Finding Cricket All-Rounders by Clustering

Nitin Yadav

The Problem

Jacques Kallis (shown above) was one of the best All-Rounders to have played Cricket

Can we find some structures in sports data using Data Mining techniques? Here, we present a method to find All-Rounders in Cricket. All-Rounders are those players who are good at both the aspects of Cricket: Batting and Bowling.

The Data

The data was obtained from Cricinfo's Statsguru **[1]**, as per the following criteria:

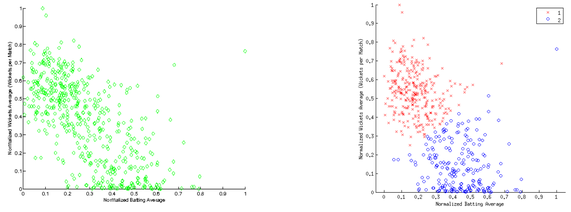
- A Player should have played more than 20 matches.
- A Player should have played between January 1, 1992 to February 1, 2014.
- Only One Day Internationals are counted.

Player	Span	Mat	Runs	HS	Bat Av	100	Wkts	BBI	Bowl Av	5	Ct	St	Ave Diff
SR Tendulkar (India)	1992-2012	**437**	17770	200*	45.79	49	146	5/32	45.10	2	133	0	0.68
ST Jayasuriya (Asia/SL)	1992-2011	**430**	13321	189	33.05	28	320	6/29	36.76	4	117	0	-3.70
DPMD Jayawardene (Asia/SL)	1998-2013	**407**	11401	144	33.33	16	7	2/56	79.71	0	201	0	-46.37
Inzamam-ul-Haq (Asia/Pak)	1992-2007	**376**	11659	137*	39.52	10	2	1/0	20.00	0	112	0	19.52
RT Ponting (Aus/ICC)	1995-2012	**375**	13704	164	42.03	30	3	1/12	34.66	0	160	0	7.37
Shahid Afridi (Asia/ICC/Pak)	1996-2013	**373**	7516	124	23.34	6	375	7/12	33.57	9	119	0	-10.22
KC Sangakkara (Asia/ICC/SL)	2000-2013	**362**	12116	169	40.11	16	-	-	-	-	356	87	-
M Muralitharan (Asia/ICC/SL)	1993-2011	**350**	674	33*	6.80	0	534	7/30	23.08	10	130	0	-16.27
R Dravid (Asia/ICC/India)	1996-2011	**344**	10889	153	39.16	12	4	2/43	42.50	0	196	14	-3.33
JH Kallis (Afr/ICC/SA)	1996-2013	**325**	11574	139	44.86	17	273	5/30	31.79	2	129	0	13.06

The data is stored in form of a matrix; each player being represented by a row and each feature (eg: Runs) being represented by a column. As we will see further, each player will represent a point in Euclidean space with dimensions as the columns of the matrix. To make our problem simpler, we reduce the number of dimensions to just two: Batting Average and Wickets taken per Match.

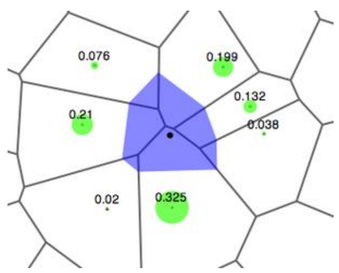
The Idea

- Cluster the data to obtain different clusters that represent bowlers and batsmen.
- For each point, find the 'strength' of assignment to each cluster. This is referred to as *affinity* in the paper **[2]**.
- Find the points with lowest maximum affinity. Those points are referred as the least *stable* points in the paper **[2]**. The least stable points are the points which cannot be said with much confidence to belong to a certain cluster. These points for our data should be the All-Rounders.



The plot on the left shows unclustered data, and the plot on the right shows the data clustered into two clusters, seemingly: Batsmen and Bowlers.

Finding Affinity



We use the weights of the Natural Neighbor Interpolant, as described in the paper **[2]**, to get the affinities for a point. The weights are also known as the area or volume (normalized) stolen from each neighboring voronoi cell.

As an example, in the figure above, the white voronoi cells have their voronoi centers as the clusters centers. The blue voronoi cell, which is added in the next step, has its center as our data point. The affinity of the data point towards any cluster is the area that blue cell steals from the cluster's corresponding white cell.

Results

We run experiments to find the least stable players in the clustering of bowlers and batsmen. Out of 509 players (our dataset size), we pick 100 least stable players, giving an accuracy of 73%. Some of the well known players in history of cricket that show up in the results, are shown below along with their stability (maximum affinity). Lower stability indicates that the player was better qualified as an All-Rounder as per our algorithm. Also, note that the least stability here can be 0.5 only.

Kapil Dev
All-Rounder
Stability 0.511

Carl Hooper
All-Rounder
Stability 0.511

Dale Steyn
Bowler
Stability 0.567

Neil Johnson
All-Rounder
Stability 0.571

Lance Klusener
All-Rounder
Stability 0.514

Jacques Kallis
All-Rounder
Stability 0.574

S. Jayasuriya
All-Rounder
Stability 0.530

Shane Watson
All-Rounder
Stability 0.500

Graham Thorpe
Batsman
Stability 0.585

Scott Styris
All-Rounder
Stability 0.534

References

[1] Cricinfo Statsguru:
http://stats.espncricinfo.com/ci/engine/current/stats/index.html

[2] Parasaran Raman, Suresh Venkatasubramanian:
"Power to the Points: Validating data memberships in clusterings"