Branden Keck
Project Proposal

Data Mining with Cell Automata

For my project, I'd like to use cellular automation as a classification method. A paper by Tom Fawcett titled "Data mining with cellular automata" [https://dl.acm.org/citation.cfm?id=1412738] describes a method by which the classification regions can be "grown" using rules like those commonly found in Conway's game of life. These simple rules seem to generate very accurate results. However, I think that it would be interesting to extend the concepts covered in this paper by developing a growth model which uses more advanced concepts such as multilayer neural networks and/or other topics covered in this class. I would also like to extend this model to "higher dimensional" categorization problems that go beyond the simple case of having two features to one output.

A major challenge will be finding an interesting "classification" data set for which this approach is a viable alternative to other classification methods. Additionally, computing power is a limiting factor in which datasets can be chosen. Due to these restrictions, data will have to be chosen such that the number of data points is not excessive. Also, the number of "features" used in categorizing the data points will have to be low in number (but as previously stated I'd like to have more than two or three features to each data point). Image data will definitely be unusable. There are several data sets in the Kaggle repository that I believe will fit these criteria. Options that appear to be viable include NHL and NBA player statistics, Education statistics, and Wine reviews – with the goal being to predict which players/teaching methods/wines will be successful vs unsuccessful given a sample set of data.

In terms of interesting results, I hope to develop a method that can outperform other categorization methods in specific scenarios. I think that cellular automation is a very interesting topic because it is often explored in "visually pleasing" cases like Wolfram's 1-dimensional simulations. I hope that extending the rules of this type of theory to higher dimensions and using learning algorithms like multilayer neural networks to optimize "growth" patterns will reveal previously unconsidered benefits of cellular automation to the field of data mining.