Data Mining Applied to Major League Baseball Batters

Michael Taylor
CS 522: Data Mining
Hood College, Computer Science Department
Frederick, Maryland, USA

Purpose of Experiment

- To extrapolate predictive analysis from current baseball statistics
- To seek out simpler and possibly better solutions to the age old problem of where to pitch the ball for a specific outcome
- We will look into each batter's success based upon pitch type and pitch location, using cluster analysis and association rules

Clustering

- DBSCAN
  - Used to produce cluster based upon neighbors
  - Ignores outliers
- Use pitch location

Association

- Produce association rules
- Apiori algorithm
- Use pitch type, speed and location

Data

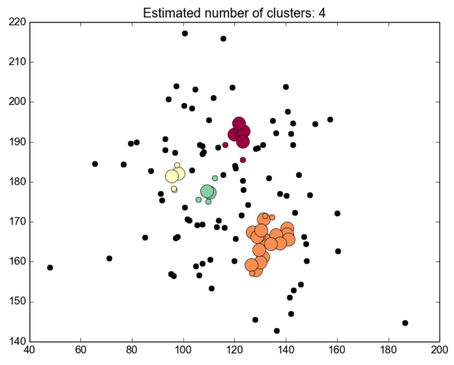
http://gd2.mlb.com/components/game/mlb/year_2015/

Field	Type	Null	Key	Default	Extra
id	varchar(50)	Yes		Null	
name	varchar(50)	Yes		Null	
des	varchar(50)	Yes		Null	
pitch_type	varchar(50)	Yes		Null	
x	Float	Yes		Null	
y	Float	Yes		Null	
start_speed	Float	Yes		Null	
end_speed	Float	Yes		Null	

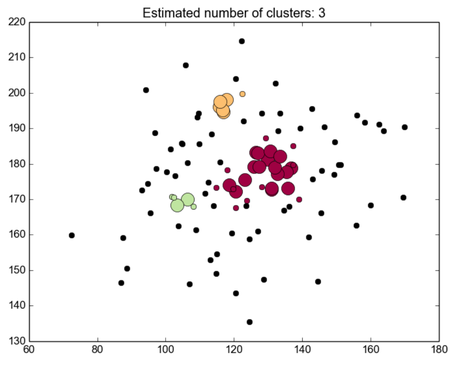
- **id**: Unique identify for each player.
- **name**: Players first and last name.
- **des**: Result of the pitch.
- **pitch_type**: Specific type of pitch, for example curveball.
- **x**: Pitch location on the x axis.
- **y**: Pitch location on the y axis.
- **start_speed**: Speed of the pitch when leaving the pitcher's hand.
- **end_speed**: Speed of the pitch when reaching the catcher.

Results

Mike Trout



Gerardo Parra



Conclusions

- From the results of this experimentation one can reach the conclusion that data analytics, specifically clustering and association analysis can be good tools for the use in major league baseball.
- From player scouting to game time decisions, the algorithms appearing here have shown that there is valuable information to be gleaned.

Future Work

- Expand this data to each player’s entire career:
  - which would produce a much larger sample size
  - more interesting and complete analysis can be obtained
  - This data would have to be weighted giving more weight to more recent data, because players tend to evolve over the course of their careers
- Apply weights to the total number of bases each hit reaches. A players home run zone is more important to be avoided than singles. This could also lead to changes in association rules.
- Include specific situations:
  - It would be important for pitchers to have a predictive analysis of what pitch to throw in order to induce a double play.