IDEA: ITERATIVE DICHOTOMIZATION ON EVERY ATTRIBUTE

1. Todo List

This is a rough timeline taking dependencies into consideration.

- Cosmetics
  - ~~Thinner lines for block outlines.~~
  - ~~Use a color scale to represent cluster scores.~~
  - ~~Graph legend showing colors with scores.~~
  - Only display clustered quadrants.
  - Try only showing the top 80% densest clusters.
- Performance Scores
  - ~~Eliminate self testing by setting aside train/test sets.~~
- Clustering
  - Takeaways from the Teak Experiment [1]
    - ~~Stop splitting if the variance increases, lives=3.~~
  - Mark (gray out) clusters which have a large delta between their performance scores (neighboring clusters which have high/low scores.)
  - Instead of using a 10% similarity rule in the GRIDCLUS algorithm [2], find the largest drop in block similarity and use that as a basis for similar.
- Competency
  - Compare performance for logged vs. unlogged coordinates.
  - Compare performance for pruned vs. unpruned with lives=3 system. Performance should increase once we prune.
- Interactive Clustering
  - Import data.
  - Assign attributes.
  - Show statistics for active cluster / all clusters.
- Feature Subset Selection
  - What's best among the best clusters?
- Contrast Sets
  - What makes cluster X different from unclustered or neighboring clusters?
- Clean up and profile code.

References

[1] Ekrem Kocaguneli, Tim Menzies, and Jacky W. Keung. Teak: Learning better case selection strategies for analogy based software cost estimation. IEEE Transactions on Software Engineering, 6, 2007.

[2] E. Schikuta. Grid-clustering: An efficient hierarchical clustering method for very large data sets. Pattern Recognition, International Conference on, 2:101, 1996.