MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

PETER IRUNGU MWANGI

SCT321-C004-2079/2018

MIT 3102 BUSINESS INTELLIGENCE

SPORTS ANALYTICS: REVIEW PAPER

Understanding sport analytics, a general perspective

Sports is well understood term, meaning an activity that involves physical exertion and skill in which an individual or a team competes another (others) for entertainment, while Analytics is the discovery and communication of meaningful patterns in data, in this case from sporting activities.

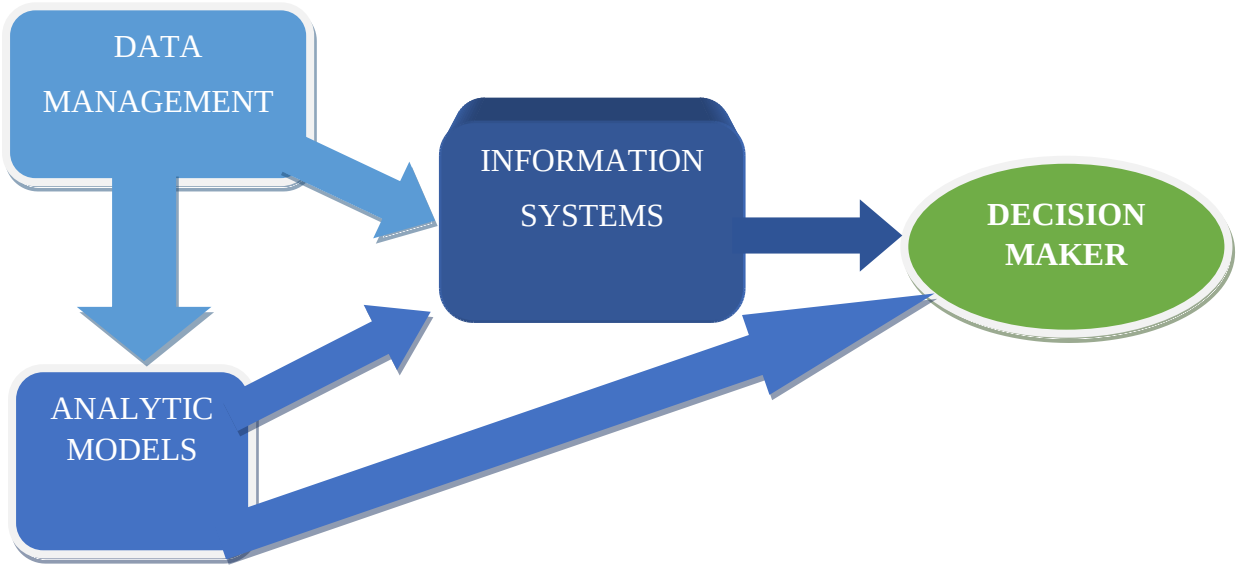
Therefore, Sports analytics are a collection of relevant, historical statistics that when properly applied can provide a competitive advantage to a team or individual. Through the collection and analysing of these data, sports analytics inform players, coaches and other staff in order to facilitate decision making both during and prior to sporting events.

Sports analytics was popularized in mainstream sports culture following the release of the 2011 film, Moneyball, in which Oakland Athletics General Manager relies heavily on the use of analytics to build a competitive team on a minimal budget.

Data analytics and Sports Data analytics is being used by businesses to analyze data and for other purposes but the effect of data analytics can be seen on the sporting platforms as well as other platforms.

Sport analytics been a process of using sports-related data (anything from player statistics to game day weather) to find meaningful patterns (strong correlations, hidden trends, etc.) and communicate those patterns (using graphs, charts, essays, etc.) to help make decisions.

However the definition is both expansive (in the sense that it includes not only statistical models but also the broader information value chain that surrounds these models) and restrictive (this because it excludes traditional analytics applications such as demand forecasting, revenue management and financial modelling, all of which are certainly relevant in the business of professional sports). This can be represented as shown below



A framework for sports analytics

Data management includes any and all processes associated with acquiring, verifying and storing data in an efficient manner. In a sports organization, data can come from a variety of sources and may be presented in many different forms. The data management function will feed both the predictive analytics function and the information systems that support decision-makers.

Predictive analysis, is the process of applying statistical tools to data to gain insight into what is likely to happen in the future. In sports, this can involve the projection of the pro careers of amateur players, identifying how the strengths and weaknesses of an opponent will play out against your own team's strengths and weaknesses, or assessing whether a free agent would fill a need on a team at an appropriate cost.

Information systems, the next component in the framework, are increasingly common in the world of sports. When implemented correctly, such information systems allow for visualization and interactive analysis of relevant information from multiple sources in one place, organized in a meaningful way to provide insights for decision makers.

Decision-makers are the ultimate customers for all components in the sports analytics framework. However, the modern professional sports organization typically has many different decision-makers, including the general manager, coaches, scouts, trainers, salary cap managers and other personnel executives.

(batter|pitcher)2vec: Statistic-Free Talent Modeling With Neural Player Embeddings

1.Introduction

This paper introduces (batter|pitcher)2vec, a neural network algorithm that adapts representation learning concepts (like those found in word2vec) to a baseball setting, modelling player talent by learning to predict the outcome of an at-bat given the context of a specific batter and pitcher. Unlike many Sabermetrics statistics, (batter|pitcher)2vec learns from "raw "baseball events as opposed to aggregate statistics, which allows it to incorporate additional contextual information when modelling player talent.

The field of Sabermetrics was developed in an effort to address some of the inherent limitations of standard baseball statistics. For example, Wins Above Replacement (WAR) "offers an estimate to answer the question, 'If this player got injured and their team had to replace them with a freely available minor leaguer or an AAAA player from their bench, how much value would the team be losing? However, the WAR formula is, itself, somewhat ad hoc, reflecting the intuition of the statistic's designer(s)

$$WAR = \frac{BR + BRR + FR + PA + LA + RR}{RPW}$$

Where BR is batting runs, BRR is base running runs, FR is fielding runs, PA is a positional adjustment, LA is a league adjustment, RR is replacement runs, and RPW is runs per win. Whereas the WAR statistic uses a combination of conventional baseball statistics to quantify an individual player's impact on his team's overall record, the Player Empirical Comparison and Optimization Test Algorithm (PECOTA) forecasts player performance by identifying a neighborhood of players with historically similar statistics (both traditional and of the Sabermetrics variety) and then performing an age correction.

2. Methods

2.1. Data

Play-by-play data for each game from the 2013, 2014, 2015, and 2016 seasons were obtained from the Retrosheet website [12]. Each play description was converted into a tuple consisting of the batter, pitcher, and at-bat outcome; for example, (Mike Trout, Anthony Bass, HR), where HR is the symbol denoting a home run.

On this paper, Michael A. Alcorn, uses two dataset of baseball sourced from http://www.seanlahman.com/baseball-archive/statistics/. The following are the dataset used to develop the model;

Pitching.csv dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	playerID	yearID	stint	teamID	lgID	W	L	G	GS	CG	SHO	SV	IPouts	H	ER	HR	BB	SO	BAOpp	ERA	IBB	WP	HBP	BK	BFP	GF	R
2	bechtge(	1871	1	PH1	NA	1	2	3	3	2	0	0	78	43	23	0	11	1		7.96		7		0	146	0	42
3	brainas0	1871	1	WS3	NA	12	15	30	30	30	0	0	792	361	132	4	37	13		4.5		7		0	1291	0	292
4	fergubo0	1871	1	NY2	NA	0	0	1	0	0	0	0	3	8	3	0	0	0		27		2		0	14	0	9
5	fishech0	1871	1	RC1	NA	4	16	24	24	22	1	0	639	295	103	3	31	15		4.35		20		0	1080	1	257
6	fleetfr01	1871	1	NY2	NA	0	1	1	1	1	0	0	27	20	10	0	3	0		10		0		0	57	0	21
7	flowedi01	1871	1	TRO	NA	0	0	1	0	0	0	0	3	1	0	0	0	0		0		0		0	3	1	0
8	mackdeC	1871	1	RC1	NA	0	1	3	1	1	0	0	39	20	5	0	3	1		3.46		1		0	70	1	30
9	mathebo	1871	1	FW1	NA	6	11	19	19	19	1	0	507	261	97	5	21	17		5.17		15		2	876	0	243
10	mcbridi0	1871	1	PH1	NA	18	5	25	25	25	0	0	666	285	113	3	40	15		4.58		3		0	1053	0	223
11	mcmuljo(	1871	1	TRO	NA	12	15	29	29	28	0	0	747	430	153	4	75	12		5.53		44		0	1334	0	362
12	meyerle0	1871	1	PH1	NA	0	0	1	0	0	0	0	3	1	1	0	2	0		9		1		0	6	1	1
13	paborch(	1871	1	CL1	NA	0	2	7	1	1	0	0	88	50	22	4	6	0		6.75		2		0	160	5	53
14	pinkhed(	1871	1	CH1	NA	1	0	3	0	0	0	1	31	10	4	0	3	0		3.48		5		0	51	3	8
15	prattal01	1871	1	CL1	NA	10	17	28	28	22	0	0	674	296	94	9	47	34		3.77		48		0	1117	1	288
16	spaldal0	1871	1	BS1	NA	19	10	31	31	22	1	0	772	333	96	2	38	23		3.36		11		0	1279	3	272
17	stearbi01	1871	1	WS3	NA	2	0	2	2	2	0	0	54	10	5	0	8	0		2.5		0		0	75	0	11
18	wolteny01	1871	1	NY2	NA	16	16	32	32	31	1	0	849	345	108	7	39	22		3.43		15		0	1349	1	283
19	wrighha0	1871	1	BS1	NA	1	0	9	0	0	0	3	56	34	13	0	4	0		6.27		3		0	105	6	31
20	zettlge01	1871	1	CH1	NA	18	9	28	28	25	0	0	722	298	73	6	25	22		2.73		24		0	1143	0	233
21	bentley0	1872	1	MID	NA	2	15	18	17	14	0	0	432	259	97	4	12	5		6.06		6		0	794	1	253
22	brainas0	1872	1	WS3	NA	2	7	9	9	9	0	0	237	148	56	0	5	1		6.38		12		0	449	0	140
23	brainas0	1872	2	MID	NA	0	2	2	2	1	0	0	24	13	5	1	0	0		5.63		0		0	50	0	17
24	brittji01	1872	1	BR2	NA	9	28	37	37	37	0	0	1008	568	169	6	21	13		4.53		26		1	1762	0	473
25	buttefr01	1872	1	MID	NA	3	2	8	5	5	0	0	177	94	28	1	3	5		4.27		0		0	295	3	78
26	clintji01	1872	1	BR1	NA	0	1	1	1	1	0	0	27	25	7	1	1	1		7		0		0	68	0	36

Batting.csv dataset

playerID	yearID	stint	teamID	lgID	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	I
abercda01	1871	1	TRO	NA	1	4	0	0	0	0	0	0	0	0	0	0	
addybo01	1871	1	RC1	NA	25	118	30	32	6	0	0	13	8	1	4	0	
allisar01	1871	1	CL1	NA	29	137	28	40	4	5	0	19	3	1	2	5	
allisdo01	1871	1	WS3	NA	27	133	28	44	10	2	2	27	1	1	0	2	
ansonca01	1871	1	RC1	NA	25	120	29	39	11	3	0	16	6	2	2	1	
armstbo01	1871	1	FW1	NA	12	49	9	11	2	1	0	5	0	1	0	1	
barkeal01	1871	1	RC1	NA	1	4	0	1	0	0	0	2	0	0	1	0	
barnero01	1871	1	BS1	NA	31	157	66	63	10	9	0	34	11	6	13	1	
barrebi01	1871	1	FW1	NA	1	5	1	1	1	0	0	1	0	0	0	0	
barrofr01	1871	1	BS1	NA	18	86	13	13	2	1	0	11	1	0	0	0	
bassjo01	1871	1	CL1	NA	22	89	18	27	1	10	3	18	0	1	3	4	
battijo01	1871	1	CL1	NA	1	3	0	0	0	0	0	0	0	0	1	0	
bealsto01	1871	1	WS3	NA	10	36	6	7	0	0	0	1	2	0	2	0	
beaveed0	1871	1	TRO	NA	3	15	7	6	0	0	0	5	2	0	0	0	
bechtge01	1871	1	PH1	NA	20	94	24	33	9	1	1	21	4	0	2	2	
bellast01	1871	1	TRO	NA	29	128	26	32	3	3	0	23	4	4	9	2	
berkena0	1871	1	PH1	NA	1	4	0	0	0	0	0	0	0	0	0	3	
berryto01	1871	1	PH1	NA	1	4	0	1	0	0	0	0	0	0	0	0	
berthha01	1871	1	WS3	NA	17	73	17	17	1	1	0	8	3	1	4	2	
biermch0	1871	1	FW1	NA	1	2	0	0	0	0	0	0	0	0	1	0	

2.2. Model

The (batter|pitcher)2vec model takes one-hot encodings of a batter and pitcher as input and then selects the corresponding player weights from the batter and pitcher weight matrices, respectively. The player weight vectors are then passed through a "sigmoid"/logistic activation function.

$$w_p = \sigma(W_p \cdot h_p)$$

$$w_b = \sigma(W_b \cdot h_b)$$

Where, h_b is the NB-dimensional one-hot vector (where NB is the number of batters) for the batter indexed by b, WB is the batter embedding matrix, σ is the logistic activation function and w_b is the batter's embedding. Likewise, h_p is the NP-dimensional one-hot vector for the pitcher indexed by p, WP is the pitcher embedding matrix, and w_p is the pitchers embedding.

Maximizing the likelihood of this model is equivalent to minimizing the model's average cross entropy (7), which is the standard loss function used in machine learning classification tasks:

$$\mathcal{L}(D) = \frac{1}{N}\sum_{i=1}^{N} H(p_i, q_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_i(o_j)\log(q_i(o_j))$$

Where Dis the training data, Nis the number of training samples, H (p_i ,q_i)is the cross entropy between the probability distributions pi and q_i, p_i is the true at-bat outcome distribution for training sample i, and qi is the predicted outcome distribution for training sample i.

3. Results

3.1. Visual inspection of the player embedding's

Visually inspecting low-dimensional approximations of neural network representations can often provide some intuition for what the model learned. A number of trends are readily apparent; for example, left-handed hitters are clearly distinguishable from right-handed hitters, and batterswith high single rates are antipodal to batters with low single rates (a similar pattern is visible for home run rates). At the least, (batter|pitcher)2vec appears to be capable of capturing the information contained in standard baseball statistics.

3.2. Nearest neighbors

Probing the neighborhoods of individual embeddings can also yield deeper insights about the model. The t-SNE algorithm was used to map the batter and pitcher embeddings into two dimensions so that they could be visualized. Intriguing player clusters are readily apparent, with close pairs including: Mike Trout/Paul Goldschmidt, Dee Gordon/Ichiro Suzuki, and Aroldis Chapman/Dellin Betances

When calculating nearest neighbors in the learned embedding space, Paul Goldschmidt is indeed Mike Trout's nearest neighbor; an unsurprising result considering how each athlete is known for his rare blend of speed and power. Similarly, Ichiro Suzuki is Dee Gordon's nearest neighbor, which is to be expected as both have a reputation for being able to get on base [16]. Notably, when clustering players on common MLB stats (e.g., HRs, RBIs), Paul Goldschmidt is not among Mike Trout's ten nearest neighbors, nor is Ichiro Suzuki among Dee Gordon's ten nearest neighbors.

3.3. Modelling previously unseen at-bat matchups

The representations learned by neural networks are theoretically interesting because they suggest the neural networks are discovering causal processes when the models are able to generalize (or transfer) well. In the case of (batter|pitcher)2vec, the ability to accurately model at-bat outcome probability distributions for previously unseen batter/pitcher pairs would indicate the neural network was extracting important aspects of baseball talent during learning. To test this hypothesis, at-bat outcomes were collected from the 2016 season for previously unseen matchups that included batters and pitchers from the training set. In all, there were 21,479 previously unseen matchups corresponding to 51,580 at-bats.

Future directions

These results prove neural embedding algorithms offer a principled means of modelling talent from “raw data”, i.e., without resorting to ad hoc statistics. Just as pre-trained word embeddings can be used to improve the performance of models in various natural language processing tasks, player embeddings could be used to better inform baseball strategy. For example, by swapping the embeddings of players in a proposed trade and “back simulating “games from earlier in the season, teams would be able to assess how many more wins (or losses) they would have obtained with the candidate player(s) on the roster (effectively establishing a counterfactual). Likewise, after first applying (batter|pitcher)2vector minor league baseball players, a second model could be trained that learns to map a player's minor

league representation to his MLB representation. Such a model would allow teams to scout prospects by surveying their neighboring MLB players in the mapped space (this framework is conceptually similar to the multimodal model described in, which learns a map between audio and video representations).

1. Lindsey, G. R. “Statistical Data Useful for the Operation of a Baseball Team,” *Operations Research,* Vol. 7, No. 2, March-April 1959, pp. 197-207.
2. www.amazon.com/Moneyball-Art-Winning-Unfair-Game/dp/0393057658.
3. What is WAR? URL:http://www.fangraphs.com/library/misc/war/(visited on 21/01/2019).
4. Joseph Kuehn. “Accounting for Complementary Skill Sets When Evaluating NBA Players’ Valuesto a Specific Team”.In: MIT Sloan Sports Analytics Conference(2016)
5. Retrosheet Event Files. URL:http://www.retrosheet.org/game.htm (visited on 21/01/2019).
6. Jonathan Judge. DRA: An In-Depth Discussion. 2015. URL:https://www.baseballprospectus.com/news/article/26196/prospectus-feature-dra-an-in-depth-discussion/ (visited on 21/01/2019).
7. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: Neural Information Processing Systems (2013), pp. 1–9. URL:https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
8. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. “A Neural Probabilistic Language Model”. In: The Journal of Machine Learning Research3 (2003), pp. 1137–1155. URL:http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.
9. Sean R. Hackett and John D. Storey. “Mixed Membership Martial Arts: Data-Driven Analysis of Winning Martial Arts Styles”. In: MIT Sloan Sports Analytics Conference (2017), pp. 1–17. URL:http://www.sloansportsconference.com/wp-content/uploads/2017/02/1575.pdf.
10. Oliver Schulte and Zeyu Zhao. “Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact”. In: MIT Sloan Sports Analytics Conference (2017). URL:http://www.sloansportsconference.com/wp-content/uploads/2017/02/1625.pdf.
11. Min-Hwan Oh, Suraj Keshri, and Garud Iyengar. “Graphical Model for Basketball Match Simulation”. In: MIT Sloan Sports Analytics Conference (2015).

URL:http://www.sloansportsconference.com/wp-content/uploads/2015/02/SSAC15-RP-Finalist-Graphical-model-for-basketball-match-simulation.pdf.

12. François Chollet. Keras. 2015. URL:https://github.com/fchollet/keras.[14] L J P van der Maaten and G E Hinton. “Visualizing High-Dimensional Data Using t-SNE”. In: Journalof Machine Learning Research9 (2008), pp. 2579–2605. URL:https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf.
13. Ian Goodfellow, Yoshua Bengio, and Aaaron Courville. “Representation Learning”. In: Deep Learning. MIT Press, 2016. Chap. 15, pp. 526–557. URL:http://www.deeplearningbook.org/contents/representation.html.
14. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. “Multimodal Deep Learning”. In: Proceedings of the 28th International Conference on Machine Learning (ICML) (2011), pp. 689–696. URL:https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf.
15. Quoc V. Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents”. In: International Conference on Machine Learning -ICML2014 32 (May 2014), pp. 1188–1196. URL:http://arxiv.org/abs/1405.4053.
16. Michael A. Alcorn. Learning to Coach Football. 2016. URL:https://sites.google.com/view/michaelaalcorn/blog/learning-to-coach-football(visited on 21/01/2019).[28] Hoang M. Le, Peter Carr, Yisong Yue, and Patrick Lucey. “Data-Driven Ghosting Using Deep Imitation Learning”. In: MIT Sloan Sports Analytics Conference (2017). URL:http://www.sloansportsconference.com/wp-content/uploads/2017/02/1671-2.pdf.
17. Kuan-ChiehWang and Richard Zemel. “Classifying NBA Offensive Plays Using Neural Networks”. In: MIT Sloan Sports Analytics Conference (2016). URL:http://www.sloansportsconference.com/wp-content/uploads/2016/02/1536-Classifying-NBA-Offensive-Plays-Using-Neural-Networks.pdf.