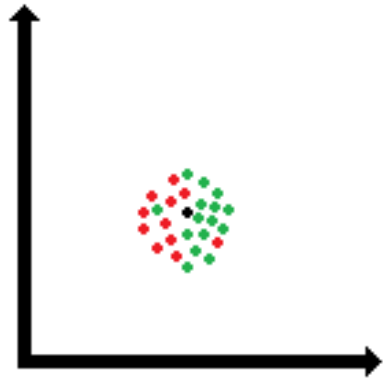
Selecting Limited Items Considering Entropy (SLICE)

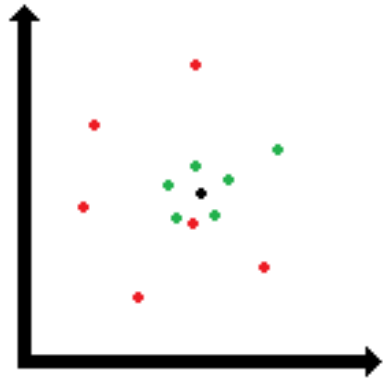
Akhila Gunjari, Yeshwanth Kuchimanchi, Avilash Rath

Motivations

➢ *The K-Nearest Neighbors algorithm struggles to handle scenarios in which the nearby data are evenly distributed or in which outlying data are less informative*



➢ *The Iterative Dichotomiser 3 (ID3) algorithm struggles in scenarios where attribute boundaries are murky and data proximity is important*



Goal

Determine whether combining k-NN and ID3 can show the strong points of both algorithms:

- ➢ *Leveraging proximity-based similarities*
- ➢ *Cutting across attribute boundaries*

Using entropy to determine scenarios in which each algorithm is best applied

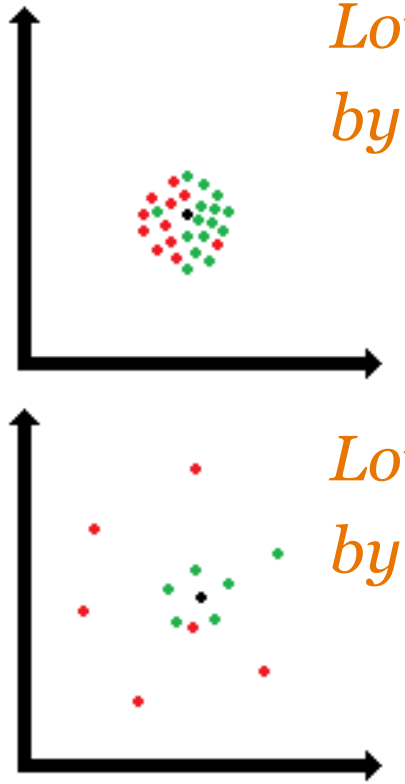
Why Entropy?

Entropy of a set of examples gives us a measure of information and similarity:

High entropy = low information (low similarity)

Low entropy = high information (high similarity)

We can use entropy to determine which algorithm to use and which points to classify



Low entropy by attribute

Low entropy by proximity

How can we leverage entropy?

Determine the entropies of k/2 nearest neighbors and k nearest neighbors (H_S and H_L, respectively)

If $H_S \approx H_L$, we use ID3 then k/2-NN

If $H_S \ll H_L$, we use k/2-NN

If $H_S \gg H_L$, we use ID3 then k-NN

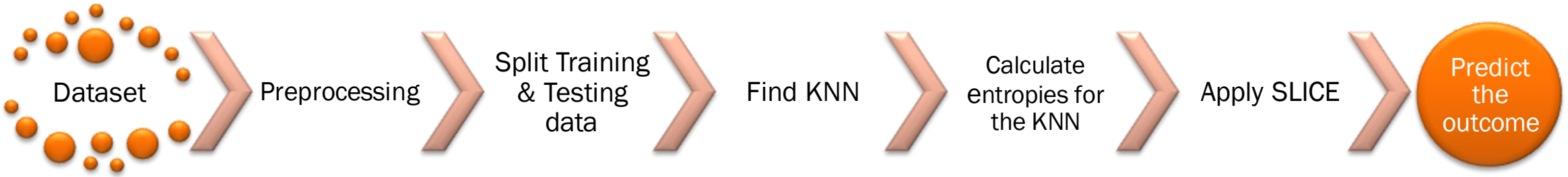
Dataset

We used Haberman's Survival Dataset to test our data against other algorithms

Our preprocessing steps were…

- *Binning: group a number of more or less continuous values into smaller number of bins*
- *Generate testing and training datasets.*

Methodology



Results

- SLICE: *Train set size: 489, Test set size: 123, Accuracy: 84.62%*
- ID3: *Train set size: 490, Test set size: 122, Accuracy: 79.5%*
- KNN: *Train set size: 489, Test set size: 123, Accuracy: 77.24%*
- K-means: *Train set size: 490, Test set size: 122, Accuracy: 76.5%*

Results

- SLICE: *Train set size: 244, Test set size: 62, Accuracy: 77.42%*
- KNN: *Train set size: 244, Test set size: 62, Accuracy: 75.81%*
- K-means: *Train set size: 245, Test set size: 61, Accuracy: 73.5%*
- ID3: *Train set size: 245, Test set size: 61, Accuracy: 37.7%*

Conclusions

Using the SLICE Algorithm, we are able efficiently use the ID3 technique and KNN to get a better accuracy with time complexities of :

Worst case: $O(n*\log(n) + n*f + K^2 *f)$

Best case: $O(n*\log(n) + n*f)$

Where n is number of examples, f is number of features, K is number of nearest neighbors.

Future Work

Potential ideas for future work include:

- *Finding the optimal parameters*
- *Use information gain to determine whether an ID3 classification step was good enough to classify the data*
- *Test the algorithm where ID3 runs on 2k points, k-NN is used for H_L, and k/2-NN is used for H_S*

Thank you

Akhila Gunjari

Avilash Rath

Yeshwanth Kuchimanchi