CS 495 – Assignment 9: Clustering

Go to http://euclid.nmu.edu/~mkowalcz/cs495/notes/datasets/Soccer players/ and download the dataset fifa19CleanedNumericDiscreteOnly.arff. Instances are soccer players (over 16,000 of them). Features include skill levels, physical attributes, what position they play, how much they are paid, etc.

Load up the soccer dataset in Weka. Click on the "cluster" tab of the workbench.

Use SimpleKMeans for the algorithm with all the defaults. Click start. Weka will attempt to group similar instances together into clusters of similar instances.

Right click the result in the results list and select "visualize cluster assignments".

By default it will color datapoints according to cluster assignments (though you can change this with the drop-downs if you want). The algorithm detected two clusters of soccer players. Select different features for the players along the top (for the X and Y axis), and you can see for yourself that the two different players are very easy to distinguish by looking at various player characteristics. For example, players in one cluster almost always have significantly lower stamina than players in the other cluster. It may help to adjust the "jitter" slider so that multiple instances are not all plotted in the same spot on the screen. You can also right click or left click the table on the right hand side to select features to display on the X-axis or Y-axis, respectively.

In terms of soccer, what do the two clusters represent?

One may ask: how well would this perform if we didn't let it see that give-away feature, "positon"? Close the visualization window, select "classes to clusters evaluation", and select "(nom) position". Now it will do clustering and use the "position" feature to determine how well clustering performed. What sort of accuracy does it get for the class GK? What sort of accuracy does it get overall?

But we are still giving it an easy learning task. To make it a little more challenging for the clustering algorithm, click "ignore attributes". Scroll down to the bottom of the list and select the five goalie features; the ones that begin with "GK" (use ctrl-click or shift-click). Click "select" and run the algorithm again. From these results, which position appears to be most similar to GK?

Click on text box just right of where you choose the algorithm. Change the number of clusters to 4. Run the algorithm again. Looking at the results, it is apparent the algorithm is good at distinguishing goal keepers from other players, but not especially good at distinguishing among other player positions.

Let's try another algorithm. Select "EM" for the clustering algorithm. Unfortunately, there are so many instances that it would take a while to run this one. To remedy that, go back to the "Preprocess" tab, choose the following filter: unsupervised -> instance -> RemovePercentage. By default, this filter removes half of the instances. Apply the filter 4 times and you should see the number of instances drop down to 1,040. Go back to the cluster tab and apply the algorithm (it should take about 2 minutes – you can trim it down more if necessary). It should get an accuracy of about 50%. Not bad considering there are 27 classes and the training was unsupervised.

There should be a few positions that correspond almost perfectly with clusters. What are those positions?

There should be 4 positions (LB, RB, RWB, LWB) that are all neatly classified into the same cluster – with only a few instances falling into other clusters. What might this tell us about those soccer positions?

You should see the GK position split cleanly over 2 or 3 clusters. That's interesting! Try and figure out what that's about. Hint: Use the filter "unsupervised -> instance -> RemoveWithValues" in order to trim down the data to just goal keepers (use attribute=8, nominalIndices=4, invertSelection=true). Click on the "position" feature to verify it only has GKs now. Feel free to use different cluster algorithms, settings, and visualizations. What is going on with the GK position; what do you discover?