What is the Best Binning Scheme for our Data?

Andrew Zheng[1] , Shing Fung[2]
River Hill High School
NASA Goddard Space Flight Center, Code 673, Greenbelt, MD, 20771, United States

Abstract

There are many factors that change the state of the magnetosphere from a quiet to a disturbed state. When data was collected about each parameter, the data was used to conduct data analysis. However, after data analysis was conducted, there was a huge discrepancy within the plots that were created. It was then concluded that there were three things that could have caused the discrepancy. One, someone could have made a mistake when conducting data analysis; or, the different amount of data used to orchestrate the data analysis affected the plots produced; lastly, it could be the binning scheme used to conduct the analysis that is causing the discrepancy. For the topic we studied, It turned out that the binning scheme was causing the discrepancy within the data, and that before, the binning scheme was made arbitrarily. Thus, in this study we look for a binning technique with suitable binning criteria to apply to our data to produce an appropriate binning scheme. For this task, entropy-based binning was chosen as it is one of the best supervised binning methods. After the entropy-based binning technique was applied to bin the data, they were compared to arbitrarily binned equal width histograms that show the general shape of the data.

Introduction

Entropy-based binning is a bin boundary determination technique that is based on how much effect a parameter has on a set of results generated. Because we are looking at other data or results to help us bin the data, the entropy-based binning is a form of *supervised binning. Unsupervised binning* is where there are no outside factors affecting how the attribute at hand is being binned. Because supervised binning techniques bin one variable to give the most information about another variable, the produced binning schemes from supervised binning techniques generally produce better results than unsupervised binning techniques.

Entropy-based binning method

For entropy-based binning, you would need:

1. The set of data you want to bin
2. Another column of your ”result” data

Step 1: Order the data to be binned from lowest to highest value
Step 2: Using your result from Step 1, calculate the “Entropy” of the data

$$entropy = -\sum_{k=1}^{n} p_i \log_2 p_i$$

where n is the number of different result outcomes, and p_i is the probability of the nth result happening given the data.

Step 3: Determine possible splits for the data

Compared to other steps, this step is relatively easy. Because there are in principle endless possibilities for possible splits within the data, initial guesses on splits (or bin boundaries) are usually determined by taking the average of two ordered consecutive data records (of course omitting repeats). But if you are using a program, you can consider all cuts!

Step 4: Calculate the information gain

Taking into consideration partition a, your split groups are always <=a and a<. Splitting your data into those groups, you want to use the formula

$$Information\ Gain_{\leq a\ and > a} = \sum_{k=1}^{n} p_{\leq a} \log_2 p_{\leq a} + \sum_{k=1}^{n} p_{>a} \log_2 p_{>a}$$

to calculate the individual entropies for bins <=a and >a. Then, keeping the numbers separate, you want to apply the formula

$$Net\ entropy\ gain = p_{\leq a} Entropy_{\leq a} + p_{>a} Entropy_{\leq a}$$

to calculate net entropy gain, where $p_{\leq a}$ and $p_{>a}$ represents the percentage of data within the bin groups <=a and >a respectively

Step 5: Calculate the entropy gain

All we need to do is to subtract the original entropy and the net entropy gain from each other to calculate the entropy gain.

Step 6: Repeat steps 4-6 for all splits

Once you finish each split, if your entropy gain for the split you just did is higher than the old entropy split, scratch the old one and keep the new one, and repeat those steps until you have gone through all splits found in step 3. Else, your old entropy gain value prevails.

Step 7: Now that you have your split, you will repeat the process on the separate bins

You would repeat steps 2-7 with the 2 new bins that you have created to generate more bins and eventually come up with your binning scheme.

Method applied

Kp (n/a)	SW Velocity (km/sec)
0.7	332
2	321
1.7	329
0.7	344
2	338
3	339
2.7	354
2	365
2.7	404
4.3	454

Suppose we want to use Kp as our result.

When Kp is 1, 2, 3 or 4, those values are significant enough to represent their own respective values. Let’s set our result boundaries as 0-1, 1-2, 2-3, and 3-4, and 4-5 where the left boundary is included.

Kp (n/a)	SW Velocity (km/sec)
0-1	332
2-3	321
1-2	329
0-1	344
2-3	338
3-4	339
2-3	354
2-3	365
2-3	404
4	454

Step 1: Order the data from highest to lowest

	SW Velocity (km/sec)
2-3	321
1-2	329
0-1	332
2-3	338
3-4	339
0-1	344
2-3	354
2-3	365
2-3	404
4-5	454

Step 2: Using your result, calculate the “Entropy” of the data si

$$Entropy = -\left(\frac{2}{10}\log_2\frac{2}{10} + \frac{1}{10}\log_2\frac{1}{10} + \frac{5}{10}\log_2\frac{5}{10} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{10}\log_2\frac{1}{10}\right)$$

$$Entropy = 1.961$$

Step 3: Determine possible splits for the data

By taking the average of consecutive values from the SW Velocity column, we get:

325, 330.5, 335, 338.5, 341.5, 349, 359.5, 384.5, 429

Step 4: Calculate the information gain

Let’s take the split 335:

$$Entropy_{\leq 335} = -\left(\frac{1}{3}\log_2\frac{1}{3} * 3\right)$$

$$Entropy_{\leq 335} = 1.58$$

$$Entropy_{>335} = -\left(\frac{1}{7}\log_2\frac{1}{7} + \frac{4}{7}\log_2\frac{4}{7} + \frac{1}{7}\log_2\frac{1}{7} + \frac{1}{7}\log_2\frac{1}{7}\right)$$

$$Entropy_{>335} = 1.664$$

$$Information\ Gain = \frac{3}{10} * 1.58 + \frac{7}{10} * 1.664$$

$$Information\ Gain = 1.640$$

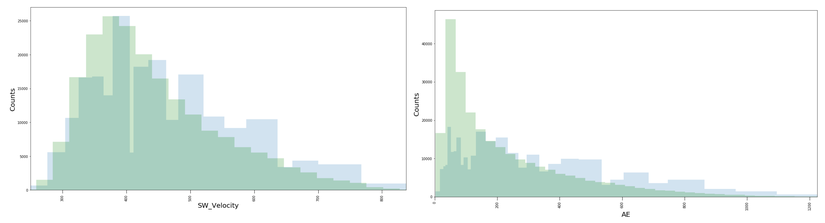
Step 5: Calculate Entropy Gain

$$Entropy\ Gain = 1.961 - 1.640 = .321$$

Step 6: Repeat steps 4-6 for all splits
Step 7: Now that you have your split, you will repeat the process on the separate bins

Conclusion

This binning scheme has given some expected results, but it has also given some unexpected results.



In these histograms, the blue part represents the binning scheme generated by the entropy-based binning technique, the light green part represents an arbitrarily determined equal bin width histogram used solely to generate the probability distribution of the data, and the darker green represents the overlap between the two histograms.

In the left, the histograms seem to match to a degree. On the contrary, the figure to the right shows the two histograms don’t nearly match as well as the ones on the left.

However, this is to be expected. Entropy-based binning generates histograms that have unequal bins, changing the entire shape of the histograms, so this would explain the graph on the right.

Another thing worth noting is how in the blue histograms of each graph, there is a portion of the graph that has been chipped out, like in the left graph at around x = 410. As it turns out, entropy-based binning works by binning the data based on how much informational value that part of the graph gives to the data analyst. Because entropy-based binning is based off of informational value, it is well known to be a reliable binning technique.

Acknowledgements

I would like to thank Shing Fung, my mentor, and Sophia Charles, my co-worker for working with me to achieve my goal this summer. I would also like to thank the National Space Club Scholars Program for giving me the opportunity to work at NASA.

References

Meurer, Kevin. “A Simple Guide to Binning Data Using an Entropy Measure.” *Kevin Meurer*, Kevin Meurer, 19 Nov. 2015, kevinmeurer.com/a-simple-guide-to-entropy-based-discretization/.
Sadawi, Noureddin. “Transforming Numerical to Categorical: Entropy-Based Binning.” *YouTube*, YouTube, www.youtube.com/watch?v=gmiINKkYc.
“C4.5 Algorithm.” *Wikipedia*, Wikimedia Foundation, 6 July 2018, en.wikipedia.org/wiki/C4.5_algorithm.