Cluster

----Siyu Zhou

- **3-A-b**

In preprocessing , I removed ‘player’ attribute. Since I can ignore hall of game when I do cluster, I didn’t remove it. I just ignored the missing value in “Strikeouts” attribute. I used “AttributeSelection” to do feature selection. Normalized data. Finally, I get ten attributes and one class as my proprecessing result.

- **3-A-c**

I used cluster number,model built time, incorrectly clustered percentage to evaluate the result of the cluster analysis methods. I want to use visualization as one of criteria, but I didn’t see the obvious difference in these five models.

MakeDensityBasedClusterer>SimpleKMeans>EM>Hierarchy>DBSCAN

- **3-A-d**

Since I know the class of this dataset is three, the numCluster which I set in weka should be no less than three in order to give this data a better cluster. If I choose three, sometimes my cluster label is ‘no class’. So, I choose four as numCluster in order to show original dataset class 0,1,2 in my cluster when I do

evaluation.

When I contrast these five methods, I think DBSCAN is the worst one. The cluster number of DBSCAN is generated by algorithm itself. In this dataset, DBSCAN just generate one cluster and depend on two parameters, so I think it is the worse one in this case , even though the incorrectly clustered percentage(ICP) is 9.3284% and time is 0.27s.As for hierarchy, it also always generate less clusters than the value I set for numCluster which is drawback of hierarchy. Here I always set numCluster as four. However, when I chose ChebyshevDistance as distancefunction and complete linktype, I can get the second lowest ICP 11.5672%. Another drawback of hierarchy is the longest running time as 3.81s. Which means it cannot be scale. With the change of distanceFunction, the ICP of SimpleKMeans are different. I chose the best result of SimpleKMeans with EuclideanDistance.The ICP of MakeDensityBasedCluster, SimpleKMeans and EM are 51.5672%, 51.791% and 51.2687% respectively, while the running time is 0.03s, 0.10s and 0.61s. Since the ICP gaps of these three are tiny, I think running time is more important when contrast. Although EM has the lowest ICP, but the running time

is far long than other two. So, EM isn’t the best choice of this dataset. With the most fast speed, I think MakeDensityBasedCluster is the best choice of this dataset.

- **3-B-f**

I got four rules when I did associate rule method. I set minimum support equal to 0.02 while minimum confidence is 0.9. In my four rules, two of them confidence equal to 1 while lift equal to 1.1. The two rules are as following:Triples=0.038835 28 ==> Hall_of_Fame=0 28 ,Batting_average=0.487805 28 ==> Hall_of_Fame=0. The third rule is Batting_average=0.536585 31 ==> Hall_of_Fame=0 28 which has 0.9 confidence and 1 lift. The last rule is Batting_average=0.517073 30 ==> Hall_of_Fame=0 27 which confidence is 0.9 and lift is 0.99. I think the former two is better because of higher confidence. I tried to change delta parameter, it seems don’t affect my rules result except number of cycles performed.