Predicting a T20 Cricket Match Result While The Match is in Progress

by

Student Name
Student ID
Student Name
Student ID
Student Name
Student ID
Student Name
11000000

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
August 2015

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Student Name

Student ID

Student Name

Student ID

Student Name

Student ID

Student Name

Student ID

Approval

The thesis/project titled "Predicting a T20 Cricket Match Result While The Match is in Progress" submitted by

1. Student Name (Student ID)
2. Student Name (Student ID)
3. Student Name (Student ID)
4. Student Name (Student ID)

Of Summer, 2015 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)

Name of Supervisor

Senior Lecturer
Department
Institution

Program Coordinator:
(Member)

Name of Program Coordinator

Designation
Department
Brac University

Head of Department:
(Chair)

Name of Head of Department

Designation
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

This is optional, if you don't have an ethics statement then omit this page

Abstract

Data Mining and Machine learning in Sports Analytics, is a brand new research field in Computer Science with a lot of challenge. In this research the goal is to design a result prediction system for a T20 cricket match while the match is in progress. Different machine learning and statistical approach were taken to find out the best possible outcome. A very popular data mining algorithm, decision tree were used in this research along with Multiple Linear Regression in order to make a comparison of the results found. These two model are very much popular in predictive modeling. Forecasting a T20 cricket match is a challenge as the momentum of the game can change drastically at any moment. As no such work has done regarding this format of cricket, we have decided to take the challenge as T20 cricket matches are very much popular now a days. We are using decision tree algorithm to design our forecasting system by depending on the previous data of matches played between the teams. This system will help the teams to take major decision when the match is in progress such as when to send which batsman or which bowler to bowl in the middle overs. It significantly expands the exposure of research in sports analytics as it was previously bound between some other selected sports.

Keywords: Data Mining; Machine Learning; Cricket; T20; Prediction; Decision tree; Linear Regression Analysis

Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our co-advisor Teacher Name sir for his kind support and advice in our work. He helped us whenever we needed help.

Thirdly, Name and the whole judging panel of Conference Name. Though our paper not accepted there, all the reviews they gave helped us a lot in our later works.

And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ	Epsilon
υ	Upsilon
IPL	Indian Premier League
LBW	Leg before Wicket
MR	Runs scored by Home team
MRN	Home Team Run Rate
ODI	One day International
OR	Runs scored by the opponent team
ORN	Opponent Team Run Rate
$T20$	Twenty Twenty

Chapter 1

Introduction

1.1 Thoughts behind the Prediction Model

There have been research done on ODI and Test match cricket but very few on T20 cricket, which is currently more favourite than its older brothers. And that's why we decided to do research on this format of the game. The result of a T20 cricket match depends on lots of in game and pre-game attributes. Pre-game attributes like condition, venue, pitch, team strength etc. and in game attributes like wickets in hand, run rate, total run, strike rate etc. influence a match result predominantly. We gave more emphasis on in game attributes as our prediction will be when match is in progress. Our intentions would be to finding out the attributes which is most affecting the result in different phases of the game. We broke an innings into three phases: Power-play (1-6 overs), Mid-overs (7- 16) and - final overs (17-20). Prediction will be active till the last over of mid overs phase. We consider an entire cycle of process of data mining, decision making and preparing a model to predict. Mining the data according to the attributes and different phases we have divided important to construct meaningful statistics. Modeling a problem for prediction requires several intelligent assumptions and molding the problem with collected data-sets. As we already mentioned cricket is a game of uncertainty and T20 format is the most unpredictable format rather than the other two format because it is the shortest format of the game and one over can change the result of a game. In this research we tried to design a prediction model which can go with this unpredictability and try to make a result prediction.

1.2 Aims and Objectives

The aim is to prepare a model which will predict the result of a T20 cricket game while the match is in progress. Our main objective is to combine pre-game data and in-game data in order to design a good predictive model. Understanding the different attributes is also needed in order to get more accuracy in result.

1.3 History of T20 Cricket

From the sixteenth century to first official Test match in 1877 to first ODI World Cup 1975 with 60 overs to 50 overs world in 1987, cricket world has changed a lot.

Now, we have a new shortest format in cricket - Twenty20 or T20. T20 cricket is fun, entertaining and more thrilling than other two formats. It has brought glamour and instant popularity to the fans and helped marketing Cricket to the rest of the world. England Cricket Board (ECB) was looking for a cricket competition to fill the void after the conclusion of Benson and Hedges Cup in 2002. ECB was looking for something new to at-tract more sponsors and viewers. Marketing Manager of ECB, Stuart Robertson was the first to came up with the idea of playing cricket match with each team getting only 20 overs to play. Thus came the name Twenty20. First official T20 matches were played in English counties. Though first official T20 international match took place be-tween Australia and New Zealand. In 2007 we witnessed ICC World Twenty20, which generated immense support for this newest form of cricket. Introduction of T20 cricket gave birth to franchise league in many countries. Among those Indian Premier League (IPL) is the most watched and expensive cricket league. Big Bash, Caribbean Premier League (CPL) and Bangladesh Premier League (BPL) other popular franchise leagues. T20 cricket showed great innovation in batting style, improved fielding. Bowlers were also trying their hardest to make them useful in a format which was made to give preference to the batsmen. With more viewers and sponsors, T20 cricket brought more money to the Boards and players. But this format also attracted more illegal activities as matching fixing, betting, miss con-duct of players. Very recently in July 2015 BCCI banned Chennai Super Kings and Rajasthan Royals for match fixing. Ironically this two teams are two of the most successful team in IPL with a large number of fan base. The West Indies regional teams completed in tournament named Stanford 20/20 which was funded by a convicted fraudster Allen Stanford.

1.4 Game Method

After Football, Cricket is the second most popular sports with a fan base of around 2.5 billion (according to Top End Sports) and mostly popular in South Asia, Australia, The Caribbean and UK. In international level Cricket is played in three formats- Test, ODI and T20I cricket. This game is played on a 22 yards clay pitch with 2 sets of stamps, each set with 3 stamps and each set having two bells on top of them. Two batsmen come to pitch with two wooden bats and bowler bowls with a cricket ball which outer part is made of lather. Test Cricket is played Red ball which is slightly heavier than the White bowl played in the limited overs. There is no fixed size of the outfield, but usually its diameter usually varies between 137 meters and 150 meters. In limited over cricket there is a circle of 30 yards around the pitch which work as a field restriction for players. Test cricket is played for 5 days with each team having at most 2 innings. ODI played for 50 overs per innings and T20 played in 20 overs. Each team play with 11 players. A coin toss decides who is going to bat or ball first. In limited over cricket team batting first scores as many run possible before the overs are finished or they all get out. If team batting next score more runs they wins and failure to score required runs in allocated overs or getting all out result in loss for team batting second. Some basic idea how the game is played: Field Restriction: According to the latest rule change in 50 overs cricket, there is only one Power play from over 1-10 with only two fielders outside of the 30 yards circle. Between 11 to 40 overs four fielder are allowed and five allowed outside the 30 yards circle in the final 10 overs. Like the ODI format T20 also have only 1 power play form over 1 to 6 with 2 fielders outside the circle. Scoring Runs: The striking batsman must hit the ball with his bat and must change his position with his partner to score 1 run. Number of runs scored depend on the number of time the batsmen change position. If the batsman hit ball and its goes outside the boundary 4 runs are added and 6 runs are added when the ball fly over the boundary. Batting team gets extra runs form No ball, Leg bye, Bye, Wide, Overthrows and Penalty runs when the ball hits keeper's helmet or cap lying on the field. Out Types: Batsmen usually get out by being bowled, caught, leg before wicket (LBW), stumped and run out. There are some rare occasion where batsmen get out by hit wicket, intentionally hitting the ball twice, handled the ball, obstructing the field and timed out. Tie match result: If the match is tie, such as both the team scored same runs then there is a rule called super over. Super over played for only one over for each team. Each team can play with two wickets when they are batting and one single bowler when they are bowling. Batting first team set a target and second team chase it. In Test cricket there is no restriction on how many overs a bowler can bowl. But in limited over cricket number of overs bowled by a single bowler is fixed. in ODI's each bowler can bowl up to 10 overs in a match and in T20 cricket bowlers are allowed to bowl only 4 overs each.

Chapter 2

Related Work

Better predictive modeling depends on better understanding of the data and attributes selection. We have to choose between some data mining algorithm. We have chosen data mining as it is very much flexible in predictive modeling. Prediction when the game is in progress is a tough ask and it need finding the best attributes that influence the match outcome. Some research was done previously on predictive modeling in sports like Basketball, Baseball along with Test and One Day International cricket. In basketball, Bhandari et al.[1] developed a knowledge discovery system and data mining framework for National Basketball Association (NBA). It was aimed to discover several interesting patterns in basketball games. This and related system have been used by several basketball teams over the past decades. Such solutions designed for offline usage and no in game effects were taken care of. There has been some recent works (20) about in-game decision making to find how much time remaining in the game without making any prior prediction model. There were several works done in cricket. Bailey and Clarke[2] and Sankaranarayanan et al.[7] used machine learning approach to predict the result of a one day match depending on the previous data and in game data. Akhtar and Scarf[4] used multinomial logistic regression in their work on predicting a outcome of a test matches played between two teams. Choudhury et al.[3] used Artificial Neural Network to predict result of a multi team one day cricket tournament depending on the past 10 years data. They used training set in order to model the data in neural network. Again there was no in play effects were taken care of. For baseball, Ganeshapillai and Guttag[5] developed a prediction model that decides when to change the starting pitcher as the game progresses. It is very much similar to our work-flow, where they used the combination of previous data and in game data to predict a pitchers performance. Tulabandhula and Rudin[6] were designed a real time prediction and decision system for professional car racing. Model makes the decision of when is the best time for tire change and how many of them. These works supplied a huge encouragement and informative ideas in our research.

Chapter 3

Prediction Modeling using Decision Tree

Decision tree algorithm is a very popular way to design a predictive modeling. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Runs, Wickets and Run-Rate). Leaf node (e.g., Result) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree built on the calculation of Entropy and Information gain.

3.1 Entropy Calculation

Entropy is a measure of unpredictability and uncertainty of a data-set. Entropy is generally considered to determine how disordered a data-set is. The higher rate of entropy refers to the uncertainty and more information needed in these cases to improve the predictability. One outcome is very much certain when the entropy is zero.

$$Entropy(S) = \sum_{i=1}^{C} Pi \log_2 Pi \tag{3.1}$$

Where Pi is the proportion of instances in the data set that take the i-th value of target attribute, which has C different values. This probability measure give us the idea of how uncertain we are about the data. We use a log2 measure as this represents how many bit we would need in order to specify what the class is of a random instance.

3.2 Information Gain

Now we want quantitative way of splitting the data-set by using a particular attribute. We can use a measure called Information Gain, which calculates the reduction in entropy that would result in split-ting the data on an attribute, A. Information Gain is actually a procedure to select the particular attribute to be a decision node of a decision tree.

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon A} \frac{S_v}{S} Entropy(S_v) \tag{3.2}$$

where v is a value of A, Sv is the subset of instances of S where A takes the value v and S is the number of instances With the help of this node evaluation technique we can proceed recursively through the subset we create until leaf nodes have been reached throughout and all subsets are pure with zero entropy. This is how a decision tree algorithm works.

3.3 Data Training

After collecting the data we converted those data into an attributed relation file format (.arff) and then we have used Weka for classification. After classification using some algorithm we got some result and later we have analyzed those result. Here is the simple work flow chart given.

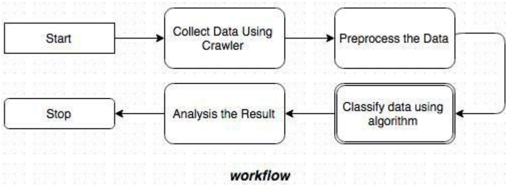

Figure 3.1: Workflow

Chapter 4

First and Second

4.1 Result Prediction based on First and Second Segment (Bat First)

As we have divided our total model into three segment and we actually consider first two segment for predicting the match outcome as we wanted to find out the final match result when match is in progress. We have taken total 91 match for making our model using multiple linear regression and we have merged all the attributes from those matches based on different segment. After analyzing those two segment our model has given 75% accuracy. So, we can predict any match outcome when the match is in progress based on our model. As we did not take any attributes from the team who will bat second and considering the attribute which we got from first segment, our predicted model is quite good.

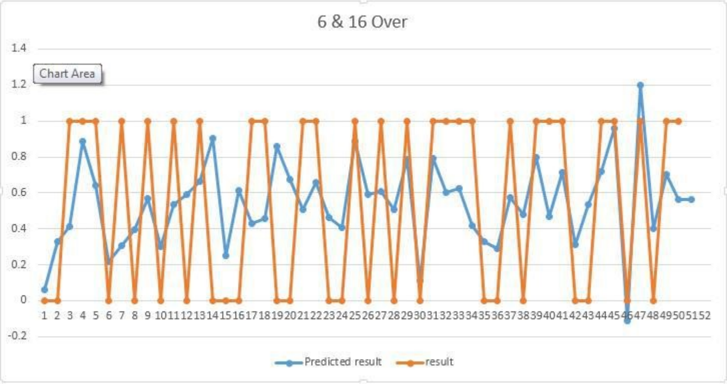

Figure 4.1: First and Second Segment Prediction (Bat First)

From the figure above we can see the graph view of our model, here 0 means lost and 1 means win. So, if the predictive final value is less than 0.5 then the result would be consider as lose and if the predictive value is greater than 0.5 then it would be consider as win.
Coefficient: These are the coefficients values for all the attributes from Win prediction based on bat first.
P-values: These are the p values for all the at-tributes from Win prediction based on bat first.

Attributes	Coefficients
Intercept	0.092219
Venue	0.242112
M6ORN	0.039573
M6OW	-0.12872
M16ORN	0.05121
M16OW	-0.07214

Table 4.1: First and Second Coefficient (Bat First)

Attributes	P-value
Intercept	0.87596
Venue	0.088577
M6ORN	0.323375
M6OW	0.084286
M16ORN	0.246463
M16OW	0.254117

Table 4.2: First and Second Segment P-value (Bat First)

4.2 Result Prediction based on First and Second Segment (Bat Second)

While calculating for 2nd innings segments we get the run rate value from team batting first. Which makes a better impact on a prediction model and that time our model has given 85.5% accuracy which is really good.

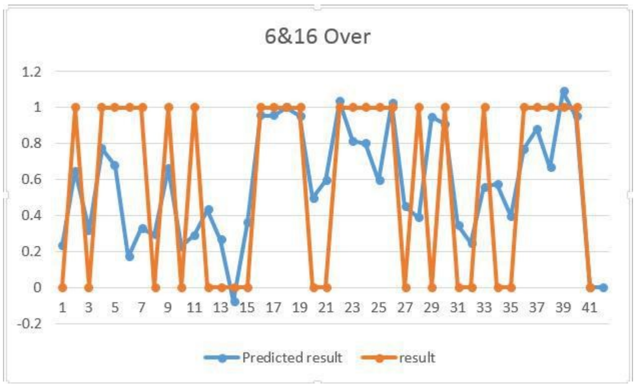

Figure 4.2: First and Second Segment Run Prediction (Bat Second)

Coefficient: These are the coefficients values for all the attributes from Win prediction based on bat second.

P-values: These are the p values for all the at-tributes from Win prediction based

Attributes	Coefficients
Intercept	2.282567
Venue	-0.04063
M6ORN	-0.02869
M6OW	-0.22694
M16ORN	-0.12705
M16OW	-0.10903
O6ORN	0.008056
060W	0.066748
O16ORN	0.028731
O16OW	-0.0978

Table 4.3: First and Second Segment Coefficient (Bat Second)

on bat second.

First and Second Segment P-value (Bat Second)	
Attributes	Coefficients
Intercept	0.007165
Venue	0.811504
M6ORN	0.482433
M6OW	0.019258
M16ORN	0.112474
M16OW	0.090761
O6ORN	0.878555
060W	0.429492
O16ORN	0.633494
O16OW	0.179298

Table 4.4: First and Second Segment P-value (Bat Second)

Bibliography

[1] I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap, and K. Ramanujam, "Advanced scout: Data mining and knowledge discovery in nba data," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 121–125, Mar. 1997, ISSN: 1573-756X. DOI: 10.1023/A:1009782106822. [Online]. Available: https://doi.org/10.1023/A:1009782106822.

[2] M. Bailey and S. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress," *Journal of sports science & medicine*, vol. 5, pp. 480–7, Dec. 2006.

[3] D. R. Choudhury, P. Bhargava, Reena, and S. Kain, "Use of artificial neural networks for predicting the outcome of cricket tournaments," *International Journal of Sports Science and Engineering*, vol. 1, no. 2, pp. 87–96, 2007, ISSN: 1750-9823.

[4] S. Akhtar and P. Scarf, "Forecasting test cricket match outcomes in play," *International Journal of Forecasting*, vol. 28, Jul. 2012. DOI: 10.1016/j.ijforecast.2011.08.005.

[5] G. Gartheeban and J. Guttag, "A data-driven method for in-game decision making in mlb: When to pull a starting pitcher," Aug. 2013, pp. 973–979. DOI: 10.1145/2487575.2487660.

[6] T. Tulabandhula and C. Rudin, "Tire changes, fresh air, and yellow flags: Challenges in predictive analytics for professional racing," *Big Data*, vol. 2, no. 2, pp. 97–112, 2014, PMID: 27442303. DOI: 10.1089/big.2014.0018. eprint: https://doi.org/10.1089/big.2014.0018. [Online]. Available: https://doi.org/10.1089/big.2014.0018.

[7] V. Veppur Sankaranarayanan, J. Sattar, and L. Lakshmanan, "Auto-play: A data mining approach to odi cricket simulation and prediction," Apr. 2014. DOI: 10.1137/1.9781611973440.121.

How to install LATEX

Windows OS

TeXLive package - full version

1. Download the TeXLive ISO (2.2GB) from https://www.tug.org/texlive/

2. Download WinCDEmu (if you don't have a virtual drive) from http://wincdemu.sysprogs.org/download/

3. To install Windows CD Emulator follow the instructions at http://wincdemu.sysprogs.org/tutorials/install/

4. Right click the iso and mount it using the WinCDEmu as shown in http://wincdemu.sysprogs.org/tutorials/mount/

5. Open your virtual drive and run setup.pl

or

Basic MikTeX - TEX distribution

1. Download Basic-MiKTEX(32bit or 64bit) from http://miktex.org/download

2. Run the installer

3. To add a new package go to Start ¿¿ All Programs ¿¿ MikTex ¿¿ Maintenance (Admin) and choose Package Manager

4. Select or search for packages to install

TexStudio - TEX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads

2. Run the installer

Mac OS X

MacTeX - TEX distribution

1. Download the file from https://www.tug.org/mactex/
2. Extract and double click to run the installer. It does the entire configuration, sit back and relax.

TexStudio - TEX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads
2. Extract and Start

Unix/Linux

TeXLive - TEX distribution

Getting the distribution:

1. TexLive can be downloaded from http://www.tug.org/texlive/acquire-netinstall.html.
2. TexLive is provided by most operating system you can use (rpm,apt-get or yum) to get TexLive distributions

Installation

1. Mount the ISO file in the mnt directory

   ```
   mount -t iso9660 -o ro,loop,noauto /your/texlive####.iso /mnt
   ```

2. Install wget on your OS (use rpm, apt-get or yum install)
3. Run the installer script install-tl.

   ```
   cd /your/download/directory
   ./install-tl
   ```

4. Enter command ‘i’ for installation
5. Post-Installation configuration: http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-320003.4.1
6. Set the path for the directory of TexLive binaries in your .bashrc file

For 32bit OS

For Bourne-compatible shells such as bash, and using Intel x86 GNU/Linux and a default directory setup as an example, the file to edit might be

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/i386-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

For 64bit OS

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

Fedora/RedHat/CentOS:

```
sudo yum install texlive
sudo yum install psutils
```

SUSE:

```
sudo zypper install texlive
```

Debian/Ubuntu:

```
sudo apt-get install texlive texlive-latex-extra
sudo apt-get install psutils
```

Overleaf: GitHub for LaTeX projects

This Project was developed using Overleaf(https://www.overleaf.com/), an online LaTeX editor that allows real-time collaboration and online compiling of projects to PDF format. In comparison to other LaTeX editors, Overleaf is a server-based application, which is accessed through a web browser.