Cricinfo Analysis to find patterns in profiles for Predicting future Captain

Muhammad Kashif Faiz[1], Dr. Malik Muhammad Saad Missen[2]
[1]Department of Computer Science and Information Technology, Islamia University Bahawalpur, Pakistan.
[2]Department of Computer Science and Information Technology, Islamia University Bahawalpur, Pakistan.
mkfaiz@gmail.com

ABSTRACT

The selection of captain is dependent on the performance of players in terms of various parameters in their areas of expertise. In this thesis we propose a method for measuring qualitative parameters in existing captains' profile and by developing an ideal profile, we set a threshold, that can be applied any of the player for predicting about whether he can be next captain or not. A real dataset of more than 76 Captain of different countries from Cricinfo is considered for analysis purpose. The findings from our research can be used to improve the quality of information retrieval and decision making. Cricinfo is a large database and having information about large number of players estimated about 50000 players from all over the world. So it can be considerable for different types of experiments and analysis. Our aim in this research work is to search and forecast about behaviors of existing captain profile's data and by building an ideal profile, forecasting about other players, where any other player is suitable for captainship or not.

Keywords: Cricket, Strategy, Ranking, Prediction, qualitative analysis

1. Introduction

Expert System are used to solve problem as a human can do. But there are reasons where it is desirable to use expert system rather to use human expert like availability of human expert is limited while in contrast expert system is always available. Human expert are available in locality while expert system can be anywhere. Other consideration like safety consideration, durability, performance are much better than a human expert. But there are some cons also of an expert system like Learning ability of expert system is very low and explanation are exact.
Before proceed further, first it is desirable to describe how an expert system work. Every expert system have domain which contains the domain knowledge stored in knowledge base, when some case is to be solved, Inferred fact are recalled and stored in working memory, then inference engine is used to conclude the results.
In this research we are developing an expert system to predicate future captain by analyzing profiles of different existing captain and then extracting an ideal profile, we can apply this profile to different players to find captains.
As information flood is due to the electronic computing, so a lot of researches are applied in this sense to provide senses to machine, so machine can itself decide relevant information's, for this purpose different techniques were adopted and continually adopting for machine learning
Due to that, a large number of people involves and are interested in cricket field, so for researchers it is well suited for different type of analysis.
e.g.: in Cricket game, as Selector of team and other committee members have problems how to decide about future captain, as it is not a straight forward decision. Different points must be considered because a captain has some sort of qualities that can make him a captain. If we don't

care such things, an unsuitable captain can dangerous for whole team, because a match totally relies on captain's decisions.

1.1 Our Objective

In this research we analyze different profiles of existing captains and then the keywords used in these profile, we will build an IDEAL profile. After building this ideal profile we will check this ideal profile with existing captain, that leads us to the finding minimum and maximum points, which will be used to set our threshold.
After threshold definition, now we are in position when we can put any player profile and check, if that player fits in that threshold, then he can be the captain.

Here, we have questions like

- How the profile keywords will be gathered?
- How we check which keyword is for our interest
- How to check negative or positive skew-ness of the keyword
- How much keywords are common in most of the captains?
- Data is not in quantitative but in a qualitative form, how we decide and use it for our purpose.
- How ideal profile will build.
- How threshold will be set

For profile keyword gathering, Cricinfo website will be used to gather different existing captain's profile, because Cricinfo is the largest dataset available at the time and containing about 50000 player's data, both new and old, and it is increasing day by day and updated. So the result we gathered from here will be more reliable and judge able by any non-scholar person also.

As profile is build using English sentences, so all rules of English are applied, use of nouns, pronouns , adjectives, verbs are used. Now from here only adjective and verb will be separated. How these will be separated, here we get help from any programming language and define an algorithm that will help in extracting the keywords of our demand. We will store these keywords in our created database.

After gathering all keywords of our interest, next step is to find, whether the keyword have positive sense or negative sense. Because this would a huge impact in building our profile and it is crucial part in this research also. All our success in building a good profile will be based on this part. For this we will get help from on net software like senti-words that will provide not only the meaning of the keyword but also provides us the skewness of the keyword also. At next step, time to find how much keywords are common in most of the captain's data, why this needed, for this purpose PCA will be applied to extract most common words.

As we have no quantitate form of data, we are building all of our descriptive profiles of player, so normal statistics rules cannot apply on such type of data. We have to separate verbs, adjectives from the profiles, their common meaning will be considered and finally we can conclude and can make different type of decisions.

To build an ideal profile, first we build a vocabulary of words and then those words will be considered that are positive meanings and have common among most of the players and we will set a threshold level by comparing all captain's profile

2. Literature Review

As today's world is said to the world of information, and data is becoming more and more huge and complex day by day. So different type of analysis can be performed on this data. As my chosen work is basically based on information retrieval and finding patterns and then making some decision on this data. The work we found on internet is related to statistics and logic was discussed in computer science field.

Hermanus H.LEMMER et.al: describe about the player performance of twenty20 cricket. The authors mainly focuses on Twenty20 cricket and normally when there are not enough matches played. The authors derived a new formula for such situations, their formula ranked the players either they are batsman or bowler. Their all measurement is only on twenty20 cricket. They measure the performance in case of small number of matches. According to the authors, if ordinary measurement will be used in small number of matches, then we can't get fruitful results.[4]

Muhammad Daniyal et al: main focuses on batting performance using Moving Range Control Chart and Individual player ranking. They choose some tap ranked players and after performing different types of calculations and analysis, they compare the performance of players.[5]

Anada B.W.Manage et al: in their paper named "An Introductory Application of Principal Components to Cricket Data" focuses on IPL Competition. IPL is new cricket that is specially played in India and said to be Indian Premier League (IPL). The authors perform component analysis on sports data. Specially, the authors discuss the application of PC (Principal Components) in cricket players ranking. [6]

Phiip Scard et al: Normally focus on follow on decision in test cricket. Their main focus is on the results that are calculated from the end of first, second and third innings positions. The author's collected 391 test matches records from 1997 to 2007 and after performing analysis and calculations, They describe how different strategy about declaration of players vary from one innings to second and how the nature and strength of covariate effects each other. According to the authors, as matches' progress, their defined variable also increases from 44 percent to 80 percent at the end of third innings. Authors also mentions the follow on decision problem in detail [7]

Faez Ahmed and co-authors focuses in their research about player's selection in cricket by reducing the budget. According to the authors, it is not so straight to make such decisions and complex calculations are needed for making decisions. Bowling and bating are two key points in cricket, and these are two major points also. We can't judge player on single base, there is need a trade of in performance in bowling and batting for a good decision. A new scheme named "Novel Representation" was introduced and multi objective approach was also used based on NSGA-II algorithm. This new formulation is also introduced be the authors of this research. They suggest a multi criteria approach for team selection. The work method is simple and generic and can be easily implemented in other sports like baseball, soccer and like games. [8]

Dr. Parag Shah and co-authors introduces new formulation named "Pressure Index" for player's evaluation in terms of performance. As the authors, there are limitations in describing the performance and abilities that will based on strike, economy rates and bowling averages. They derived new measure called pressure index, which measure the pressure under batsman is batting and under which pressure a team is playing [9]

Gurshan Singh, et al: This is the paper that is based on measure using artificial intelligence. A software tool introduced by the authors. Different parameters were considered for ranking the player. Fuzzy logic is the main concern of the authors and on this base, they calculate the performance of players. [10]

Harsimranjeet Singh et al: they calculate the performance on the basis of cricket balls and bats under dynamic conditions. Authors measure the hardness and elasticity of the balls of cricket as function for calculating incoming speed.[11]

MD Shamshoddin Altamash et al: describes in their paper about data mining and how to find huge hidden pattern from huge data. They use data mining as tool to select player. Their analysis is dynamic and they used association rule. Their analysis reveals the performance of Indian cricketers. They also suggest that this same methodology can be applied for other cricket players[12].

Mastoli manjiri mahadev et al: presents election expert system through fuzzy logic. Their core research is based on to develop a rule based fuzzy expert system by which prediction of result of election can be done by the system.[13]

Gursharn Singh et al: In their publication, performance evaluation is a critical issue. Vague and imprecise are two main parameters for their study. The authors suggests a fuzzy cognitive map based player performance. They develop a tool that use this logic to perform different types of computation taking care of the parameters used. They also present their model in Simple Graphical User Interface.[14]

Kiri L.Wagstaff : his paper is based on Machine learning, in which , he describe different new paradigms to use in ML (Machine Learning). The Authors present six impact challenges to focus the field attention and energy. Their aim is to inspire ongoing focus and discussion on Machine learning. [15]

Pedro Domingos figure out to do different important tasks by generalizing from examples. They describe that Machine learning is cost-effective solution where manual programming cannot be used. According to author, for a successful machine learning requires enough amount of Black Art that is hard to find in textbooks. Authors summaries his paper by twelve key lessons that new researcher have to learn about machine learning. [16]

Ashwini Umarikar describes that fuzzy logic is one of the most popular technologies and used in technology from automobile control to medical sciences. The purpose of the paper is to present fuzzy logic based control and difference from conventional control. This paper also presents an overview of practically use of fuzzy control.[17]

Some of the studies in the similar direction can be found in the articles that are referenced in reference section of this thesis

3. Methodology

As, we are especially work on qualitative nature of data, our approach must be different as either applied approaches that are based on quantities data. First we discuss how we got this data and further described about the software that are helpful in calculating our findings.

3.1 Data Collection:

Cricinfo is our main focus and this website is a collection of thousands of player's records about their bating and bowling. But described earlier our main focus will not be of these two areas. Our main focus is on the data that is provided in profiles of players. This data is about the history, nature, player positive points and negative points like a little story of every player. As thousands of records are available on Cricinfo, we randomly select 76 captain for analysis purpose.

As we have to judge our ideal profile, a data of 10 non-captain players was also collected from the site. Which will be used for testing after building the ideal profile.

4. Experimental Work

After getting data of 76 captain from Cricinfo website, the first thing we have to do with that data is to find the verbs, nouns from these profiles, for this purpose, we use programming to split these profile's data into words and then applying Wordnet 3.0 to find the negative or positive ness of the word. When we separated the terms and applied Wordnet on these profile, following data was obtained

Table -1:Captain Profiles Words with Positive Words in their Profiles

Sr #	Captain Name	Total Words in Profile	Positive Words in Profile
1	Mohammad Hafeez	309	15
2	Salman Butt	548	33
3	Waqar Younis Maitla	406	20
4	Saleem Malik	148	10
5	Sahibzada Mohammad Shahid Khan Afridi	354	14
6	Misbah-Ul-Haq Khan Niazi	347	14
7	Shoaib Malik	303	16
8	Wasim Akram	181	11
9	Syed Zaheer Abbas Kirmani	723	49
10	Intikhab Alam Khan	188	8
11	Saeed Anwar	189	16
12	Asif Iqbal Razvi	95	7
13	Imran Khan Niazi	308	20
14	Mohammad Javed Miandad Khan	585	34
15	Inzamam-Ul-Haq	682	22
16	Majid Jahangir Khan	168	10
17	Mohammad Moin Khan	180	9
18	Mushtaq Mohammad	137	9
19	Abdul Qadir Khan	144	10
20	Ramiz Hasan Raja	269	24
21	Rashid Latif	294	16
22	Mohammad Aamer Sohail Ali	151	9
23	Wasim Bari	359	29
24	Hanif Mohammad	134	7
25	Fazal Mahmood	471	20
26	Abdul Hafeez Kardar	422	28
27	Imtiaz Ahmed	116	11
28	David William Gregory	483	24

29	George Giffen	985	35
30	Clement Hill	906	45
31	Montague Alfred Noble	1598	68
32	Herbert Leslie Collins	266	15
33	Warren Bardsley	594	29
34	Victor York Richardson	309	12
35	Donald George Bradman	1025	53
36	Arthur Robert Morris	128	8
37	Raymond Russell Lindwall	983	54
38	Ian David Craig	186	11
39	Richard Benaud	175	11
40	Robert Neil Harvey	328	19
41	Ian Michael Chappell	131	8
42	Gregory Stephen Chappell	209	8
43	Allan Robert Border	266	9
44	Mark Anthony Taylor	182	14
45	Stephen Rodger Waugh	354	20
46	Adam Craig Gilchrist	570	26
47	Ricky Thomas Ponting	437	26
48	Michael John Clarke	354	33
49	Stephen Fleming	512	22
50	Arjuna Ranatunga	217	7
51	Muhammad Azharuddin	239	12
52	Ms Dhoni	585	28
53	Aarvinda De Silva	146	10
54	Hansie Cronje	2961	118
55	Brian Charles Lara	758	28
56	Sachin Tendulkar	554	29
57	Graeme Smith	448	28
58	Darren Sammy	511	19
59	Martin David Crowe	256	19
60	David Laud Houghton	124	8
61	Alistair Douglas Ross Campbell	325	26
62	Mahela Jayawardene	453	30
63	Kumar Sangakkara	425	25
64	Angelo Mathews	296	16
65	Mohammad Mushfiqur Rahim	433	17
66	Anil Kumble	507	24
67	Jimmy Adams	240	12
68	Graham Alan Gooch	314	11
69	Ian Botham	367	20
70	Clive Lloyd	1207	59
71	Andrew Strauss	593	22
72	Rahul Dravid	509	26
73	Sourav Ganguly	514	28
74	Alec Stewart	352	18
75	Stuart Broad	1025	45
76	Ab De Villiers	423	28

In this table, all gathered data of captain are summarized in the form of total words in the profile of each player and the positive words provided in each profile, for finding positive words in profile, Wordnet helped us in this respect, because word net is categorized in such manner.

Using Wordnet, we find the words that are positive and ignore all the neutral and negative words

Now, we got all the positive words to build an ideal profile, the words collected from the profiles that have positive polarity in all the captain's profiles, but the words selected us have to further reduce to get 100% pure words used by our ideal profile

In our experiment or data analysis words collected from different profiles was used as factors and we provides points to each word according to number of occurrence a word appear in each profile.

PCA in SPSS provides two approaches to extract components

4.1 Based on Eigenvalues

In this approach, after analysis, only those factors are chosen that have eigenvalue greater that the value provided by researcher. Default value for Eigen value is 1(one).

4.2 Fixed Number of Factors

In this approach, we can specify number of factors by ourselves. This is suitable, when number of factors returned by the PCA are less or greater than our expected results.

In our analysis, when we analyze our data based on First Approach (ie. Based on Eigenvalue) provided too few components.

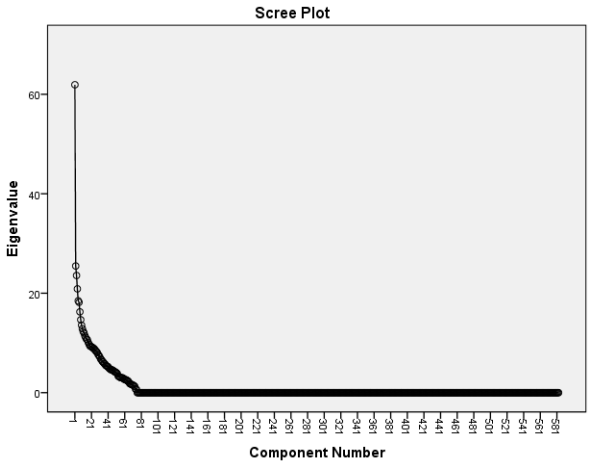
Only 73 Components returned by this approach, which doesn't cover the existing variables or generally speaking, it skips the number of players.

As we have 76 players, when these extraction is applied to the positive words of 76 players, we have to miss 9 players, only 67 players qualify on the components or factors returned by this approach

So far, we used second approach, which is suitable for our analysis, we provided different ranges, ie. 100 Components Extraction was used, it also skips the players (captains), 150 Components Extraction was used. It repeats the same, so we applied 250 components extraction, it remain sufficient, and all captain were covered in this approach.

After Putting all the data in SPSS and then by reducing terms according to the second approach, when number of fixed factors set to 250, following outcomes obtained by SPSS

Figure 1: Scree Plot



4.3 IDEAL PROFILE

Table 2: Ideal Profile

Ideal Profile Words				
love	fit	productive	compliant	doubt
first-class	proper	successful	hero	novel
good	choice	utter	Free	satisfy
like	comfortable	clever	physical	liked
praise	crucial	Pacific	similar	opposing
splendid	respect	correct	Justice	green
dedication	wonderful	master	received	tell
benefit	break	catching	off	wish
make	delightful	hope	times	fast
great	admiration	beat	bring	presence
firm	surpassing	paid	arrogance	learn
integrity	magnanimous	everlasting	save	easily
greatest	agog	deep	weakness	modern
important	formidable	tribute	critical	fame
artistic	intriguing	must	admission	flamboyant
character	fatal	moving	higher	fortuitous
nice	prepared	desired	vulnerable	formative
elegance	adventurous	vice	response	chance
majestic	fateful	maestro	Mark	amazing
right	manifest	profound	opponent	failing
worthy	tidy	class	old	commitment
clean	education	glorious	aggressive	unique
brilliant	contrived	sheer	technical	swing
consider	young	major	some	obvious
receive	encouraging	highly	help	desire
corruption	humour	add	accepted	marked
intelligent	different	disparate	very	deserved
think	interested	admire	run	lean
finished	supporting	frustrating	rate	double
easy	accept	obliged	known	High

question	appeal	calibre	one	fashionable
importance	fitness	skilful	still	active
time	certain	influenza	personal	matched
well	first	clear	approach	backward
favourable	hit	straight	triumph	powerful
maintained	groomed	indifferent	perception	grace
patience	apparent	dour	camp	back
superior	becoming	Grey	rising	decided
established	Taking	imposing	shrewd	carry
complete	Square	revered	Darling	mobility
accomplished	Testimonial	hammered	dramatic	talented
simply	Noble	lure	enterprising	secure
smart	finish	practical	astute	hold
beautiful	graphic	daily	settled	improved
better	popularity	actual	prove	check
promising	recovered	immunity	rare	hand
devious	credit	Commission	astonishing	ruined
emphasis	pleased	directed	show	catch
hot	sure	inept	reputation	season
amazed	winning	arrogant	point	sharing

4.4 RESULTS:

As each player or captain profile having different length, we have to equalize the number of positive words.

For this, we used %age:

Points = PositiveWords in Profile *100/ TotalNumberOfWordsinProfile (1)

From the above tabular data, we can built threshold value, the minimum value found in the captain data is

Min Value: 33.333

Max Value: 93.878

And threshold is defined as Minimum value i.e.: **33.333**

4.5 Testing

After building threshold, now it's time to test our result with non-captain data, for this purpose we select 15 random players profiles, the data is as follows

After applying out ideal profile to this selection we got the following point chart, giving us the point of each player with relevant to the ideal profile

Table 3: Player Data (Non-Captain)

Player Name	Words According to Ideal Profile	Positive Words	Points
Balapuwaduge Ajantha Winslow Mendis	4	20	20.000
Sean Colin Williams	10	31	32.258
Steven Smith	8	22	36.364
Azhar Ali	4	11	36.364
Paul Adams	7	16	43.750
Adam Voges	17	36	47.222
Usman Khawaja	6	12	50.000
Mohammad Amir	9	18	50.000
Michael John Lumb	13	26	50.000
Kemar Andre Jamal Roach	5	8	62.500
Hettige Don Rumesh Lahiru Thirimanne	9	14	64.286
Pragyan Ojha	6	9	66.667
Rilee Roscoe Rossouw	6	9	66.667
Umar Akmal	6	8	75.000
Hasan Raza	5	6	83.333

4.6 Results and Analysis

From above data, it can see clearly that, **Balapuwaduge Ajantha Winslow Mendis**and **Sean Colin Williams**does not qualify the ideal profile, and now we can judge that he cannot be the captain, and the remaining one are can be the captain, but they just touch the threshold value, which shows that these players can be the average captain with average capability.

5. Conclusion

As in our experiment, we gather qualitative data of different captain, and by finding words that are common in most of profiles and have positive tendency are chosen, by applying these measure we built an ideal profile, when applied to our sample captain's data, we got threshold value that provide us the minimum and maximum score / points that can be suitable for choosing captain

5.1 Future Work

1. We recommend that our research can be helpful in fields of decision making and also in machine learning, by which system can analyze any data provided to it, can extract words that are common and positive behavior

2. We also suggest Cricinfo can also apply this technique to predict about a player for suitable captain
3. This research will also be helpful in predicting behavior of English syntax, which can be helpful in social network sites to restrict abusing or negative words.

6. References

1. Players Information (2015, Apr10). Retrieved from http://www.espncricinfo.com

2. CRICINFO (2015a). Statsguru [http://stat.cricinfo.com/guru?sdb=find&search=]. Retrieved on Apr2015.

3. CRICINFO (2014b). ICC World Twenty20 [http://contentrsa.cricinfo.com/twenry20wc/engine/current/match/html]. Retrieved on Dec-2014.

4. Lemmer, H. H. (2008). An analysis of players' performances in the first cricket Twenty20 world cup series. South African Journal for Research in Sport, Physical Education and Recreation, 30(2), 71-77.

5. Daniyal, M., Nawaz, T., Mubeen, I., & Aleem, M. (2009). Analysis of Batting Performance in Cricket Using Individual and Moving Range (mr) Control Charts. Journal of Quantitative Analysis in Sports, 5(3), 5-5.

6. Manage, A. B., & Scariano, S. M. (2013). An Introductory Application of Principal Components to Cricket Data. Journal of Statistics Education, 21(3).

7. Scarf, P., & Akhtar, S. (2011). An analysis of strategy in the first three innings in test cricket: declaration and the follow-on. Journal of the Operational Research Society, 62(11), 1931-1940.

8. Ahmed, F., Deb, K., & Jindal, A. (2011). Evolutionary multi-objective optimization and decision making approaches to cricket team selection. In Proceedings of the Second International Conference on Swarm, Evolutionary, and Memetic Computing. Berlin, Heidelberg: Springer-Verlag http://dx. doi. org/10.1007/978-3-642-27242-4_9.

9. Shah, P., & Shah, M. Pressure Index in Cricket.

10. Hayward, G., & Davidson, V. (2003). Fuzzy logic applications. Analyst, 128(11), 1304-1306.

11. Al-Odienat, A. I., & Al-Lawama, A. A. (2008). The advantages of PID fuzzy controllers over the conventional types. American Journal of Applied Sciences, 5(6), 653-658.

12. Singh, G., Bhatia, N., & Singh, S. (2011). Fuzzy Logic based Cricket Player Performance Evaluator. IJCA Special Issue on "Artificial Intelligence Techniques-Novel Approaches & Practical Applications, 11-16.

13. Vasičkaninová, A., Bakošová, M., & Karšaiová, M. (2011). Neuro-fuzzy Control of a Chemical Reactor with Uncertainties. In Proceedings of the 18th International Conference on Process Control, Slovak University of Technology in Bratislava, Tatranská Lomnica, Slovakia (pp. 360-365).

14. Singh, H., Singh, G., & Bhatia, N. (2012). Election Results Prediction System based on Fuzzy Logic. International Journal of Computer Applications, 53(9), 30-37.

15. MAHADEV, M. M., & KULKARNI, R. (2013). A REVIEW: ROLE OF FUZZY EXPERT SYSTEM FOR PREDICTION OF ELECTION RESULTS. Reviews of Literature• Volume, 1(2).

16. Hussein, H. M., & Yakunin, A. G. SHORT TERM FORECASTING FOR AIR TEMPERATURE BASED ON PATTERN REPETITION.

17. Dubey, A. K. (2009). Performance Optimization Control of ECH using Fuzzy Inference Application. Journal of Advanced Mechanical Design, Systems, and Manufacturing, 3(1), 22-34.

18. Parker, D., Burns, P., & Natarajan, H. (2008). Player valuations in the Indian premier league. Frontier Economics, 1-17.

19. Emerson, E. (2001). Challenging behaviour: Analysis and intervention in people with severe learning disabilities. Cambridge University Press.

20. Magee, J., Kramer, J., & Giannakopoulou, D. (1999). Behaviour analysis of software architectures. Software Architecture, 35-49.

21. Knuth, D. E., Morris, Jr, J. H., & Pratt, V. R. (1977). Fast pattern matching in strings. SIAM journal on computing, 6(2), 323-350.

22. Singh, G., Bhatia, N., & Singh, S. (2011). Fuzzy Logic based Cricket Player Performance Evaluator. IJCA Special Issue on "Artificial Intelligence Techniques-Novel Approaches & Practical Applications, 11-16.