beads informations are also presented to the user in a tabular format.

Beads is performed over CHAMELEON dataset 1 [1]; the dataset has 8000 points and six clusters of random shape. The clusters are divided into 25 beads. The Beads visualization for one of the clusters is displayed in Table 1. The center of each bead is highlighted using a black dot. It can be noticed that the resulting composition of bead shapes on the Beads Plot share close resemblance with the original structure of the cluster.

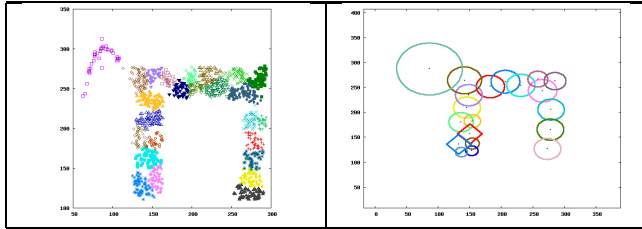


Table 1: Beads Visualization: CHAMELEON data set1

SynDECA generated high dimensional datasets: Beads Visualizations on a couple of high dimensional datasets are shown in Table 2. Beads Visualizations of two 7-D and 10-D clusters (having hyper-cube and random shapes) are displayed. The center of the beads are highlighted with small black circles. It is interesting to observe that in Beads Plot of 10-D hyper-cube, the shapes of all the beads are either square or rhombus. It can be observed that the 7-D random shaped cluster is hollow at its center; inferred by absence of a bead at the origin of the Beads Plot.

Results on Real-life data set: Beads is also used to visualize NBA players' data set. The data set contains the average match statistics of 310 players and employs RECORD algorithm [3] to obtain the prior clustering results needed by Beads. RECORD identified 4 clusters and the Beads visualization of two of the four clusters are shown in Figures 2 (a), (b). It is interesting to note that cluster represented by Beads Plots in Figure 2(a) has all circular beads and it is *hollow* to some extent at the cluster center, as there is no bead right at the origin of the Beads Plot. In the Bead Plots of cluster displayed in Figure 2(b), an interesting observation in these figures is the different sizes (radii) of the beads.

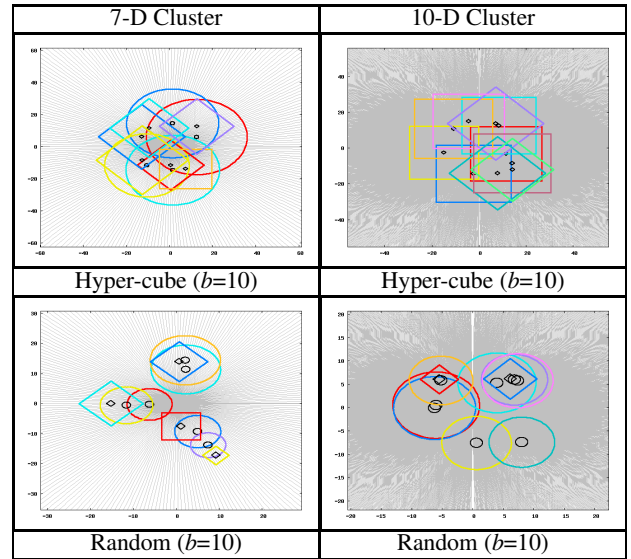


Table 2: Beads Visualization on 7-D and 10-D datasets

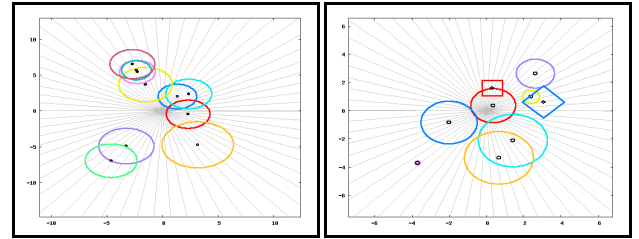


Figure 2: NBA Data Set: Beads Visualization on (a) Cluster 1 and (b) Cluster 2, $b=10$

5 Summary

In this poster paper, we present BEADS, a high dimensional data cluster visualization by having a 2-D representation of shape and spread of the cluster. The Cluster Division component, the Bead Shape Identification and Cluster Shape Composition form the core of the system. BEADS visualization consists of a 2-D plot, standard 2-D shapes which are used as metaphors to represent corresponding high-dimensional shapes of beads. The final resulting images convey the relative placement of beads with respect to the cluster center, the shape of the beads. We give a textual summary of the beads and their 2-D placement on the Beads plot in tabular format along with the image. More information on BEADS is available at: <http://cde.iit.ac.in/~soujanya/beads/>

References

- [1] G. Karypis, E. H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. pages 68–75. IEEE Computer 32(8), 1999.
- [2] S. Vadapalli and K. Karlapalem. Heidi matrix: Nearest neighbor driven high dimensional data visualization. In *VAKD*, pages 83–92. ACM, 2009.
- [3] S. Vadapalli, S. Valluri, and K. Karlapalem. A simple yet effective data clustering algorithm. In *Proc. ICDM*, pages 1108–1112, 2006.
- [4] J. Vennam and S. Vadapalli. SynDECA: Synthetic generation of datasets to evaluate clustering algorithms. In *COMAD*, 2005.