Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 55 – 60

Procedia Computer Science

2016 International Conference on Computational Science

Applications of modern classification techniques to predict the outcome of ODI Cricket

Neeraj Pathak[a], Hardik Wadhwa[b] *

"Department of Mathematical Sciences, Indian Institute of Technology, Banaras Hindu University, Varanasi-221005, India"

Abstract

Data mining and Machine learning in Sports analytics is a recent field in Computer Science. In this paper our goal is to predict the outcome of an ODI (One Day International) Cricket match. Outcome of an ODI Cricket match depends on several factors such as home game advantage, Day/Night, Toss, Innings (first or second), physical fitness of teams and dynamic strategies, a lot of which varies as the game proceeds. We have applied modern classification techniques –Naïve Bayesian, Support Vector Machines, and Random Forest, and conducted a comparative study based on their outcomes and performances. Based on the outcome of these models we have developed a tool COP (Cricket Outcome Predictor), which outputs the win/loss probability of an ODI match. The target audience of this tool involves teams playing cricket, and Sports Analysts in general.

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016

"Keywords: ML and Data mining; Classification; Random Forest; SVM; Naive Bayes; Sports Analytics; Predictive Modeling"

1. Introduction

Cricket is a bat and ball game played between two teams of eleven players. It is the world's second most popular sport. Each team turns into bat, attempting to score runs, while other team fields. Each turn is known as innings. The objective is to score more runs than the opponent.

There are three internationally recognized formats of Cricket matches — Test match, One Day International (ODI) and T20. The main difference between the three formats is the scheduled duration of the game. Test cricket's duration is five days, ODI is scheduled to complete in a Day or a Day/Night combination and T20 is a shorter version where each team plays for twenty over (an over is a set of six balls bowled from one end of a

* Corresponding author. Tel.:+91 8808939572; +91 8295117656
*E-mail address:*neeraj.pathak.apm12@iitbhu.ac.in (Neeraj Pathak), hardik.wadhwa.apm12@iitbhu.ac.in (Hardik Wadhwa)

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.126

cricket pitch) each, hence the name T20. An ODI is a form of a limited over cricket, played between two teams with international status, in which each team faces fifty over. Multiple championships and competitions are conducted in these formats of cricket, around the world.

Cricket world cup is an international championship of ODI conducted once in every four year. As per ICC (International Cricket Council), ten full member nations qualify for the tournament without having to appear in any qualifying rounds. In 2015 world cup a total of 14 nations participated in the tournament. In this paper we have focused our study around the ten full member nations of ICC. Those nations are India, Australia, England, Pakistan, South Africa, Bangladesh, West Indies, Zimbabwe, New Zealand and Sri Lanka.

Since its inception ODI cricket has evolved a lot in terms of the rules and playing styles and stadium facilities such as better lighting conditions for day/night matches. Some old rules of ODI are no longer in practice. We have focused on relatively newer data (from 2001 to 2015). Even though Cricket is a highly followed Sport yet we found very little work done on the topic of outcome prediction in ODI format Most of the work is of statistical nature such as De Silva [3] which estimates the magnitude of victory in ODI rather than predicting the outcome as a whole. Bandulasiri [2] with the objective of outcome prediction of ODI and also analyzes the impact of individual factors in outcome prediction. Kaluarachchi and Varde [1] focuses on applications of machine learning in outcome prediction of ODI match. They concluded that Naïve Bayesian was best classification technique out of the four methods they tried.

Outcome of ODI match is influenced by a large no. of factors, for our study we considered the factors analyzed by Bandulasiri [2] and proven to have a significant impact on outcome of ODI match. The factors are

- **Toss outcome :** A coin is tossed at the beginning of the match and the winner (captain of the team) decides whether to bat first or second
- **Home Game Advantage :** It refers to whether the game is being played on home grounds or in a different country
- **Day/Night Effect :** It considers the effect of whether the game is played during day or at night
- **Bat First :** This factor refers to whether the concerned team batted first or second

Our classification models are built using these factors. To predict the outcome of ODI matches we have applied three classification techniques - Naïve Bayesian, Random Forest and Support Vector Machines. We then built a software tool called COP (Cricket Outcome Predictor) based on emerged results of classification. The tool provides a choice of models to predict the outcome of match based on above factors. There is also a provision of getting probability of winning or losing.

The remaining sections are organized as follows. In Section 2, we have discussed our approach to tackle the problem. Section 3 talks about the comparative study of various classifiers used. Section 4 is about the implementation and functionality of the software tool (COP). Section 5 covers future work and conclusions.

2. Approach for classification

Data of ODI matches during the time period 2001-2015 for each team was collected from www.cricinfo.com [4]. We collected data of only those matches which were successfully completed and a clear winner was declared. The factors chosen were Home Game Advantage, Day/Night Effect, Bat First, and Toss outcome. A separate model for each team is prepared, in which each team is analyzed with respect to every other team. The rationale behind this is to avoid any duplicate data entry [1].

Three different classification models, for each team, are built and implemented using open source statistical tool R [5]. The choice of using R is that it provides a large no. of highly optimized open source libraries to implement and build machine learning models. The packages used were caret [6], e1071 [7], randomForest [8].

In all the classifiers the data was divided into training and testing set in 80:20 ratios.

The idea behind choosing a Naïve Bayesian classifier is that our factors are independent and Naïve Bayesian classifier is known to perform best in such a situation. It is based on the Bayes' theorem-

Bayes' theorem is stated mathematically as

$$P(A|B) = \big(P(B|A)P(A)\big)/P(B)$$

- Where $P(A)$and $P(B)$ are the probabilities of A and B without regard to each other
- $P(A|B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B|A)$ Is the probability of observing event B given that A is true.

Define $B = \{b1, b2, \dots bn\}$ as a set of variables, we want to construct a posterior probability for the event A_i among a set of possible outcomes $A = \{a1, a2, \dots am\}$.

Naïve Bayesian [9] classifier employs Bayes' theorem to compute the probability of winning or losing and hence the outcome.

Random Forest [10] is an ensemble method used for classification, regression and other tasks, that operate by constructing a large number of decision trees during training phase. In classification the output is the class which is mode of the classes predicted by the individual trees, while in regression mean prediction of individual trees is given as final output. Random Forests have the advantage of correcting the flaw of decision trees which tend to overfit to their training set.

Support Vector Machines (SVM) [9] are supervised learning models used for classification and regression tasks. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

3. Comparative evaluation of classifiers

To evaluate and compare the performance of classifiers for the task of predicting the outcome of ODI cricket match, we make use of kappa statistic and balanced accuracy. Kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. Higher kappa denotes better classification performance. Balanced accuracy [11] avoids inflated performance estimates on imbalanced datasets. It is defined as the arithmetic mean of sensitivity and specificity, or the average accuracy obtained on either class. Higher values of balanced accuracy shows better performance of classifier

	Classifiers					
	Naïve Bayes		**Random Forest**		**SVM**	
Teams	**Balanced Accuracy**	**Kappa statistic**	**Balanced Accuracy**	**Kappa statistic**	**Balanced Accuracy**	**Kappa statistic**
India	0.6700	0.3377	0.7053	0.4215	0.6355	0.2803
England	0.6396	0.2773	0.6230	0.2436	0.6707	0.3190
Australia	0.5256	0.0625	0.50	0*	0.50	0*
Bangladesh	0.6106	0.2312	0.6106	0.2312	0.7373	0.5043
Pakistan	0.5382	0.0751	0.5555	0.1100	0.6204	0.2149
West Indies	0.6670	0.3484	0.6316	0.2812	0.6250	0.25
Sri Lanka	0.5942	0.1858	0.5763	0.151	0.5985	0.1922
Zimbabwe	0.5622	0.1235	0.5403	0.11	0.5455	0.1286
South Africa	0.5538	0.1134	0.5688	0.1565	0.6538	0.3851
New Zealand	0.6569	0.3125	0.6902	0.3774	0.5802	0.1584
Average	0.6018	0.2067	0.6002	0.2025	0.6167	0.2619

The results obtained showed that on average SVM outperformed the other classifiers, followed by Naïve Bayesian and Random Forest which had relatively similar performance. In case of Australia we noticed that kappa of Random Forest and SVM methods was zero, the reason of this being highly imbalanced data among the lost and won classes, however Naïve Bayesian model was still able to predict without any noticeable anomaly, which is in agreement with the ability of Naïve Bayesian to perform good with imbalanced data as well.

Random Forest and SVM are required to be tuned further to deal with the issue of class imbalance; however it demands a large no. of observations.

4. COP (Cricket Outcome Predictor) software tool

We used gWidgets[12] library of R to develop the Graphical User Interface of this tool. The tool is built by keeping the computation part independent so as to facilitate adding more classifiers in future. It gives the prediction of the ODI match even before the match has started. This is due to the fact that our features do not change during the course of match and their values are available before the beginning of the match.

The tool provides a choice of model to get the prediction, user can select any one of the three classifiers to get the probability or outcome of the concerned match.

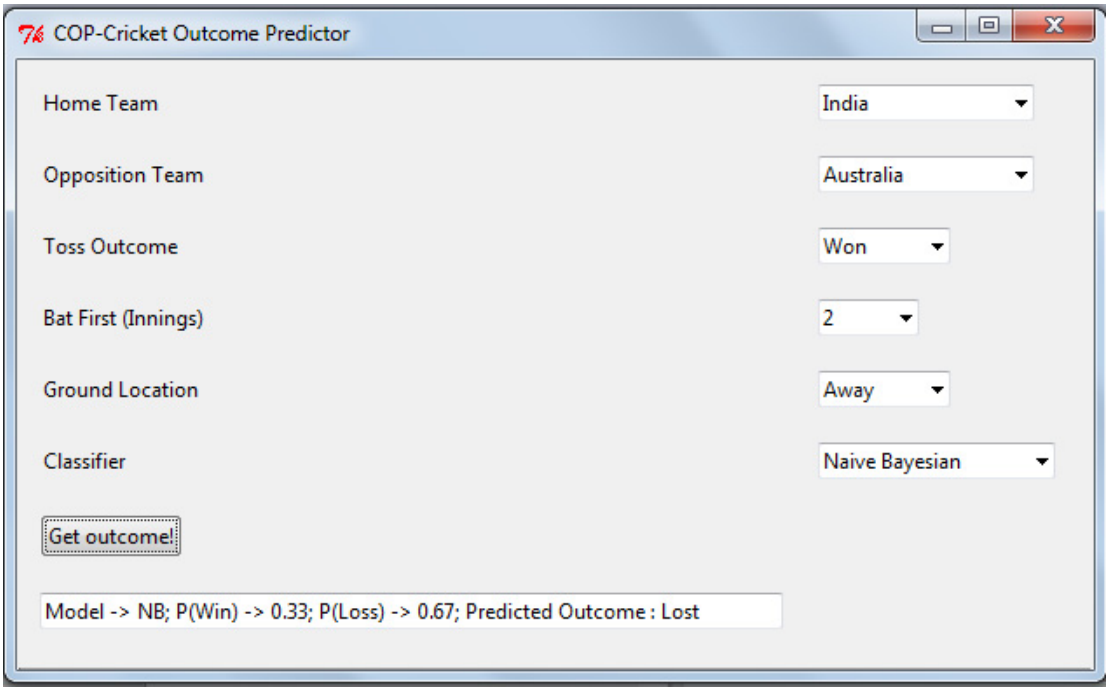

Fig. 1 Snapshot of outcome of COP with Naïve Bayesian as classifier choice

We used COP to predict the outcome of the matches being played between India and Australia, in Australia in during January 2016.

	Naïve Bayesian		SVM		Random Forest		Actual Outcome
Match Date	Probability of Winning	Predicted Outcome	Probability of Winning	Predicted Outcome	Probability of Winning	Predicted Outcome	
12 Jan	0.37	Lost	0.49	Lost	0.10	Lost	Lost
15 Jan	0.48	Lost	0.49	Lost	0.30	Lost	Lost
17 Jan	0.42	Lost	0.52	Won	0.25	Lost	Lost
20 Jan	0.46	Lost	0.49	Lost	0.08	Lost	Lost
23 Jan	0.51	Won	0.49	Lost	0.15	Lost	Won

Table showing predicted outcomes along with probability of winning for India, and actual outcome of the match.

We were able to successfully predict correct outcome of all the matches using Naïve Bayesian model which shows its prowess in making correct predictions when the attributes are independent.

We conclude that Naïve Bayesian is most suited approach in such a scenario.

5. Future work and conclusions

Our motivation for using Naïve Bayesian classifier was based on the fact that it is the most suitable approach when the predictors are independent and is also known to perform decently even in the case of severe class imbalance in the dataset, which happened to be the case in our study.

However, a well tuned Random Forest and SVM classifier using cross-validation and bootstrap sampling, on average are at par with Naïve Bayesian in terms of balanced accuracy. In Section 3, we observed the balanced accuracy of the three methods to be close enough—

Model	**Average Balanced Accuracy**
Naïve Bayesian	0.6018
Random Forest	0.6002
SVM	0.6167

However, in the case of severe class imbalance Random Forest and SVM fail to perform, as observed in the Random Forest and SVM model for Australia.

Future work could be done in the following ways:

- As we know Machine Learning and Data Mining are developing at a rapid pace with several new techniques being developed and old techniques being modified to enhance performance, keeping this in mind our work can be expanded to incorporate new methods of classification for outcome prediction.
- More features could be added along with the ones currently considered.
- Although our study is done for ODI matches only, however similar approach could be applied to predict outcome in other versions of Cricket matches as well.
- Classification techniques can be applied to other sports such as baseball, football as well, although the method of implementation might differ from one sport to another.

References

[1] CricAI: A classification based tool to predict the outcome in ODI cricket, Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on 17-19 Dec. 2010, pp. 250-255
[2] A. Bandulasiri, "Predicting the Winner in One Day International Cricket", Journal of Mathematical Sciences & Mathematics Education, Vol. 3, No. 1.
[3] B.M De Silva, and T.B. Swartz, Estimation of the magnitude of the victory in one-day cricket. Australia and New Zealand Journal of Statistics, 2001, Vol. 43, pp. 1369-1373.
[4] CricInfo, Website for cricket data, [online] http://www.cricinfo.com
[5] R, https://www.r-project.org/
[6] caret, https://cran.r-project.org/web/packages/caret/index.html
[7] e1071, https://cran.r-project.org/web/packages/e1071/index.html
[8] randomForest, https://cran.r-project.org/web/packages/randomForest/index.html
[9] C. M Bishop, "Pattern Recognition and Machine Learning", Springer New York, 2006.
[10] Random Forest, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
[11] Balanced Accuracy, http://ong-home.my/papers/brodersen10post-balacc.pdf
[12] https://cran.r-project.org/web/packages/gWidgets/index.html
[13] Wikipedia on the Game of Cricket, website [Online] http://en.wikipedia.org/wiki/Cricket