Cricket-Insights Project Deliverable Cluster Analysis

Abhinav Chawla(IMT2013002)

Aditya Naidu(IMT2013003)

Shivam Kumar(IMT2013042)

S.S. Padhi(IMT2013043)

1 Introduction

This document presents the results that we obtained for Cluster Analysis of our dataset. We tested three main clustering algorithms on our dataset and analyzed the Silhouette plot for each output. The algorithms that we experimented with are as follows:

- K-means Clustering
- Hierarchical Clustering
- Density-based Clustering

It was deduced that K-means gave the most appropriate and interpretable clusters on our dataset. We then focused our attention on K-means algorithm and further experimented on it with varying number of clusters and modifications to our dataset via methods such as introduction of new columns, removal of existing ones and applying filters on data. In the subsequent sections, we present the details of the dataset used for K-means clustering and the results that we obtained.

2 Dataset

Our raw data consists of 577 yaml files where each file contains ball-by-ball details of IPL matches. The data was scraped from CricInfo for all the IPL seasons i.e. 2008-2016. We extracted the following two sets of data for the purpose of Cluster Analysis:

- Cumulative Batting Performance: This dataset consists of overall batting performance for several players over all the IPL seasons so far. The attributes extracted are as follows:
  - Name: Name of the player
  - Runs: Total runs scored by player so far
  - Wickets: Total number of times that the player has lost his wicket
  - Balls: Total number of balls faced by the player
  - Strike rate: Overall strike rate of the player
  - Avg: Batting average of the player
  - Batting style: To state if the player is Left-handed or Right-handed
- Cumulative Bowling Performance: This dataset consists of overall bowling performance for several players over all the IPL seasons so far.The attributes extracted are as follows:
  - Name: Name of the player
  - Runs: Total runs scored on player's bowling so far
  - Wickets: Total number of wickets taken by the player
  - Balls: Total number of balls bowled by the player
  - Economy: Economy of the bowler

For cluster analysis using K-means, we used three attributes for both the above mentioned datasets namely Runs,Wickets and Balls.

3 Results

K-means was run on both the datasets for several different number of clusters and corresponding cluster plots and silhouette plots were generated. For both the datasets, 2 to 4 clusters produces the highest mean silhouette coefficients. The two best plots, one for each dataset and their corresponding silhouette plot are included below.

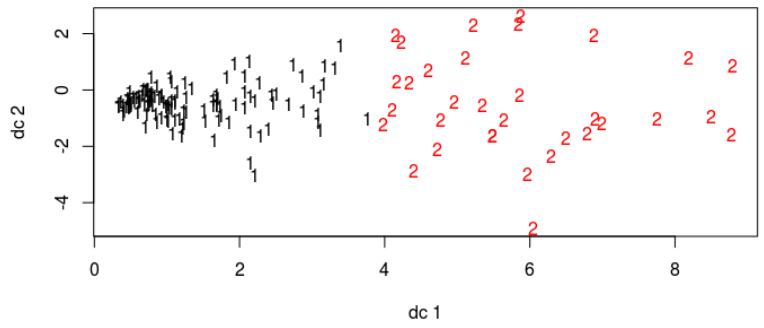

Figure 1: K-Means Cluster plot for 2 cluster on Cumulative Batting Performance

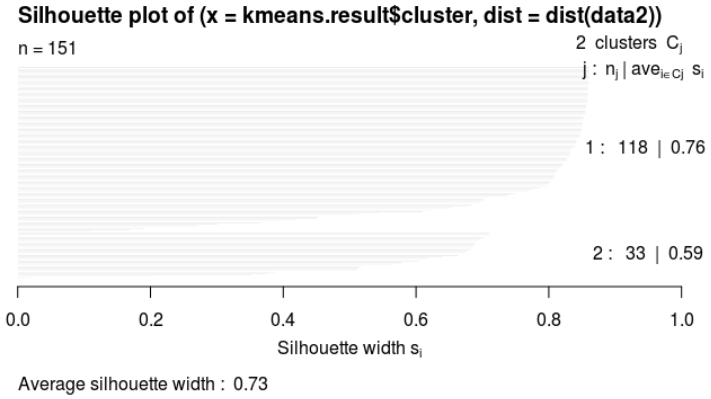

Figure 2: Silhouette plot for 2 cluster on Cumulative Batting Performance

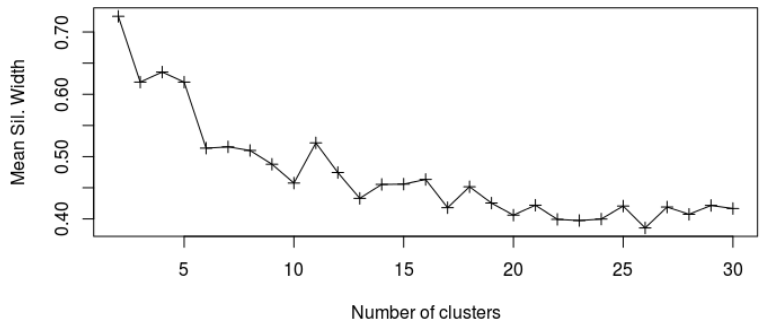

Figure 3: Plot for variation of Mean Silhouette Width with number of cluster for Cumulative Batting Performance

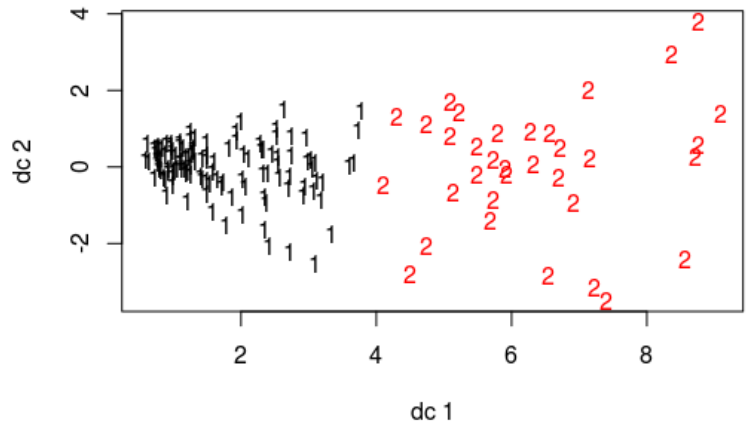

Figure 4: K-Means Cluster plot for 2 cluster on Cumulative Bowling Performance

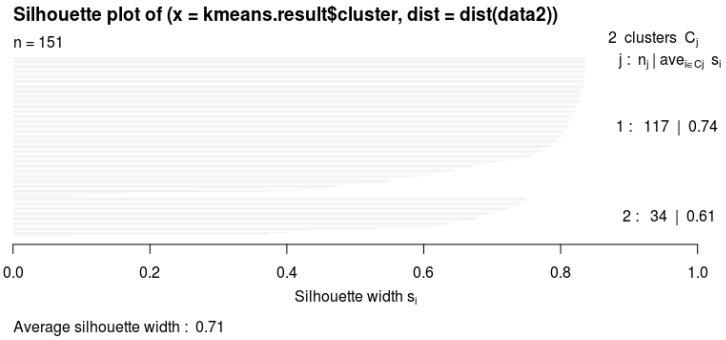

Figure 5: Silhouette plot for 2 cluster on Cumulative Bowling Performance

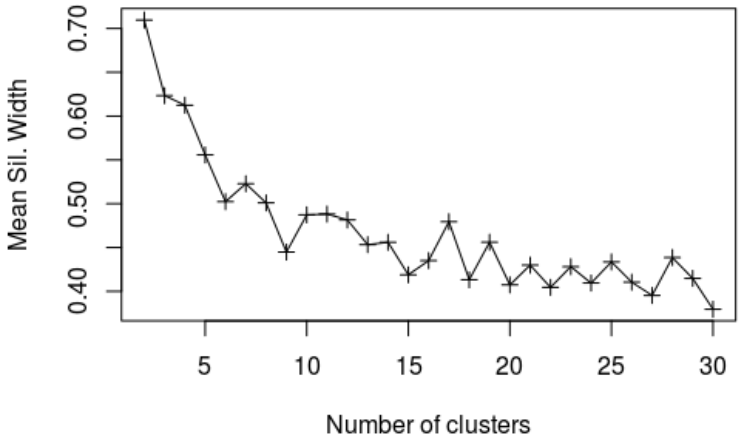

Figure 6: Plot for variation of Mean Silhouette Width with number of cluster for Cumulative Bowling Performance