CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket

Amal Kaluarachchi, Aparna S. Varde
Department of Computer Science
Montclair State University, Montclair, NJ, USA
(amalkal@hotmail.com, vardea@montclair.edu)

***Abstract* - Victory is the ultimate goal in any sport. In this work we address the winning factors in the sport of One Day International (ODI) cricket. Winning an ODI cricket match depends on various factors related to scoring as well as physical strength of the two teams. Some of the factors have been described in the literature but there is scope for further research on analyzing them, especially with reference to predicting victory. Interesting factors include home game advantage, day / night effect, winning the toss and batting first. In this article, we have used artificial intelligence techniques, more specifically Bayesian classifiers in machine learning, to predict how these factors affect the outcome of an ODI cricket match. Based on the emerged results, we have developed a software tool called CricAI. This tool outputs the probability of victory in an ODI cricket match using input factors such as home game advantage available at the beginning of the match. The CricAI tool can be used in real-world applications by teams playing cricket. It can accordingly be helpful in adjusting certain factors in order to maximize the chances of winning the real game.**

Keywords: AI and Automation, Bayes Thoerem, Classifiers, Predictive Analysis, Probability, Sports Applications

I. INTRODUCTION

Cricket is a bat-and-ball team sport first documented as being played in southern England in the 16th century. By the end of the 18th century, cricket had developed to the point where it had become the national sport of England. The expansion of the British Empire led to cricket being played overseas and by the mid-19th century the first international matches were being held. Today, the sport is played in more than 100 countries.

Standard limited overs cricket (where an *over* is a series of six bat-and-ball rounds) was introduced in England in the 1963 season in the form of a knockout cup contested by the first-class county clubs. In 1969, a national league competition was established. The concept was gradually introduced to the other major cricket countries and the first limited overs international match was played in 1971. The first Cricket World Cup took place in England in 1975.

A "one day match", so called because each match is scheduled for completion in a single day, is the most common form of limited overs cricket played on an international level. In practice, matches sometimes continue on a second day if they have been interrupted or postponed by bad weather. The main objective of a limited overs match is to produce a definite result and so a conventional draw is not possible, but matches can be undecided if the scores are tied or if bad weather prevents a result. Each team plays one innings only and faces a limited number of overs, usually a maximum of 50 (300 deliveries). [17]

The Cricket World Cup is held in one day format. The last World Cup in 2007 was won by Australia. That world tournament was enjoyed the participation of 16 nations, all qualified from a larger pool of potential qualifiers. The final participating nations included Australia, Sri Lanka, South Africa, New Zealand, West Indies, England, Pakistan, India, Zimbabwe, Bangladesh, Ireland, Bermuda, Scotland, Netherlands, Canada, and Kenya. The next World Cup will be hosted by India, Bangladesh and Sri Lanka in 2011.

Match data since the beginning of the ODI game is available. However our literature search found no previous machine learning work on this topic. Some work could be found on topic of optimal scoring rates by Clarke [15] and Preston and Thomas [7]. They utilize dynamic programming methods.

Some studies, such as those conducted by De Silva [3] analyze the magnitude of the victory. It is found that most of these studies describe the factors affecting winning but do not focus on the analysis of the factors with the goal of predicting the probability of victory. In the real-world scenario, however, there are cases where the magnitude of the victory is important especially when betting is involved.

In this work, we focus on analyzing data related to the factors affecting the outcome of cricket. Based on our analysis, we design a software tool that implements the Bayes Theorem whose basic formula is stated here.

$$P(\theta \mid \mathrm{X}) = \frac{P(x \mid \theta)P(\theta)}{P(\mathrm{X})}$$

where $$P(\mathrm{X}) = \int P(x \mid \theta)P(\theta)d(\theta)$$

such that $P(\theta)$ is the prior distribution of parameter θ, x shows collected data, $P(\theta|x)$ is the posterior distribution of θ and is known later, given the knowledge of the data. We will discuss more details on this in our analysis and evaluation in the paper.

The factors being considered for analysis include:

- *Home Game Advantage:* This refers to whether the game is played on home grounds or in a different country.
- *Day / Night Effect:* This factor considers the effect of whether the match is played during the daytime or at night.
- *Winning the Toss:* A coin is tossed at the beginning of the game and the captain of the team that

wins the toss decides which team should bat first. This factor refers to whether the team wins the toss.

- *Batting First:* This factor determines whether the concerned team batted first (or bowled first) in the given match.

Using these factors, we have first applied different classification approaches, to predict whether a given team wins the concerned match. Naïve Bayes, Decision Tree Classifiers using C4.5, Bagging and Boosting were the algorithms considered. We have conducted comparative studies among various classifiers and summarized the results in this paper.

Based on these results, we have then developed a tool called CricAI that can be used to determine the probability of victory in an ODI cricket match given the concerned factors as inputs. This tool provides an implementation of Bayesian classifiers which were found to yield the best results in our analysis The CricAI tool is an interesting application of machine learning techniques. It can be of value to real cricketers and cricket analysts in terms of analyzing various scenarios in advance and working towards maximizing their chances of victory.

The rest of this paper is organized as follows. Section 2 explains in detail the approach we have used for conducting the analysis. Section 3 presents a comparative study of the classifiers used. Section 4 describes the implementation of the CricAI tool based on our results. Section 5 presents a critique of related work in the area. Section 6 gives the conclusions.

II. APPROACH FOR ANALYSIS

A Data Collection

For conducting our research, we collected data on all the ODI matches played since 1971 (the year ODI was introduced) till date from the website www.cricinfo.com [4]. The attributes selected were Team, Opponent team, Home/Away, Day/Night, Toss, Bat 1st and results. Each team was analyzed individually against every other team. The reason for doing so is the following. When we have one data set for all matches, one match forms two records in the data set. For instance, if Sri Lanka played against Australia, the data set is as shown in Table 1 below

TABLE I
Sample Data Set of Two Teams which explains the over-fitting situation

Team	Opponent	Toss	Bat1st	Result
Sri Lanka	Australia	Won	Yes	Won
Australia	Sri Lanka	Lost	No	Lost

This created over-fitting results with the given data set. The best way to avoid that was to select one team against other at a time.

B. Choice of Machine Learning Techniques

We tried to use three machine learning techniques: association rule mining, clustering and classification. Selected algorithms from each technique were trained using the WEKA tool. Data set was organized in Attribute Relation File Format (ARFF) as required by WEKA. (Figure 1)

```
%Author:AmalKaluarachchi
%Date:March10th2009
%------------------------------------------------------
@RELATION     Lanka
@ATTRIBUTE    Year                  string
@ATTRIBUTE    Game_Date             string
@ATTRIBUTE    Month                 {Q1,Q2,Q3,Q4}
@ATTRIBUTE    Score                 numeric
@ATTRIBUTE    Wickets               numeric
@ATTRIBUTE    Overs                 numeric
@ATTRIBUTE    RPO                   numeric
@ATTRIBUTE    Inns                  {1,2}
@ATTRIBUTE    Result                {lost,won,no,tied}
@ATTRIBUTE    Opposition {PK,WI,AU,NL,EN,SA,IN}
@ATTRIBUTE    Ground
{Other,Delhi,Leeds,Nairobi,Centurion,Manchester,Pune,TheOval,Welli
ngton,Hyderabad,Kolkata,Mumbai,PortofSpain,SLOth,Hobart,Dhaka,Br
isbane,PSS,Karachi,Lahore,Sydney,Adelaide,Melbourne,Perth,Dambull
a,SSC,RPS,Sharjah}
@ATTRIBUTE    Home                  {Yes,No}
@ATTRIBUTE    Extra                 numeric
@ATTRIBUTE    byes                  numeric
@ATTRIBUTE    leg_Byes              numeric
@ATTRIBUTE    wide                  numeric
@ATTRIBUTE    Noball                numeric
@DATA
1985,26-Jan-85,Q1,204,6,50,4.08,1,lost,WI,Adelaide,No,17,2,9,2,4
1985,28-Jan-85,Q1,91,10,35.5,2.53,2,lost,AU,Adelaide,No,9,1,5,3,0
1985,12-Jan-85,Q1,180,10,48.1,3.73,2,lost,WI,Brisbane,No,8,0,3,3,2
```

Fig 1. Data Set in ARFF Format

Using Apriori algorithm [13] for association rule mining, we could get rules as given below:

Toss=lost Batting=1st => Day/Night=Day

Considering the practical interpretation of this rule, losing a toss and batting first practically does not imply the match to be a day match. However, wining a toss in a day match may imply batting first. Let us consider another rule.

HOME=Y => Day/Night=Day

This rule has 100% support and confidence as the same was produced by the dataset of West Indies where there is no ground for night matches. Thus, the rule is obvious.

Our goal is to find match winning factors. Thus, we would be looking for an Association rule such as: Toss=won Batting=1st => Result=won

On the other hand clustering did not make any contribution to our research either as we dealt with multiple independent attributes, therefore placing them in groups based on their similarity did not seem feasible.

As we expected, classification produced significant results in our research. Our attributes are independent. Furthermore, the result of a cricket game is dichotomous (ignoring a few games with ties). Thus, it is a binary classification.

In machine learning, Naïve Bayes and Decision Trees are two popular classification approaches. We also looked into two other

algorithms, AdaBoost and Bagging. The results obtained were interesting in terms of improving the performance.

The Naive Bayes Classifier technique is based on the Bayes theorem. The weak decision stumps can usually give as good results as C4.5, while boosting C4.5 generally gives the decision-tree algorithm a significant improvement in performance. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Moreover, Naive Bayes classifiers can handle an arbitrary number of independent variables (e.g. Opponent team and Toss) whether continuous or categorical.

Given a set of variables, X = {x1,x2,...,xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1,c2,c...,cd}. In a more familiar language, X forms the predictors and C is the set of categorical levels present in the dependent variable.

Decision trees serve as another simple binary classifier. The term "decision tree" is used in decision theory and project management to depict a series of decisions for choosing alternative activities. The tree is generated using probabilities and benefits of outcomes of the activities.

Practically, AdaBoost has many advantages. It is fast, simple and easy to program. It has no parameters to tune (except for the number of rounds). It requires no prior knowledge about the weak learner and so can be flexibly combined with *any* method for finding weak hypotheses.

III. COMPARATIVE EVALUATION OF CLASSIFIERS

A. Classification Results

The collected data set was trained using Naïve Bayes, Decision Trees, Bagging and Boosting. The results are given in Table II. It was found that Naïve Bayes produced the best results. This is well-justified. Since the selected attributes are independent, we in fact expected Naïve Bayes to perform better in this exercise.

TABLE II
Results of Various Classifiers

Team	Naïve Bayes		Decision Tree		Bagging		Boosting	
	ROC	RMSE	ROC	RMSE	ROC	RMSE	ROC	RMSE
Australia	0.558	0.487	0.537	0.495	0.560	0.488	0.560	0.486
England	0.574	0.498	0.515	0.509	0.526	0.511	0.576	0.497
India	0.587	0.493	0.517	0.504	0.574	0.496	0.517	0.504
New Zealand	0.655	0.478	0.559	0.494	0.624	0.487	0.638	0.480
Pakistan	0.581	0.494	0.556	0.505	0.584	0.498	0.584	0.494
South Africa	0.590	0.486	0.534	0.510	0.575	0.494	0.588	0.485
Sri Lanka	0.613	0.482	0.551	0.484	0.589	0.494	0.620	0.481
West Indies	0.589	0.489	0.602	0.491	0.627	0.490	0.584	0.485
Average	0.593	0.488	0.546	0.499	0.582	0.495	0.583	0.489

Comparison of different classification techniques was performed using Receiver Operator Curve (ROC) and Root Mean Squared Error (RMSE). Better learning techniques produce highest ROC and lowest RMSE.

As it can be seen from this Figure 3, Boosting and Bagging did not perform significantly well in our research.

B. Analyzing Winning Factors

As Naïve Bayes produced the best result on the selected datasets, our analysis of match winning factors was done using Naïve Bayes. The concerned Naïve Bayes formulae are as follows.

Let P(θ) be the prior distribution of some parameter, θ. This depicts what is known about θ before the data, x, are collected. P(θ |x) is the posterior distribution of θ, and is what is known later, given the knowledge of the data. Bayes' Theorem for a single continuous random variable is then:[5]

$$P(\theta \mid \mathrm{X}) = \frac{P(x \mid \theta)P(\theta)}{P(\mathrm{X})}$$

Where,

$$P(\mathrm{X}) = \int P(x \mid \theta)P(\theta)d(\theta)$$

Accordingly, the Naïve Bayes' formula for multivariate variables can be written as:

$$P(\theta_m \mid \mathrm{X}) = P(x_1 \mid \theta_m)P(x_2 \mid \theta_m)P(x_3 \mid \theta_m)....P(x_n \mid \theta_m)\frac{P(\theta_m)}{P(\mathrm{X})}$$

Using this formula of Naïve Bayes for multivariate variables, the probability of winning a match P(Won|X) for given criteria can be calculated as follows;

Let assume the given criteria as follows:
- Opponent team = Australia (AU)
- Day night = N
- Played = Home
- Toss = Won
- Bat = 1st
- Month = Q1

Thus,

PoP= P(Won|X)

PrP Won = P(AU|Won) x P(N|won) x P(Home|Won) x P(Toss|Won) x P(1st|won) x P(Q1|Won)

PrP Lost = P(AU|Lost) x P(N|Lost)xP(Home|Lost) x P(Toss|Lost) xP(1st|Lost) x P(Q1|Lost)

P(X) = Prp Won + PrP Lost

$$P(Won|X) = \frac{PrP\ Won\ \ x\ P(Won)}{P(X)}$$

PoP = Posterior Probability, PrP = Prior Probability

Applying this formula to historical match data collected, we tried to study some match winning factors such as home game advantage, winning toss, day night effect etc. (Table III-V).

C. Home Game Advantage

The analysis of this factor is shown in Table III. This illustrates that except for Australia and South Africa, all the other teams have higher probability of wining in home grounds.

TABLE III
Analysis of Home Game Advantage

Team	Home		Away	
	Won	Lost	Won	Lost
Australia	60.33%	39.68%	57.76%	42.24%
England	53.49%	46.51%	38.11%	61.89%
India	53.05%	46.95%	39.47%	60.53%
Pakistan	54.28%	45.72%	47.43%	52.58%
New Zealand	52.71%	47.29%	30.06%	69.94%
South Africa	70.40%	29.60%	51.13%	48.87%
Sri Lanka	61.53%	38.47%	33.47%	66.53%
West Indies	62.57%	37.43%	56.59%	43.41%

D. Winning Toss and Batting First

This factor is analyzed as shown in Table IV. As the figure indicates, wining toss does not have major impact on the match outcome. It produced mixed output for each team. Some teams (e.g., for Pakistan winning possibility increased from 54% to 56%) it made positive impacts while for other have negative impact (E.g. for England wining possibility decreased from 53% to 49%).

TABLE IV
Analysis of winning toss and batting 1st

Team	Won Toss		Lost Toss	
	Won	Lost	Won	Lost
Australia	60.33%	39.68%	60.07%	39.93%
England	53.49%	46.51%	49.02%	50.98%
India	53.05%	46.95%	49.64%	50.36%
Pakistan	54.28%	45.72%	56.10%	43.90%
New Zealand	52.71%	47.29%	56.19%	43.81%
South Africa	70.40%	29.60%	68.17%	31.83%
Sri Lanka	61.53%	38.47%	60.53%	39.47%
West Indies	62.57%	37.43%	60.23%	39.77%

E. Winning Toss and Batting Second

The results of analyzing this factor are shown in Table V. Comparing Table IV and V we can clearly see that winning toss and batting second reduce the chance of winning. Interestingly, losing toss and batting 2nd increases chances of winning.

TABLE V
Analysis of winning toss and batting 2nd

Team	Won Toss		Lost Toss	
	Won	Lost	Won	Lost
Australia	57.09%	39.68%	56.83%	43.17%
England	65.98%	34.02%	61.86%	38.14%
India	59.88%	40.12%	56.56%	43.44%
Pakistan	52.50%	47.50%	54.33%	45.67%
New Zealand	60.28%	39.72%	63.59%	36.41%
South Africa	71.97%	28.04%	69.80%	30.20%
Sri Lanka	61.53%	38.47%	56.75%	43.25%
West Indies	62.57%	37.43%	64.62%	35.38%

Using Naïve Bayes formulae with different combinations of attributes, probability tables (similar to Table IV and V) were generated. Those tables helped us to study the impact of those attribute combinations on outcome of the match.

Since our research produced some interesting results, we moved to the next step of developing an application to predict out come of future matches using the knowledge discovered from our research.

IV. IMPLEMENTATION OF THE CRICAI TOOL

The CricAI software tool is developed based on the results of our detailed analysis with real data. It is a Java implementation of Naïve Bayes, since Bayesian classifiers yielded the best results in our analysis. MySQL is used as the database to store match related data.

CricAI can be used to predict the outcome of an ODI cricket match even before the match has started. All the required information is made available to this tool at the beginning of the match. (e.g., two teams, toss etc.). Upon selecting different attribute values, the system produces the probability of wining and loosing the match using Naïve Bayes formula. The calculation is real time and no calculated information is stored in the system. We implemented it that way as we can add change more attribute to the system with minor changes to the program. A snapshot of this tool appears in Figure 2.

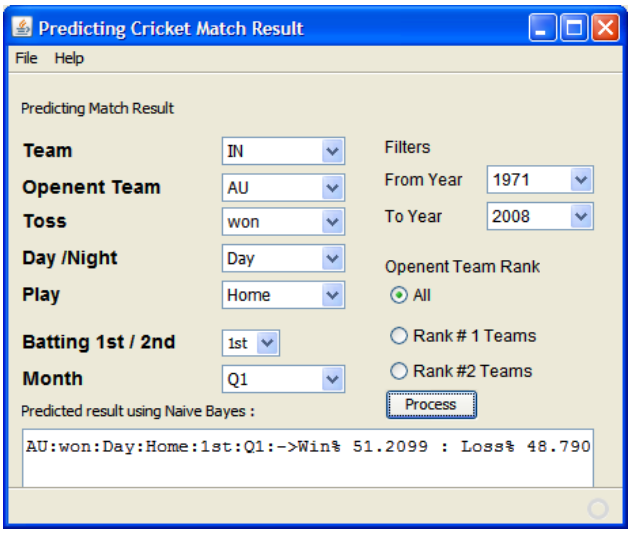

Fig. 2 Snapshot of CricAI tool

The CricAI tool has some filters to obtain better results. Currently all international teams are almost equally talented. Winning a game therefore could be based on the factors and decisions made by the individuals (e.g., winning a toss and selecting batting first). However, in the early stage of the game just after the game was introduced in 1970s, some teams were far more talented than the others. Thus, including those matches for calculation could have adverse effects on the results. The Year filter helps overcome this drawback as shown in Figure 3.

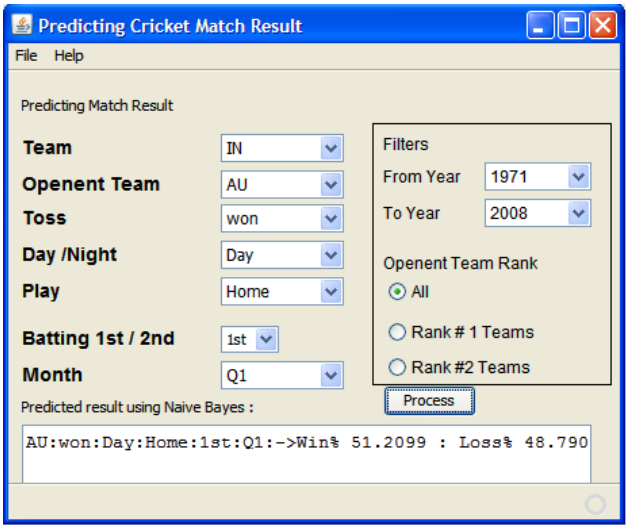

Fig 3. Using filters in CricAI

Another filter is selecting the rank of the opponent. The International Cricket Council (ICC) team ranking is used. This also helps make better decisions on the match winning criteria. For example, Sri Lanka has not lost any match to Rank 2 teams such as Canada and the Netherlands.

V. RELATED WORK

From our literature survey, we found that very limited machine learning work has been done on game of cricket. Though cricket shares some attributes with other sports such as baseball, it still remains unique in certain respects and deserves to be analyzed independently. Most of analyzing studies on cricket so far have been conducted using statistical methods. Furthermore, many of them have addressed the five day long test matches but not the One Day Internationals. We present some relevant studies below.

The statistical research on Cricket has been started very early stage of the cricket. In 1945 Wood used the geometric distribution to model the total score in cricket [6]. This was not a study on the ODI form of the game but has been recognized among the pioneering research in the game of cricket.

Bailey and Clarke conducted a study to predict the outcome in one day international cricket while the game is in progress [10]. This study was performed using statistical models. The interesting fact about this article is that the authors have statistically proved how the match resources (number of overs and batsmen left) affect the final result. However, they deal with analysis during the current game. They do not predict in advance the chances of winning a new game based on previous matches.

Chedzoy studied the issue of umpiring errors in cricket matches [11]. An umpire is the term used for a referee in cricket. This article focuses on umpiring decisions and how they affect the outcome of the match. This study was also based on a statistical approach. Moreover, it focused only on one aspect of the game, namely, the effect of umpires.

Sparks and Abrahamson developed a mathematical model to predict award winners [14] in a game. This study was conducted using machine learning techniques and focused on baseball matches. They have employed this model in the national league and correctly predicted winners prior to the award announcement. Smith and Lipscomb [9] have done similar study for Predicting CY Young award winners for Baseball Pitchers. Interestingly, they have found that Naïve Bayes performed well in their research as well.

Bandulasiri [1] has written an interesting article on predicting the winner in an ODI cricket match. This article addressed similar datasets as we used in our research. In this paper, the author has used statistical methods to find wining factors for an ODI match. We have explored the machine learning path, considering popular classifiers, and developed a software tool based on our results. This AI-based tool would be very helpful in predictive analysis in cricket.

VI. CONCLUSIONS

In this article, we have addressed the problem of predicting the chances of victory in a One Day International cricket match. By analyzing different attributes related to the ODI game, we have been able to predict the winning criteria formulated using attributes from the dataset. We have developed a software tool called CricAI based on our study.

The main contributions of our work are:

- Comparison of machine learning techniques which revealed that classification is the best approach to solve the problem.
- Evaluation of various classifiers over real data which proved that Naïve Bayes works best over the concerned datasets.
- Analysis of individual factors that affect the outcome of the game which for example, showed that winning the toss (the most controversial factor in the cricket community), does not have a major impact on the outcome of the match.
- Development of the CricAl tool that can be used in real-world scenarios to predict the chances of victory in a given match, using attributes and filters.

As future work, we are planning to expand our analysis using more attributes such as the previous match result of the selected team and the opponent team, the number of known batsmen in the selected team and the opponent team and more. It is also possible to apply the machine learning techniques we used in our research to predict the outcome in other outdoor sports such as baseball. The specific approach used may depend on the nature of the given datasets and applications.

REFERENCES

[1] A. Bandulasiri, "Predicting the Winner in One Day International Cricket", *Journal of Mathematical Sciences & Mathematics Education, Vol. 3, No. 1.*

[2] A. Ceglar and J.F. Roddick, "Association mining", ACM Computing Surveys, 2006 Vol. 38, No. 2.

[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] C. M Bishop, "Pattern Recognition and Machine Learning", Springer New York, 2006.

[6] G.H. Wood, "Cricket scores and geometrical progression", Journal of the Royal Statistical Society, 1945, Series A, 108: pp. 12–22.

[7] I. Preston and J. Thomas "Batting Strategy in Limited Overs Cricket", The Statistician (Journal of the Royal Statistical Society: Series D), 2000, 49, 95-106.

[8] I. Witten and E. Frank, "Data Mining: Practical Machine Learning Algorithms with Java Implementations", Morgan Kaufman Publishers, California, USA (2000).

[9] L. Smith, B. Lipscomb, and A. Simkins, "Data Mining in Sports: Predicting CY Young Award CCSC", Central Plains Conference April 13 - 14, 2007.

[10] M. Bailey and S.R. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress", *Journal of Sports Science and Medicine,* 2006, Vol. 5, pp. 480-487.

[11] O.B. Chedzoy, "Issue of the effect of umpiring errors in cricket Statistician", 1997, Vol. 46, No. 4, pp. 459-527.

[12] P. Lutu, "An Integrated Approach for Scaling up Classification and Prediction Algorithms for Data Mining", Proceedings of the annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, 2002.

[13] R. Agrawal, T. Imielinski and A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases" SIGMOD. June 1993, pp. 207-216.

[14] R. Sparks and D. Abrahamson, "A mathematical model to predict award winners", Math Horizons, April 2005, 5-13

[15] S.R. Clarke, "Dynamic programming in one-day cricket—optimal scoring rates", *Journal of the Operational Research Society, 1988, Vol. 39, No. pp. 331–337.*

[16] University of Waikato, WEKA, Waikato Environment for Knowledge Analysis.

[17] Wikipedia on the Game of Cricket, website [Online] http://en.wikipedia.org/wiki/Cricket