Vector Quantization, Density Estimation and Outlier Detection on Cricket Dataset

Kamalaruban Parameswaran
Department of Electronic & Telecommunication Engineering
University of Moratuwa
Sri Lanka
pkamalaruban@gmail.com

Abstract—**This study aims to apply unsupervised machine learning algorithms on Cricket players' career statistics dataset. K-means clustering algorithm is used to find the natural grouping that exists within the cricket players using player's batting average, strike rate, bowling average, economy etc. as input features – in this case players are grouped into 3 groups. Further separate probability density models are fitted for batsmen, bowlers and all-rounding players using appropriate player's performance metrics as input features and using these models, outstanding players are identified. Similar method is used to identify match winning players, where the differences between player's performance metrics and team's average performance metrics are used as input features. The results obtained from this study seem to correlate with expert generated results where they used point based system to rank the players. This kind of statistical analysis of sports data plays a vital role in team planning and exploiting opponents' weakness.**

Keywords- Cricket; Density estimation; K-means clustering; Outlier detection

I. INTRODUCTION

Cricket is a very popular game in the subcontinent. One-Day Internationals (ODI) is a version of cricket that is completed in one day, as distinct from Test cricket which can take up to five days to complete. Batting, bowling and fielding are the three main elements of cricket; depending on the player's level of contribution in these activities they are identified as a batsman or a bowler or an all-rounding player. However in cricket there is no hard and fast rule to assign a player into one of the above categories. Basically it depends on the analyst's view therefore we can consider it as a qualitative measurement.

Table I describes some of the performance metrics that are used in the clustering algorithm and outlier detection process. These metrics are related to the individual player performances. For match winning player detection, team's average performance metrics like team's batting average, run rate, bowling average, opposing team's run rate etc. are also taken into account.

This paper presents the application of unsupervised machine learning techniques on cricket players' career statistics dataset to find the natural grouping of players and to identify outstanding and match winning players.

TABLE I. INDIVIDUAL PLAYER PERFORMANCE METRICS

Performance metric	*Measurement*
Batting average	Player's consistent run scoring ability
Strike rate	Player's rapid scoring ability
Bowling average	Player's bowling efficiency
Economy	Player's ability to restrict the batsman
Batting innings percentage	For regular batsman this will be high
Bowling innings percentage	For regular bowler this will be high

Since classifying a player is a qualitative process, I did not attempt to label players manually; rather I attempt to find the natural clustering that exists within the players automatically. Thus this automatic clustering will group players as the ones who have contributed mostly with bat only, with ball only and with both.

Squad selection decisions and award nomination are some important processes which need much care and in depth analysis. International Cricket Council (ICC) and related bodies annually hold award ceremony for cricketers – ICC normally ranks the players using a point based system which mainly depends on the individual performance of the player [8]. However this point system does not take into consideration how much that player stands out from the others from statistical point of view.

Thus in this study separate probability density models are developed for batsmen, bowlers and all-rounding players using appropriate input features for each model. Then the outliers of each model are identified separately and ranked based on their probability. Appropriate domain knowledge based filtering is applied on the outliers to make sure that the filtered outliers are in fact outstanding players. It is found that the obtained results mostly comply with some expert generated rank list. As this system effectively measures how outstanding a player is, this could be used in award nomination process.

During the squad selection process, management needs to identify the key players of their team and when a team plans its game plan against its opponent team in a match it needs to identify the key players of the opponent team. Thus identifying outstanding match winning players (batsman, bowler) of each team would be very useful information. For this type outlier detection process, the difference between player's performance

978-1-4673-2907-1/13/$31.00 ©2013 IEEE

metrics and team's average performance metrics are used as input features. A proper squad selection with the right mix of batsmen, bowlers and all-rounding player will greatly determine the success of the team.

II. LITERATURE REVIEW

Cluster analysis is the organization of a collection of patterns into clusters based on similarity [1], thus the clustering algorithms are used in several applications like image segmentation, news grouping and object recognition etc. K-means clustering algorithm is one of the several clustering techniques that minimize the within cluster sum of squares [2].

For many applications like identification of system faults, network anomaly detection and exceptional human behavior detection etc., the discovery of outliers leads to more interesting and useful results than the discovery of inliers [3, 4]. Density-based outlier detection methods estimate the density distribution of the input space and then identify outliers as those lying in regions of low density [5].

There are studies that have been performed on cricket dataset to predict outcome of a cricket match using neural networks [6] and Bayesian classifiers [7]. However these studies did not analyse the cricket players' career statistics dataset, thus in this paper unsupervised machine learning techniques are applied on the players' career statistics dataset to extract some useful information.

III. METHODOLOGY

The dataset used for this project is collected from ESPN Cricinfo website [9]. For K-means clustering and outstanding batsmen, bowlers and all-rounding players detection I have used career statistics of each player such as batting average, strike rate, bowling average, economy etc. For the match winning player detection I have used player's statistics and team's aggregate statistics in the won matches in which the player also played.

I have applied K-means clustering algorithm on players' career statistics dataset to group them into three clusters. This algorithm does not require any supervision to find the natural clustering in the input data. Intuitively the algorithm tries to cluster players as the ones who have fair amount of contribution with bat only or with ball only or with both. However the K-means algorithm does not guarantee to find the global minimum error, but will certainly find a local minimum. To overcome this problem I have used the standard K-means Assignment-Update algorithm with random restarts as described below.

Input:

player dataset $= \{player^{(1)}, player^{(2)}, \ldots player^{(i)}, \ldots, player^{(m)}\}$

where $player^{(i)}$ *= (batting average, strike rate, bowling average, economy, batting innings percentage, bowling innings percentage)*

Algorithm:

for $j = 1$ *to several random initializations*

randomly choose $K = 3$ *centroids :* $\{centroid^{(1)}, centroid^{(2)}, centroid^{(3)}\}$

loop until cost function stop decreasing:

for each $player^{(i)}$*:*

$cluster\_of\_player^{(i)} = arg\ min_k \| player^{(i)} - centroid^{(k)} \|$

compute new centroids based on new players allocated

$$cost = \sum_{1}^{m} \| player^{(i)} - cluster\_of\_player^{(i)} \|$$

choose the clustering with minimum cost

Input features are normalized prior to clustering in order to compensate the scale variations in the input data. After normalization **more weight** is given to batting innings percentage and bowling innings percentage features – this would help in overcoming exceptional occasions like one who is not a regular bowler who bowled in very few matches possessing very good bowling figures(and vice versa for one who is not a regular batsman).

After clustering the clusters are named automatically based on the average *batting innings percentage* and average *bowling innings percentage* for all K clusters as follows,

- Cluster with both average *batting innings percentage* and *bowling innings percentage* greater than 70% is named as *all-rounding cluster*.
- Cluster with only average *batting innings percentage* greater than 70% is named as *batting cluster*.
- Cluster with only average *bowling innings percentage* greater than 70% is named as *bowling cluster*.

I have used this automatic cluster assignment information in the following parts like density estimation and outlier detection on players' career statistics dataset.

Then I have used the above clustered output to develop separate probability models for batsmen, bowlers and all-rounding players using appropriate input features, in order to detect outstanding batsmen, bowlers and all-rounding players from the respective models. In order to develop a model for batsmen I have considered players from *batting cluster* and *all-rounding cluster*, for bowlers I have considered players from *bowling cluster* and *all-rounding cluster*, and for all-rounding players I have considered players only from *all-rounding cluster*. For the batsmen model I have used batting average and strike rate as the input features, for the bowlers model I have used bowling average and economy, and for the all-rounding players model all these four metrics are used as input features.

I have developed these separate probability density models assuming that the input data is approximately distributed according to Gaussian distribution. By plotting the histogram of each input features that I have used in the density estimation process, I have ensured that the above assumption is indeed valid. Incase if any input features do not obey Gaussian

behavior we may have to do some conversion like 'log (input feature + constant)', on those input features to make sure that they are distributed according to Gaussian distribution. In these experiments all the input features that I have used, approximately distributed according to Gaussian distribution. Algorithm for fitting a multivariate Gaussian density model for batsmen is described below.

Input:

batsmen dataset = {batsman$^{(1)}$, batsman$^{(2)}$, ... batsman$^{(i)}$, ... , batsman$^{(m)}$}

where batsman$^{(i)}$ = (batting average, strike rate)

Algorithm:

$$\mu = \frac{1}{m}\sum_{1}^{m} batsman^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{1}^{m} (batsman^{(i)} - \mu)(batsman^{(i)} - \mu)^T$$

$$p(batsman) = \frac{1}{2\pi |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(batsman - \mu)^T \Sigma^{-1} (batsman - \mu) \right)$$

The mean and variance parameters for each input features could have been calculated separately and constructed the model from the independent probability distribution of each features when there is negligible correlation between input features. In this case both approach produced almost similar results.

Then I have used the above developed models to figure out the outstanding players. But this problem is different from regular anomaly detection problem where we have labeled anomalous and non anomalous dataset which can be cross validated and tested using F-score. Where as in this case we can not exactly say whether a player is really outstanding since it is a qualitative measure – so I do not attempt to cross validate or test the results using F-score, rather I have tried to rank the players based on the probability that is assigned to them in the fitted density model i.e. the lower the probability, more the being outlier, thus higher the rank. However when we attempt to detect outliers using low probability points, we will get unwanted outliers also (who are really not outstanding players) – thus we have to filter outliers appropriately in order to figure out outstanding players correctly.

Filtering for Outstanding batsman:

- batting average > mean(batting average) – lower limit constant for batting average
- strike rate > mean(strike rate) – lower limit constant for strike rate

Filtering for Outstanding bowler:

- bowling average < mean(bowling average) + lower limit constant for bowling average
- economy < mean(economy) + lower limit constant for economy

This rank list is qualitatively compared with some expert (ICC, Star cricket etc) generated rank lists which use some other methodologies to rank the players.

I have used the similar approach to find the match winning batsmen and bowlers, where the input features are the differences between player's performance metrics and team's average performance metrics, for example for match winning batsmen detection batting average difference (player's *batting average* – team's *average runs per wicket*) and strike rate difference (player's *strike rate* – team's *runs per over* x 100/6) in the won matches are used as input features.

However players who have played very few batting or bowling innings in the won matches could posses very good batting or bowling figures – ultimately end up as an outstanding match winning player during the outlier detection process - which is unreasonable. Thus prior to density estimation process this kind of entries should be removed from the data set. Actually what I do in this outlier detection is identifying dominant batsman/bowler in teams' point of view.

IV. Results and Discussion

Career statistics of 186 current cricketers who had played at least 20 ODI matches is selected as the input dataset. By applying K-means clustering algorithm on this dataset; 52 players are assigned to the *all-rounding cluster*, 71 players are assigned to the *batting cluster* and 63 players are assigned to the *bowling cluster*. By careful inspection it has been found that the clustering seems to correlate with some expert clustered list like ICC ODI batsman, bowler and all rounder lists.

Then this clustered data is used to fit probability density models for batsmen, bowlers and all-rounding players separately. Then appropriately filtered outliers from these models are ranked based on the probability assigned to them in these models. The outstanding batsmen, bowlers and all-rounding players list generated by the outlier detection process seems to tally with ICC ODI players rank list [10]. Gaussian density models fitted for batsmen, bowlers, match winning batsmen and match winning bowlers are illustrated in Figure 1. The filtered outstanding players of each category are depicted by circles in the respective models. In the batsmen model, players who are positioned higher along the strike rate axis are well known hitters, whereas players who are positioned higher along the batting average are more consistent players.

The TABLE II presents the all-rounding player rank list obtained from the outlier detection process – most of the players in this rank list are found in ICC top 10 all-rounding players list.

V. Conclusion

In this study I have applied several unsupervised machine learning techniques like K-means clustering, density estimation and outlier detection on cricket players' career statistics dataset to automatically group the players based on their level of contribution in batting and bowling, and to detect the

outstanding players and dominant team members. It has been found that the results obtained from this study seem to correlate with some expert generated rank lists where they have used different qualitative approach to rank the players.

REFERENCES

[1] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* Vol. 31, No. 3, September 1999.

[2] D. Pollard, "Strong Consistency of K-Means Clustering," *The Anals of Statistics*, 1981, Vol. 9, No. 1, 135-140

[3] P. Gogoi, D.K. Bhattacharyya, B. Borah, and J.K. Kalita. A survey of outlier detection methods in network anomaly identification, *The Computer Journal,* 2011.

[4] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies", *Artificial Intelligence Review, Springer,* 2004, Vol. 22, 85-126

[5] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In SIGMOD'00.

[6] D. Roy Choudhury, Preeti Bhargava, Reena, Samta Kain , "Use of Artificial Neural Networks for Predicting the Outcome of Cricket Tournaments", *ISSN 1750-9823,* 2007, Vol. 1, No. 2, 87-96

[7] A. Kaluarachchi, Aparna S. Varde, "CricAI: A classification based tool to predict ODI matches", 5th International Conference on Information and Automation for Sustainability, 2010.

[8] http://www.relianceiccrankings.com/about.php

[9] http://www.espncricinfo.com/

[10] http://www.relianceiccrankings.com/

TABLE II. OUTSTANDING ALL-ROUNDING PLAYERS

Name	**Team**	**Batting average**	**Strike rate**	**Bowling average**	**Economy**
AD Russell	West Indies	36.62	123.62	27.86	5.48
Mohammad Hafeez	Pakistan	26.9	67.84	32.77	4.06
Shakib Al Hasan	Bangaladesh	35.63	78.07	28.85	4.29
JH Kallis	south affrica	45.26	72.97	31.69	4.82
Shahid Afridi	Pakistan	23.48	113.79	33.51	4.6
DJG Sammy	West Indies	21.26	104.15	45.13	4.59
SR Watson	Australia	41.48	88.27	28.83	4.8
CH Gayle	West Indies	39.43	84.45	35.08	4.73
KA Pollard	West Indies	25.8	97.86	35.55	5.38
Abdul Razzaq	Pakistan	29.7	81.25	31.83	4.69

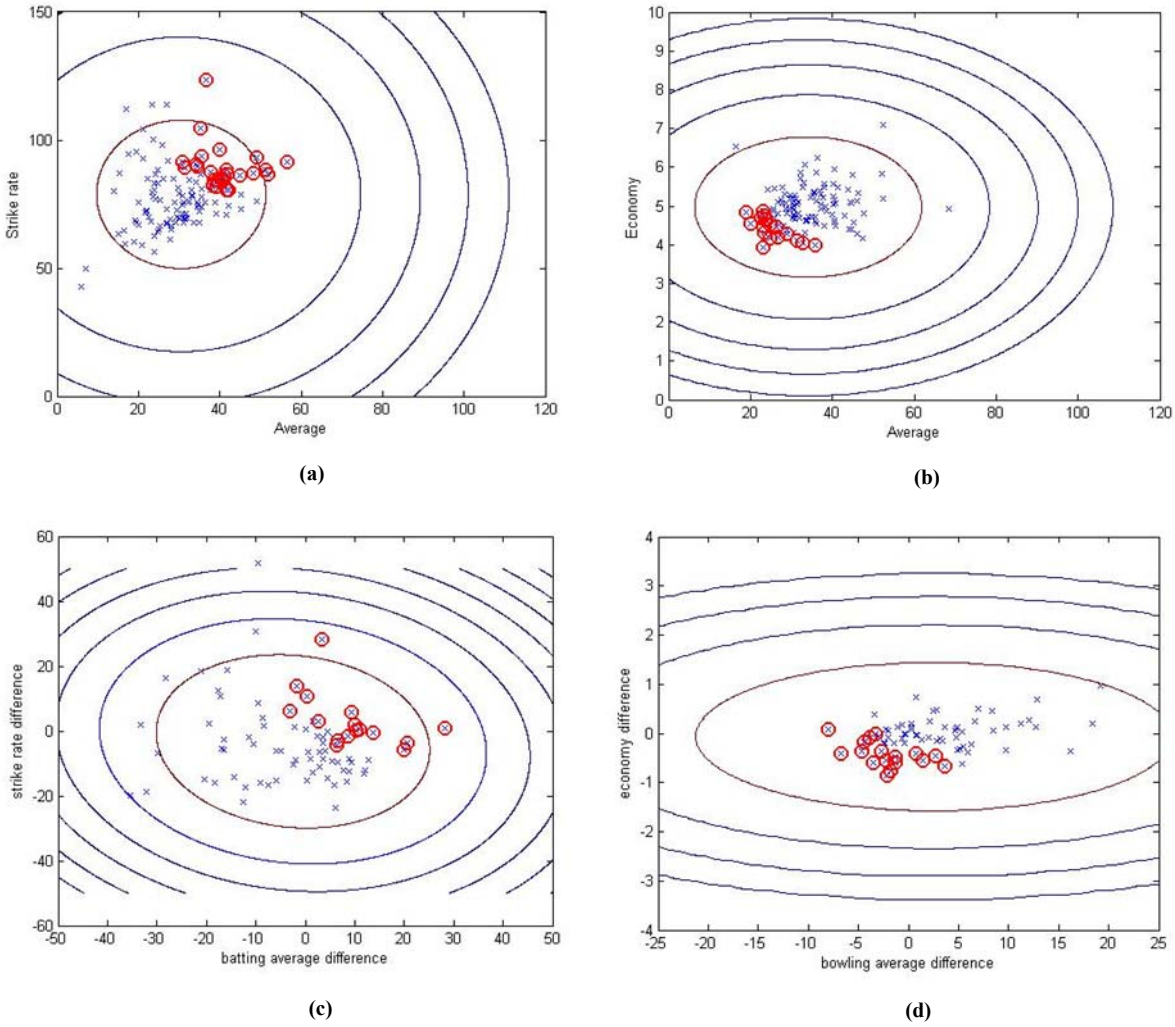

Figure 1. Gaussian density models for (a) batsmen, (b) bowlers, (c) match winning batsmen, and (d) match winning bowlers.