Analyzing Sports Training Data with Machine Learning Techniques

PURDUE UNIVERSITY

Rehana Mahfuz, Zeinab Mourad, and Aly El Gamal*

Electrical and Computer Engineering, Purdue University

Introduction

Motivation: Data Science has potential in inspecting sports data and helping sports teams make better-informed decisions.

Aim: To find structure in Purdue Women's Soccer training data.
22 players, 48 drills, 9 features.

Approach: Unsupervised learning problem: clustering, while extracting most relevant features

Methods

Data Preparation

Extracted players: 22 observations, 9 features.

Extracted drills: 48 observations, 9 features.

Features scaled according to median using formula: $\frac{value*100}{median(feature)}$

Players before scaling

1191.7363	68.12257	5.990286	155.83886	0.54285714	4.8857143	13.542857	9.105528	136.490286
1177.6661	66.24702	9.642982	135.59649	0.84210526	4.0175439	14.526316	7.509234	118.691053
1011.6468	59.8561	10.65	150.14537	1.09756098	4.0243902	8.341463	9.108663	122.798537
1003.456	60.21524	6.344762	172.47222	0.66666667	7.8571429	11.746032	10.541047	0
962.4395	55.96116	9.823953	111.29581	0.86046512	3.2093023	8.906977	6.864924	104.532558
1221.1298	66.73782	6.738727	163.89127	0.6	4.3272727	11.272727	8.926555	113.207455
883.8558	49.97733	10.528667	137.33817	1.26666667	4.8833333	11.233333	7.622221	71.418667
1151.3495	62.79691	13.215636	159.32	1.30909091	6.0909091	9.927273	8.67581	112.338364
1099.7553	65.16981	17.636981	164.73057	1.73584906	6.2830189	11.90566	10.966506	114.596792
1225.421	66.9676	17.9518	170.982	1.64	4.22	13	9.569917	128.2964
1319.2053	73.4539	25.278814	198.63254	2.10169492	5.8305085	11.372881	11.079212	130.237119
1070.3317	61.55857	14.473968	151.23698	1.36507937	3.7301587	8.809524	8.473943	107.019683
1024.4819	58.23207	6.268448	154.54414	0.70689655	4.637931	15.137931	8.83018	98.740862
666.6064	38.58872	4.757949	88.90744	0.58974359	3.2564103	9	5.071457	57.557179
1078.6988	57.66679	22.034821	156.3025	1.94642857	5.0178571	11.125	8.27927	1.143929
741.0243	44.74	11.232	113.35633	1.16666667	2.6	7.833333	6.933838	98.994
673.1279	33.38474	0.38	31.00947	0.05263158	0.1052632	3.631579	1.601257	84.635789
782.0648	45.03389	15.857037	113.47963	1.38888889	4.0555556	9.851852	6.505926	80.022407
763.419	46.19897	10.48931	98.57034	1.03448276	4.3103448	7.172414	6.404119	113.643793
1187.506	65.97103	22.206552	176.26655	1.96551724	5.8793103	12.034483	9.94725	123.102759
1107.0095	61.44155	12.282931	126.93069	1.17241379	3.7413793	8.068966	7.266498	130.475517
946.1482	53.81053	18.822456	150.18298	1.73684211	4.8421053	11.666667	8.567051	120.292807

Players after scaling

113.77969	113.47017	54.750806	103.40314	46.416286	113.126432	121.14371	106.86616	121.031094
112.43636	110.3461	88.136207	89.9718	72.003104	93.024352	129.94095	88.13141	105.247842
96.58585	99.70089	97.34028	99.62536	93.845508	93.182876	74.61615	106.90296	108.890103
95.80385	100.29911	57.990695	114.43981	57.002457	181.928473	105.07073	123.71399	0
91.88784	93.21319	89.79027	73.84767	73.572939	74.309896	79.67478	80.56953	92.692969
116.58601	111.16361	61.591511	108.74613	51.302211	100.195979	100.83692	104.76566	100.385328
84.38515	83.24607	96.231301	91.12745	108.304668	113.071302	100.48453	89.45747	63.329631
109.92381	104.59933	120.790023	105.71297	111.932097	141.032155	88.80155	101.82281	99.614672
104.99791	108.55182	161.200815	109.30302	148.421492	145.480367	106.49864	128.70735	101.617306
116.99571	111.54636	164.078238	113.45101	140.226044	97.712129	116.28774	112.31641	113.765266
125.94965	122.35043	231.046646	131.79787	179.702661	135.0027	101.73282	130.03011	115.486175
102.18873	102.53666	132.291091	100.34968	116.719317	86.370083	78.80305	99.45362	94.898397
97.81127	96.99579	57.293193	102.54406	60.44226	107.389127	135.41198	103.63457	87.557254
63.6435	64.27632	43.487329	58.9924	50.42525	75.400658	80.5069	59.52067	51.038126
102.98756	96.05421	201.396778	103.71078	166.426816	116.186138	99.51547	97.16886	1.014365
70.74847	74.52237	102.65972	75.21488	99.7543	60.201786	70.07082	81.37833	87.781721
64.26614	55.60817	3.473174	20.57559	4.500194	2.437319	32.48524	18.79301	75.049753
74.66677	75.01189	144.932246	75.29669	118.755119	93.904495	88.12689	76.35618	70.958893
72.88658	76.95253	95.871587	65.40399	88.452088	99.804021	64.15875	75.16134	100.772246
113.37582	109.8864	202.966381	116.95745	168.058968	136.132685	107.65098	116.74495	109.159868
105.6905	102.34175	112.265159	84.22182	100.2457	86.62989	72.1786	85.28256	115.697572
90.33245	89.63093	172.035976	99.65032	148.506401	112.116686	104.36079	100.54638	106.668178

Feature Reduction

Feature selection:

Correlation matrix using Pearson's method: $$\frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{\sqrt{(n\Sigma x^2-(\Sigma x)^2)(n\Sigma y^2-(\Sigma y)^2)}}$$

Correlation Matrix

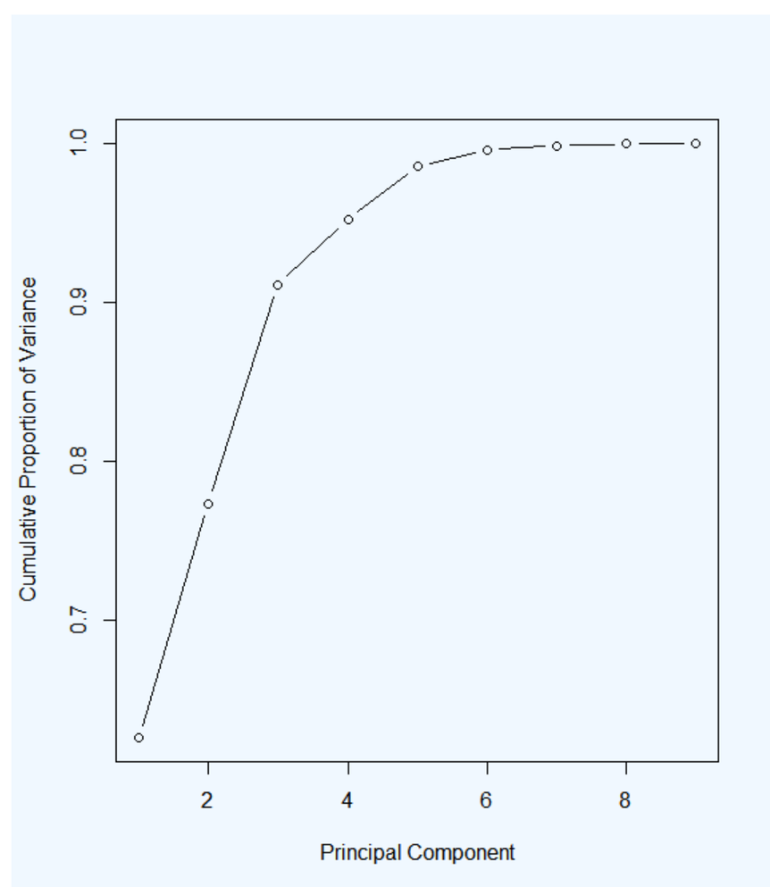
	Distance Total	Distance Per Min	High Speed Running	HML Distance	Sprints	Accelerations	Decelerations	HML Distance Per Minute	Average Heart Rate
Distance Total	1	0.8186695	0.4207638	0.8683186	0.4559951	0.4621517	0.6832496	0.6167345	0.2669202
Distance Per Min	0.8186695	1	0.3508357	0.7582125	0.3862897	0.425612	0.544916	0.7940255	0.3648052
High Speed Running	0.4207638	0.3508357	1	0.6397767	0.937304	0.3507593	0.1994807	0.5341228	0.1821072
HML Distance	0.8683186	0.7582125	0.6397767	1	0.67865	0.589473	0.6798512	0.8152752	0.251808
Sprints	0.4559951	0.3862897	0.937304	0.67865	1	0.403034	0.2317703	0.5676638	0.2018754
Accelerations	0.4621517	0.425612	0.3507593	0.589473	0.403034	1	0.5570324	0.5056643	0.1538933
Decelerations	0.6832496	0.544916	0.1994807	0.6798512	0.2317703	0.5570324	1	0.5063563	0.1224072
HML Distance Per Minute	0.6167345	0.7940255	0.5341228	0.8152752	0.5676638	0.5056643	0.5063563	1	0.2832838
Average Heart Rate	0.2669202	0.3648052	0.1821072	0.251808	0.2018754	0.1538933	0.1224072	0.2832838	1

Feature transformation:

Principal components analysis: Find mutually orthogonal axes with maximum variance.

First 6 principal components account for 99.6% of variance.

Principal Component #	Proportion of variance
1	62.6414207
2	14.71655923
3	13.77081087
4	4.064271592
5	3.373364063
6	1.051230178
7	0.24267435
8	0.101564035
9	0.038104976



Clustering

K-means:

Initialization: Each point is a center.

Expectation
Assign each point to closest center.

Maximization
Recalculate centers as means of clusters.

Terminating Condition: Centers converge.

Challenge: Random initialization of centers yielded different results at each execution.

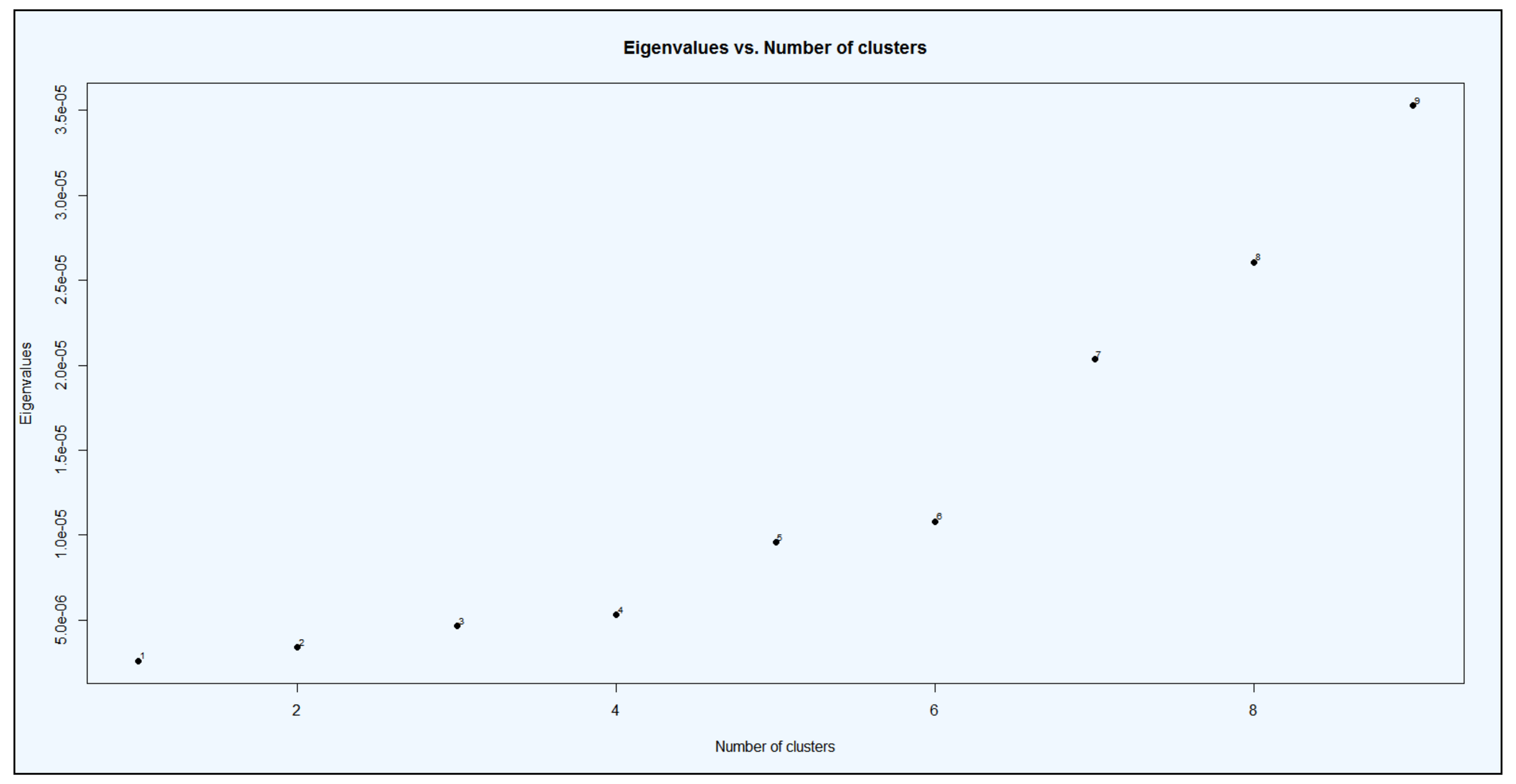
Solution: Use centers for which total within-cluster sum-of-squares is minimum.

Spectral Clustering:

Gaussian Kernel Similarity Function $$W_{ij} = e^{\frac{|x_i - x_j|^2}{2\sigma^2}}$$

Largest k eigenvalues of normalized laplacian of affinity matrix are subjected to k-means.

Eigengap method: Optimum k is given by index of first encounter of big jump in eigenvalues.
The appropriate sigma was found for which k=4.



Finding similarities in results

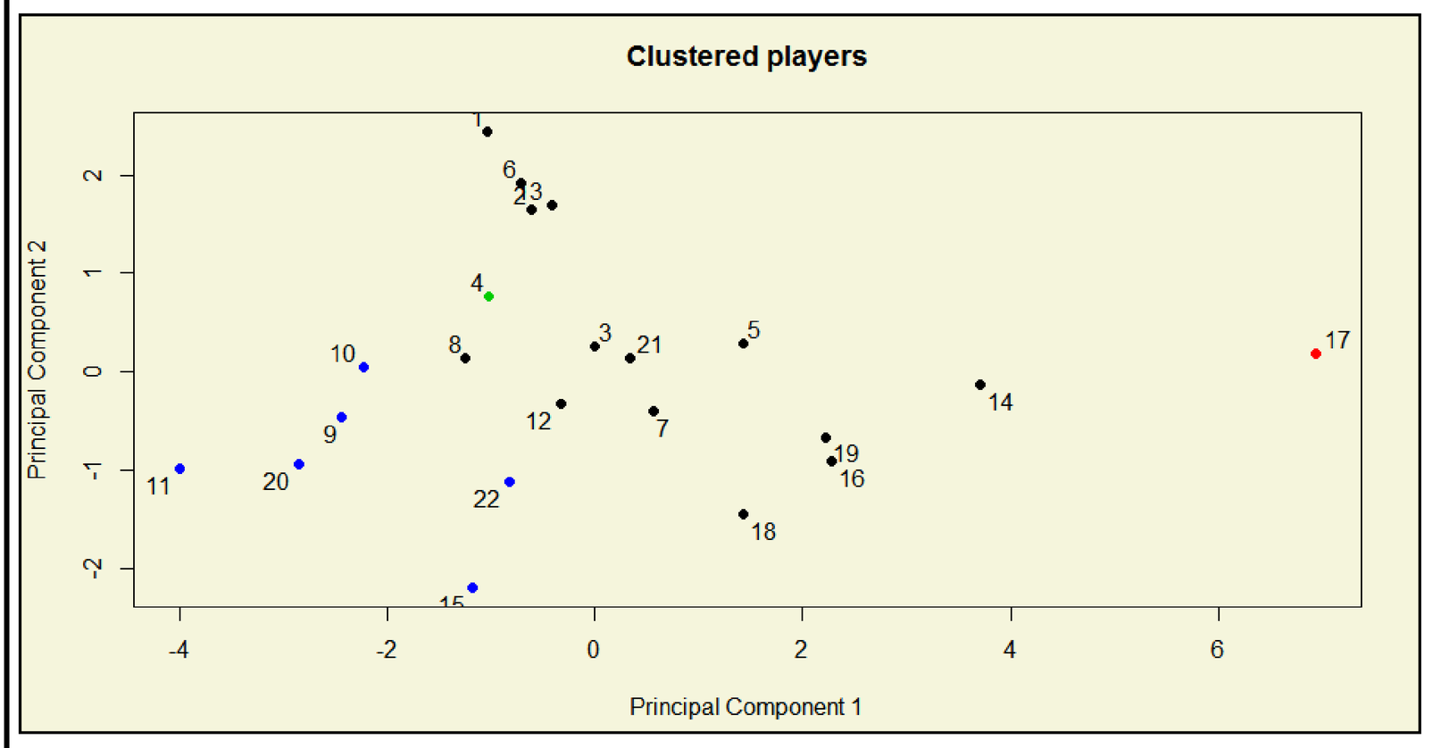
K! possibilities of matching k clusters from two clustering results.

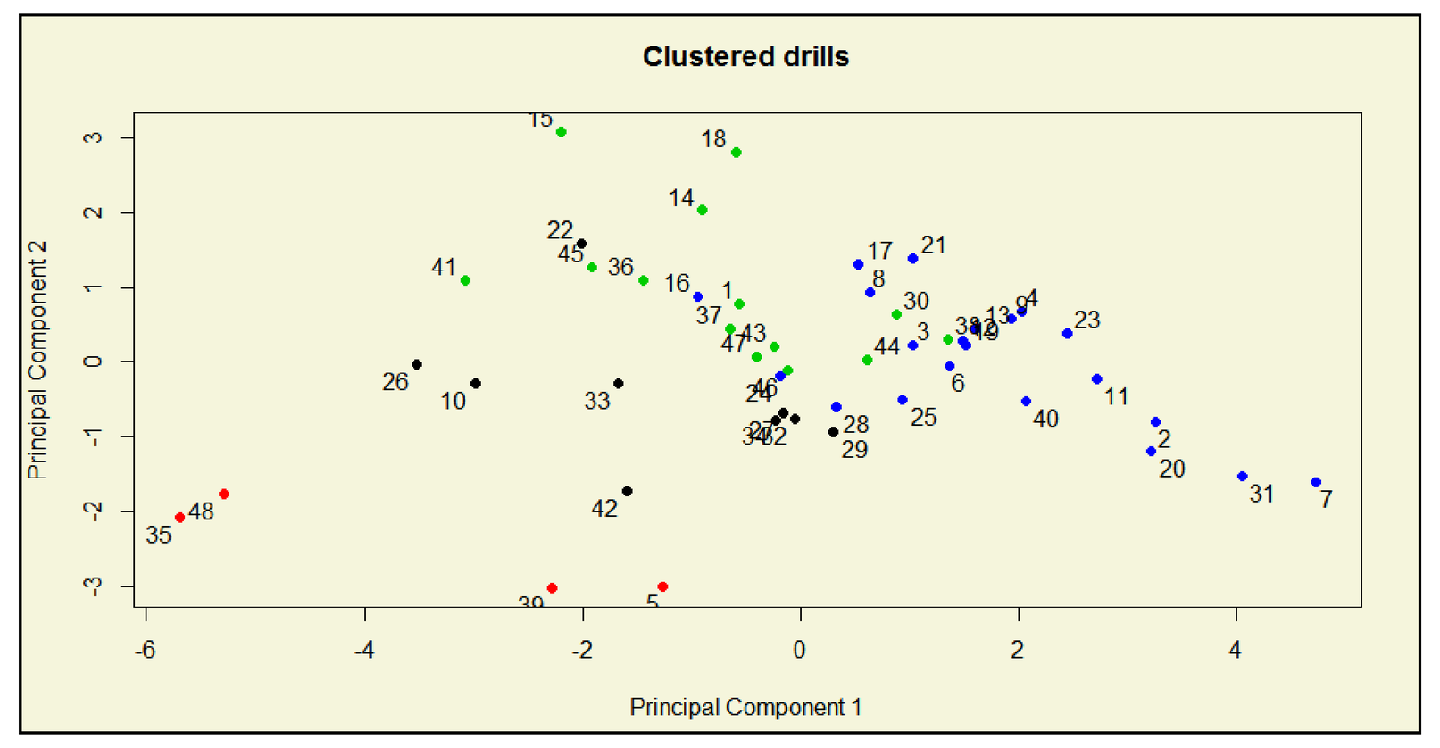
Clusters from result #1: {A,B,C,D}
Clusters from result #2: {a,b,c,d}

Hamming distance used to find best possibility.

A	a	B	b	C	c	D	d
A	a	B	b	C	d	D	c
A	a	B	c	C	b	D	d
A	a	B	c	C	d	D	b
A	a	B	d	C	b	D	c
A	a	B	d	C	c	D	b
A	b	B	a	C	c	D	d
A	b	B	a	C	d	D	c
A	b	B	c	C	a	D	d
A	b	B	c	C	d	D	a
A	b	B	d	C	a	D	c
A	b	B	d	C	c	D	a
A	c	B	a	C	b	D	d
A	c	B	a	C	d	D	b
A	c	B	b	C	a	D	d
A	c	B	b	C	d	D	a
A	c	B	d	C	a	D	b
A	c	B	d	C	b	D	a
A	d	B	a	C	b	D	c
A	d	B	a	C	c	D	b
A	d	B	b	C	a	D	c
A	d	B	b	C	c	D	a
A	d	B	c	C	a	D	b
A	d	B	c	C	b	D	a

Result





Combined results from k-means and spectral clustering of both the six selected features as well as the first six principal components.

Conclusion and Future Work

Conclusion: Fairly consistent clusters were discovered within both drills and players .

Next steps:

Short term: Explore ways to make affinity matrix for spectral clustering more accurate.

Long term: Perform semi-supervised learning with future data.

References

[1] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems, Advances in Neural Information Processing Systems, 2002.*

[2] Shlens, J. (2014). A Tutorial on Principal Component Analysis.

[3] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Acknowledgement

Purdue SURF Program: Sponsorship

Jampani Dwaraknath Reddy: Fruitful discussions