P-ISSN: 2394-1685
E-ISSN: 2394-1693
Impact Factor (ISRA): 5.38
IJPESH 2016; 3(5): 513-518
© 2016 IJPESH
www.kheljournal.com
Received: 26-07-2016
Accepted: 28-08-2016

Prasant Nair
Former Student, Jamshedpur
Co-Operative College,
Jharkhand, India

Correspondence
Prasant Nair
Former Student, Jamshedpur
Co-Operative College,
Jharkhand, India

Editing the worm graph alias ball by ball data and predicting the winner for cricket

Prasant Nair

Abstract
Data mining means extraction of hidden predictive information from large databases. And data visualization places the extracted data in a visual context to bring out patterns, trends, and correlations that might go undetected in text-based data. The cricket data, whether the ball by ball data or the over by over data of any match is large and lack predictive value in its raw state. The worm graph turns this data into a visual form of cumulative runs and fall of wickets. But again, it lacks predictive value. This paper prepares an easy predictive model for cricket which any student, housewife, teenager or old age person can try himself. Under this In-play method, you remove the 'nines' from the current score; then the same worm now zooms out (during first innings itself) which team will win the match? This is a special tribute to William Playfair, the inventor of the time series line chart.

Keywords: Predictive model cricket, in-play model, casting out nines, real time score, ball by ball data

1. Introduction

In any predictive model for cricket (or for any other fields), while the output of the model would comprise the 'score' or 'match result' prediction, the input comprises past or live score i.e. real time data on the basis of which the model propounds the prediction. For developing the model, we walk across the stages of data mining, visualization, predictive analytics, etc. Data mining means extraction of patterns and knowledge from large amounts of data. Data mining derives the name from its similarities with mining a mountain for a vein of valuable ore. Either we sift through the entire data, or else, intelligently probe it to spot only the useful hints. Data mining parameters include Association, Sequence or path analysis, Classification, Clustering, Forecasting [1]. Data visualization is a graphical display of data by placing it in a visual context. Patterns, trends, and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Predictive analytics is the practice of extracting information from existing data sets to determine patterns and predict future outcomes and trends. Predictive analytics does not tell you what will happen, rather, what might happen with an acceptable level of reliability. With the aid of predictive models, we see the probable future in the present data. In a natural way, it is inclined towards the use of 'probability function' that we have learned in high school levels. In the advanced level, often we find the extensive use of Bayesian probability in predictive modeling.

Visualization or infographics mean visual graph based on mathematics and graphics which includes pictorials. Representing quantitative information graphically began in the 17th century when the French philosopher and mathematician Rene Descartes developed a two-dimensional coordinate system for displaying values, comprising a horizontal axis for one variable and a vertical axis of another, primarily as a graphical means of performing mathematical operations. In late 18th century, Scotsman William Playfair [2] initiated the use of a line moving up and down as it progressed from left to right to show how values changed through time. Playfair extended the potential of graphics. He invented the bar graph and the pie chart. Playfair understood that statistical charts could assist human information processing by reducing demands on attention, working memory and long-term memory. He had a perfect feel of perception and cognition. Playfair's charts are constructed in such a manner that comparisons in different domains (lines, colors, labels, etc.) do not exceed attentional and

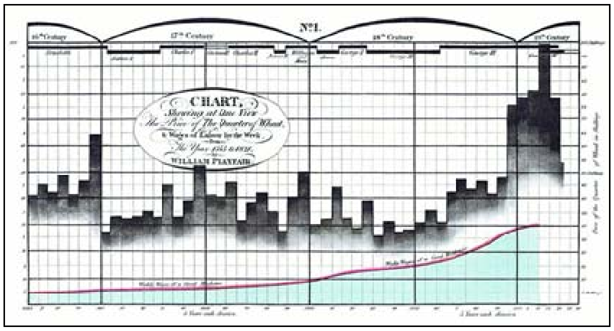

Fig 1: Playfair' Graph

Working memory capacity (Figure 1). Modern cricket infographics shown during the TV telecast, for example, the Manhattan, the worm graph, pitch map reports, the wagon wheel amply use these basic graph methods. The worm graph is one such example of the line graph, it plots the balls on the x-axis and the cumulative runs on the y-axis. The Manhattan represents the bar graph concept.

Visualization brings quick co-ordination between seeing (visual perception) and thinking (cognition). Vision is handled by the visual cortex located in the rear of the brain, it is fast. Thinking is handled by the cerebral cortex in the front of the brain, it is at a slow pace. We see instantly with ease, but think at a slower pace. Data visualization shifts the balance toward greater use of visual perception, taking advantage of our powerful eyes. [3] Efficient visualization helps in seeing and understanding, together. Humans receive input from all five of their senses (sight, touch, hearing, smell, taste), but they receive greater information from vision than any of the other four. Fifty percent of the human brain is dedicated to visual functions, and consequently, images are processed faster than text. The brain processes pictures all at once, but processes text in a linear fashion. It takes much longer to understand a text based information. In cricket also, we find that an image is easier to remember and recall than its equivalent version of 300 balls-to-ball data. In this paper, we have tried to be experimental, like Playfair, with visualization [4] and graph method. We have developed a predictive model by editing the worm graph using a simple mathematical method known as 'casting out nines'. This technique comes under the topic of 'Number theory- Digit root'.

2. Predictive Visualization in cricket

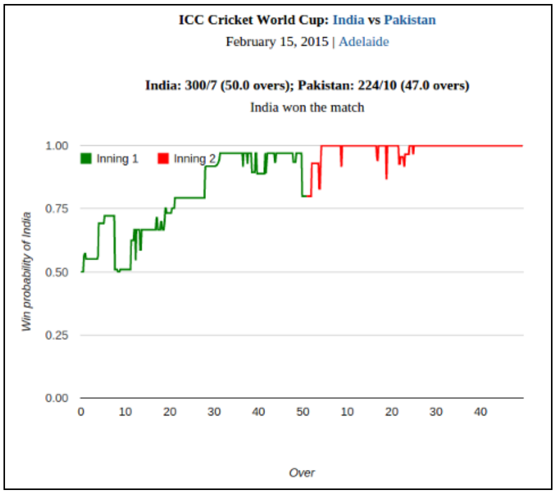

Fig 2: Cricmetric' probability function based visual predictive model.

Data mining, visualization, and predictive modeling are increasingly being used in cricket [5, 6] to analyze the performance of the players, [7] to provide the viewers with critical insights, for betting, and to know 'who is winning'? There is an upsurge of data visualization on world cups, [8, 9] and Sachin Tendulkar' career analysis. Gramener [10] creates and presents several pages of complicated past data in a compact single page visualization. In the predictive models category in cricket, the best served example is the 'Winning and Score Predictor' developed by Dr. Scott Brooker and Dr. Seamus Hogan. The WASP gives a predicted score and the winning probability in the matches shown on Sky Sports. (11)

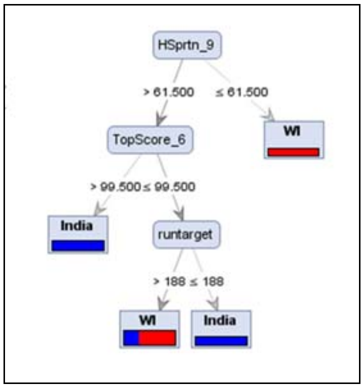

Fig 3: Visual decision tree based predictive model

Few of the visual predictive model examples in cricket includes the 'Score with data' [12] and the 'IBM Watson' [13] which predict the outcome of a cricket match in progress on feeding real time score and historical data. Cricmetric creates the 'Win Probability statistics' (Figure 2) to quantify the chances of a team winning a limit over cricket match. One can check out the scoreboard page to see the live win probability [14] of ongoing Cricket matches! Bala Deshpande uses decision trees, in a visual manner, (Figure 3) for cricket predictions. [15] All these models belong to the 'In-play' modeling category. But they require past data along with the real time data. Most of the above-stated models extract and assemble the past data, churn it using statistical formulas and then, try to trace any pattern or predictive element present in it. The mathematical techniques used are non-friendly to the common man.

3. Cricket Infographics

Infographics are extensively shown during TV telecast. Few of the cricket infographics are manhattan, worm graph, pitch map reports, bowling overview and wagon wheel. *Manhattan* is a series of bars whose heights show the number of runs scored in each over. The *'worm graph'* is an x-y axis graph showing the cumulative runs and fall of wickets of a running match. *Pitch map* is a scalar visualization which shows where the bowler pitched all the balls in his quota of overs in a match or in a spell. A wagon wheel shows which parts of the ground the batsman is playing his shots in that inning. England Cricket Board have developed the 3D visualization feature for the above infographics, and heat maps in their score app [16] for iPad users. The user can instantly get an idea of the match and the player's performance by these easy visuals.

It is thought provoking that the above TV infographics provide good insights but no clue about the probable score or the match outcome during the first team's innings. Here, Michael Lascarides [17] tried to modify the Manhattan by breaking down each bar to show the various runs scored for each ball in that over. He intends to present the steady accumulation of runs.

This is just the beginning. We intended to turn the infographics into 'predictive models'. In our research, we brainstormed over the predictive utility of the worm graph.

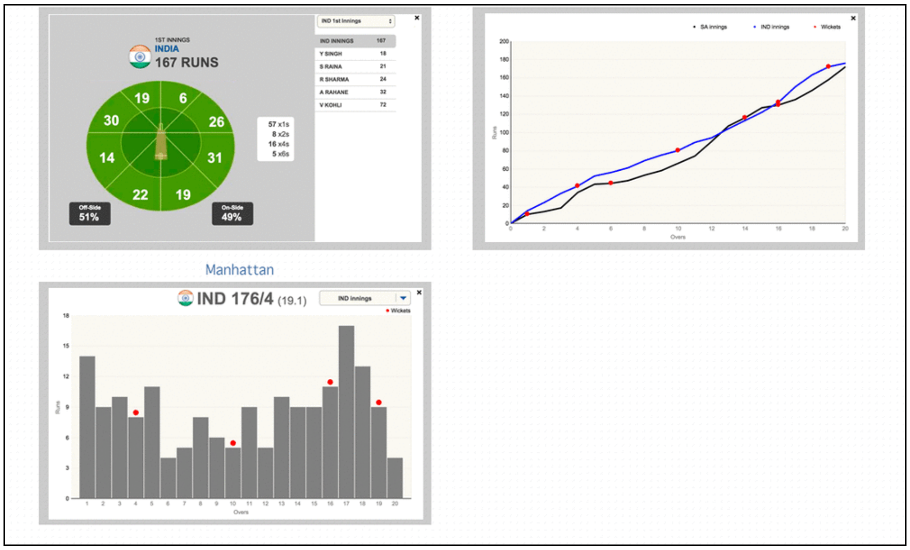

Fig 4: Wagon wheel, manhattan and worm graph visual.

4. Worm graphs

A line chart or line graph displays information as a series of data points called 'markers' connected by straight line segments. The 'worm graph' is a runs/balls line chart plotted on x-y axis showing the cumulative runs of the team and fall of wickets (Figure 4). A single creeping line visualizes the entire 120 to 300 balls data of the match! And the 'dots' reveal about the fall of wickets; the length of the gap between two dots tells the spectator whether the partnership had been long or short, without going into the actual figures. Also, it tells us whether the team is scoring at a good or bad pace. During the chasing team's innings, it provides a visual clue whether the chasing team is pacing better than the first team. A better pace, lesser 'wickets' increase the probability of winning. Despite all these clues, one doesn't get to know who will win! A rising worm may indicate a victory nearly till the end of the game, whereas a sudden, sharp decline in the worm has created an opposite result in the end [18]. In this research paper, we edit and convert the worm graph into a predictive model using the mathematics of 'remainder' values, a concept which is much used in computer science. [19]

5. Materials and method

There are lots of valuable information hidden in cricket data. But, the quantity of the data is becoming too large to handle. Over 80, 000 cricket matches have been played till now. ESPN Cricinfo, itself, carries over 3700 ODI match and 552 T20 match scorecards and ball-to-ball data details. The underlying data is generating much faster than it can be processed and made sense of. As a result, this information often remains buried and untapped. Under such circumstances, the predictive models loose the sense of individual match analysis. They would mix the entire set of 80, 000 matches data (or else any other sample size) and try to trace any consistent, meaningful pattern to predict the match outcome. In our model, we adopted the 'per match analysis'approach. We used arithmetic' equations to develop this predictive model. According to the principle of induction, if an arithmetic identity/equation is true for an integer 'n', and for 'n + 1', then it is equally valid for infinite values of integers. We needn't check too many samples.

For the research work and the model construction, we used simple 'excel' worksheet. Data mining tools scour databases for hidden patterns and find such predictive information that experts may miss because it lies outside their expectations. The surprise tools we have used is an arithmetic method known by the name 'casting out nines' and digit root. Under this approach, we remove the 'nines' and concentrate on the mini-sized remainders.

5.1 Casting out Nines

Our entire data set related to cricket scores are supposed to be only in 'integers'. Every integer i.e. whole number comprises two parts: 1) nines and 2) digit roots. The 'Casting out nines' [20] technique means we cast out/remove all the multiples of nine from the number. The resultant number is in a single digit format, known as the digital root of that number. In modular arithmetic, we express the '9' as the modulus and the digital root as the remainder [21]. When we remove 'all' of the nines, we arrive at the digit root. If 'not all' the nines are removed, the remainder value is 'digit sum'. A striking feature of digit root is that all the 'mathematical properties' of the original number resides in its 'digit root'. They are like perfect mathematical samples of the original numbers. we can use the simpler addition a' + b' as a check on the correctness of the equation on original numbers a + b. It applies to multiplication and exponentiation and many other operations where the function isn't distorting the integer nature of the numbers.

$$(17 + 61)\ (\text{mod}\,9) \equiv (8\ + 7)\ (\text{mod}\,9),$$

$$\textit{where}, \text{mod}\,9\ \textit{implies that we remove all the multiples of}\ '\textit{nine}'.$$

$$\Rightarrow\ 78\ (\text{mod}\,9)\ =\ 15\ (\text{mod}\,9)$$

$$\Rightarrow\ 6\ =\ 6$$

Likewise, for testing the multiplication,

$$a \times b\ \equiv\ a' \times b'\,(\text{mod}\,9)$$

The cricket data may pile and assume large size, and may appear to be random. But, most of the parts of that data comprises ‘nines’. If we remove the nines, it would lead to ‘data reduction’ in a major way (Figure 5). And then, we hit upon the real gold treasure which is the thin bed of remainders i.e. ‘digit roots’. The earlier said statement ‘intelligently probing it to locate the value’ fits here. Perhaps, integrating this ‘casting out nines’ in data mining algorithms leads to meaningful patterns with data visualization and provides a better way to explore the pattern.

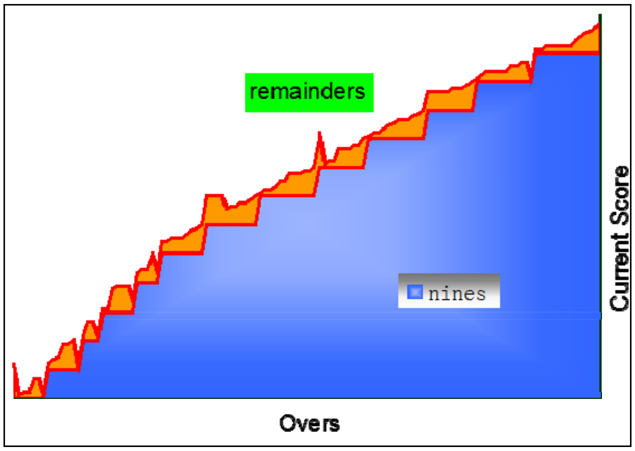

Fig 5: In our research experiment, we considered the worm graph as an area graph comprising a pile of 'nines' (that needs to be removed) and the valuable ore, the flaming 'remainders’. We found that the ‘remainders’ portion provided the clue about the probable winner team in any specific match.

5.2 Ball by ball data

We required the ball by ball score data of the match. We extracted this data from the live commentary page of the ESPN Cricinfo [22] site, cricsheet [23] dot com. For the efficient working of the model, it is preferable to design an application that can extract the real-time ball by ball score, perform the mathematics internally, and produce the modified ‘worm’ on its display screen.

6. Construction of the Model

We regard batsmen’ data (runs scored + balls faced) as the samples, and the total score as the ‘population’. This is a non-parametric predictive model as we are not proposing any statistical assumptions about the total score. Statistical metrics like ‘correlation’ and ‘regression’ can check only the ‘closeness, not equality’ between the dependent and the independent variable. In contrast, the actual match result is based on ‘exact equality’ between integers. Hence, in our model, we chose the ‘equality’ concepts and functions. We introduced a simple ‘mathematical’ assumption that under the ideal condition, at least one of the out batsmen’ data must be congruent to the total score. Differently stated, we assume an equality between their digital root values. A ‘mathematical’ co-ordination should be present between the batsman’ scoring effort and the team’s output. When the digit roots tally that would mean the team has scored in a mathematically coordinated manner. And that would increase the chance of their winning the match. In our previous paper, we proposed a ‘win/defeat’ model [23] based on equality only between the batsman’s data with the team total score/target. Here, we increased the reliability of this test by checking the congruence throughout each delivery of the match till the end of the first innings. We checked the equality between the digit roots of the batsman’s with that of the ball by ball current score. The extra runs conceded by wides, no balls were added to its next legal delivery. If (r + b) represents the batsman’s data, ‘cs’ represents the ‘instantaneous’ i.e. current score, and ‘s’ the final score, what we are expecting is this

dr (r + b) = dr (cs) = dr (s)

Or in terms of modular mathematics,

(r + b) mod 9 = cs mod 9 = s mod 9

The maximum number of times the data syncs in the above way, the greater the possibility of the first batting team’s win. For checking the ‘counts’, we use the simple ‘frequency’ function. If the frequency count is high, there is a win chance; if it is absent or low, the first team is heading for a defeat.

There is a strong base for using the above concept because in actual situations, almost in 80 percent of the matches played till now, we find that (r + b) is equal to ‘s’ or else ‘9 - s’. Hence, we have reciprocated similar kind of mathematical ‘back’ test.

We can’t actually apply the ball by ball score remainder in the graph method. Here, we attempted a variation. In a running match, we cast out nines from the team score at the end of each over. In the worm graph, the remainder portion look like the example thin line of ‘green’ bits, and the long blocks are the ‘nines’ portion in that score. We made use of these ‘remainder’ values to predict the outcome of the IPL match Chennai Super Kings v Kings XI Punjab, [25] played at Durban on May 20, 2009. In this famous match, CSK successfully defended the low target of 117.

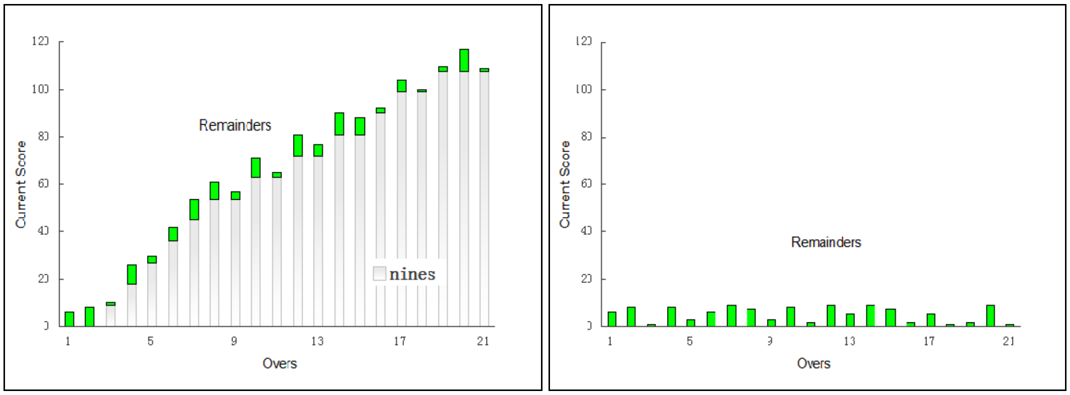

Fig 6: The 'remainders' in the over-by-over data of an example match. We remove the nines, this leaves us with only the remainders (right figure).

Add the Batsman's data lines

Batsman's data (runs scored balls faced) is assumed to work as the sample to the team score.

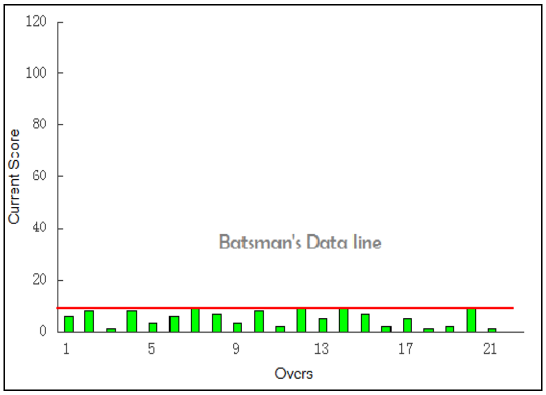

Fig 7: For a win, the maximum count of green remainders should touch the red batsman' data line.

Next, we calculated the batsmen data for the first and second wicket batsmen, termed them as Data one, Data two. We plotted them as lines in the previous graph along with the bars. And that created our visual predictive model. The graph (Figure 7) showed the scoreline green bars touching the batsman's data red line. It also showed the count of the touched, it was n=7. So, the model predicted first team's win chance. The actual result was the same, CSK defended their score and won the match.

For convenience, we tried a variation. Instead of the 'bar' chart, we plotted everything as a simple line graph (Figure 8) with data points highlighted as red dots. Along with the batsmen' data lines, we calculate the 9 - x version of their remainder values and plot it as dashed lines to represent the 'weak' lines.

As long as the scoreline touches the batsman' line, or else no lines, everything is good. But, as soon as the scoreline touches the weak data line, we can sense that the first team is losing grounds. For 'exact' results, we will calculate and check the ball by ball score remainders!

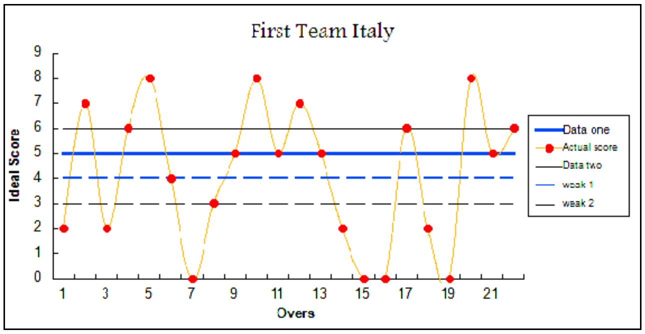

Fig 8: Italy v Canada match. Our model showed that the scoreline had touched the Batsmen' data lines more frequently than the weak lines.

We applied our predictive model in the narrow finish match, Canada v Italy, [26] played on Nov 24, 2013. The digit root of first out batsman, Northcode' data, (dr) 15 equals that of the total score (dr) 104. And furthermore, the digit root of the second out batsman, Berg' data, (dr) equals that of the target (dr) 105. The simple method suggests a win for the first team. But when the target is so low as 105 runs, it is risky to choose the first team as the probable winner. So, we checked the frequencies of these two digit roots through the entire innings of the first batting team. Their '9 - x' values, 4 and 3, represented the weakness parameters. We arrived at the following values (Table 1) in which we find 10 > 9, 19 > 13. Clearly, the 'net' value reflected the chance of winning for the first batting team Italy. And the actual result was the same.

Table 1: Italy v Canada match, frequency counts.

Win parameters	Defeat parameters
(n) 5 = 10	(n) 4 = 9
(n) 6 = 19	(n) 3 = 13

7. Results and discussion

We tested our model during the cricket World T20 2016. It correctly predicted even the most surprising match outcomes, for example, high targets chased down, low targets defended. Next, we tested it on all the matches of IPL 2016. The chasing team won most of the matches in IPL this year. In all those cases, the graph showed the scoreline touching the weak lines, during the beginning 5 overs itself. Also, the model correctly showed the first batting team's win chance in those few matches in which the team defended their low targets with success. The results of our experiments and the model are available at this site. [25] When neither of the first three batsmen' data were tallying with the total score/target but with the ball by ball current score, and had a high frequency count, this caused a win for the first team. In a few matches, the data of the middle order or the tailenders tallying with the score/target's digit root for many times lead to a their win. From these test results, we infer that 'a digital root' co-ordination should be present between the batsman's data and current scores or else the total score/target to predict a win for that team.

8. Conclusion

This is perhaps the most basic mathematical rule that in an equation the LHS should be equal to the RHS. The 'digit root' helps check this equality in an easy and accurate manner. Through this model, we have conceptualized that an equality would cause a win for that team. We combined the number theory and the graph method into one. We used the 'innovative' vision of Playfair towards graphs and developed this model which reconstructs the 'worm graph' and extracts the predictive hint from it. The model can be regarded as a useful, mini tool for the cricket lovers and those who are interested in knowing 'who will win this match?'

9. References

1. Ramageri BM. 'Data Mining Techniques and Applications', Indian Journal of Computer Science and Engineering. 2010; 1(4):301-305.
2. Spence I. William Playfair and the psychology of graphs. In: 2006 JSM proceedings, American Statistical Association, Alexandria, 2006, 2426-2436.
3. Few S. Data Visualization for Human Perception', the Encyclopedia of Human-Computer Interaction.
4. Rao CR, Rao RC, Wegman EJ, Solka JL. Handbook of statistics: Data mining and data visualization. Amsterdam: Elsevier/North Holland, 2005.
5. Ahmed W. A Multivariate Data Mining Approach to Predict Match Outcome in One-Day International Cricket. Available at: http://gsse.pafkiet.edu.pk/sites/all/files/ThesisReports/2015_08_05_MS_WaqarAhmed.pdf (Accessed: 16 June 2016).

6. Adler T. How big data has transformed cricket, 2015. Available at: http://business-reporter.co.uk/2015/03/23/how-big-data-has-transformed-cricket/ (Accessed: 3 April 2016).
7. Smith D. So what should 'the data' have told England cricket coach Peter Moores, 2015. Available at: http://blogs.sas.com/content/sastraining/2015/03/13/so-what-should-the-data-have-told-england-cricket-coach-peter-moores/ (Accessed: 26 July 2016).
8. Team E. Cricket world cup fever–analyzing the data with power query - office Blogs, 2015. Available at: https://blogs.office.com/2015/04/03/cricket-world-cup-fever-analyzing-the-data-with-power-query/ (Accessed: 2 July 2016).
9. India vs Pakistan cricket Available at: http://shwetahumnabadkar.github.io/India_Pakistan_final/ (Accessed: 7 January 2016).
10. India ODI batting, 2015. Available at: https://gramener.com/posters/ India-ODI-batting.pdf (Accessed: 29 March 2016).
11. Brad. University of Canterbury creates 'WASP' for sky sport's cricket coverage throng, 2012. Available at: http://www.throng.co.nz/2012/11/university-of-canterbury-creates-wasp-for-sky-sports-cricket-coverage/ (Accessed: 7 April 2016).
12. Naidu P. How IBM's #Scorewithdata is crunching big data cricket insights into visual content for #CWC15, 2015 Available at: http://lighthouseinsights.in/ibm-scorewithdata.html/ (Accessed: 30 March 2016).
13. Keywebmetrics. Predicting One day cricket (ODI) match winners with IBM Watson Analytics, 2015. Available at: http://www.keywebmetrics.com/2015/01/predicting-one-day-cricket-match-winners-ibm-watson-analytics/ (Accessed: 7 April 2016).
14. Azavedo C. Cricmetric, 2016. Available at: http://www.cricmetric.com/blog/ (Accessed: 26 July 2016).
15. Deshpande B. Predictive Analytics with cricket statistics: IND-WI game prediction, 2016. Available at: http://www.simafore.com/blog/bid/55072/Predictive-Analytics-with-cricket-statistics-IND-WI-game-prediction (Accessed: 26 February 2016).
16. ECB cricket live sets the bar for sports apps, 2014. Available at: http://www.othermedia.com/news/ecb-cricket-live-sets-the-bar-for-sports-apps (Accessed: 29 March 2016).
17. Lascarides M. Visualizing cricket, 2015. Available at: http://lascarides.github.io/cricket/ (Accessed: 30 March 2016).
18. 4th ODI: Australia v India at Canberra, 2016 | cricket scorecard (no date) Available at: http://www.espncricinfo.com/ci/engine/current/match/895813.html (Accessed: 20 January 2016).
19. Modular arithmetic, 2016 in Wikipedia. Available at: https://en.wikipedia.org/wiki/Modular_arithmetic#Applications (Accessed: 29 July 2016).
20. Weisstein EW. Casting out Nines, 2002. Available at: http://mathworld.wolfram.com/CastingOutNines.html (Accessed: 10 April 2016).
21. Burton David M. Elementary Number Theory. McGraw-Hill. 2010, 17-19.
22. Ltd, ESM. Searchable cricket statistics database, 2016. Available at: http://stats.espncricinfo.com/ci/engine/stats/index.html (Accessed: 2 March 2016).
23. Stephen (no date) IPL data. Available at: http://cricsheet.org/downloads/ipl.zip (Accessed: 23 April 2016).
24. Nair Prasant. Mathematical Suggestions to the D/L Rain Method (December 15, 2015). Available at SSRN: http://ssrn.com/abstract=2729437.
25. 54th match: Chennai super kings v kings XI Punjab at Durban, may 20, 2009 | cricket scorecard (2008) Available at: http://www.espncricinfo.com/ipl2009/engine/current/match/392234.html (Accessed: 30 July 2016).
26. 50th match, group A: Canada v Italy at Abu Dhabi, Nov 24, 2013 | cricket scorecard (2013) Available at: http://www.espncricinfo.com/icc-world-twenty20-qualifier-2013/engine/match/660191.html (Accessed: 23 July 2016).
27. Sri Lanka vs Australia First test - world t20 score Info (no date) Available at: https://sites.google.com/site/worldt20scoreinfo/home/graph-method/sri-lanka-vs-australia-first-test (Accessed: 31 July 2016).