Cricket-Insights Project Deliverable Predictive Analytics using Association Rules

Abhinav Chawla(IMT2013002)　Aditya Naidu(IMT2013003)

Shivam Kumar(IMT2013042)　Siddartha Padhi(IMT2013043)

1 Introduction

This document presents the results that we obtained for predictive analysis using association rules on our dataset. We used the apriori algorithm using the apriori function call in R from arules library. We analyzed the rules obtained by varying parameters such as minimum length, minimum support and/or minimum confidence. We also tried to obtain association rules by forcing certain constraints on L.H.S and/or R.H.S for the rules. In the subsequent sections, we describe the data preparation phase, the rules that we obtained for our dataset and their interpretations.

2 Data Preparation

For our previous deliverables, we had already extracted both quantitative and categorical data from our raw data using several scripts written in R and python. Further processing was required before we could use the available data for deriving the association rules. In case of columns with categorical data, processing was trivial as the values in such columns was just prepended with the column name itself. For quantitative data however, processing was carried out in the following two phases:

1. **Discretization**: Quantitative columns were converted to columns with discrete values of HIGH, MEDIUM or LOW. Initially, we used the default discretize function call in R to make our data discrete but it did not give useful association rules in many of the cases. We then discretized the columns based on our domain knowledge. For example, An individual score of 30 and above in IPL was categorized as HIGH while below that was categorized as LOW.

2. **Prefixation**: Discretized values were then prepended with the column name as was done for categorical data.

3 Results

After data preparation, we generated association rules and picked those whose confidence was atleast 0.6. Generated rules were sorted in a descending order based on lift for analysis. The rules and their interpretations are presented below.

Rule	Support	Confidence	Lift
op_partner_name=CH Gayle-TM Dilshan, partnership_label=HIGH ⇒ op_winner=Royal Challengers Bangalore	0.008591065	0.8333333	7.028986
op_partner_name=CH Gayle-TM Dilshan ⇒ op_winner=Royal Challengers Bangalore	0.012027491	0.7368421	6.2151030
op_partner_name=CH Gayle-TM Dilshan ⇒ partnership_label=HIGH	0.010309278	0.6315789	1.6520402
Interpretation: If Gayle and Dilshan open for a match, they tend to have a high partnership and if they do, RCB will most likely win.			

Rule	Support	Confidence	Lift
op_partner_name=CH Gayle-V Kohli, partnership_label=HIGH ⇒ op_winner=Royal Challengers Bangalore	0.007731959	0.75	6.3260870
op_partner_name=CH Gayle-V Kohli ⇒ op_winner=Royal Challengers Bangalore	0.012886598	0.60	5.0608696
Interpretation: If Gayle and Kohli open for a match and have a high partnership, RCB will most likely win.			

Rule	Support	Confidence	Lift
Toss=Toss_Gujarat Lions ⇒ Elected_to=field	0.01441441	1	1.843854
Interpretation: If Gujarat Lions win the toss, they always elect to field.			

Rule	Support	Confidence	Lift
Toss=Toss_Royal Challengers Bangalore,Elected_to=field, City=Bangalore ⇒ Win_team=win_team_Royal Challengers Bangalore	0.01981982	0.6111111	5.062189
City=Bangalore ⇒ Elected_to=field	0.08468468	0.8545455	1.575657
Interpretation 1: If RCB wins the toss in Bangalore and elects to field, they are likely to win.			
Interpretation 2: Teams who win the toss at Bangalore usually elect to field first.			

Rule	Support	Confidence	Lift
Toss=Toss_Pune Warriors,Elected_to=field ⇒ loser=loser_Pune Warriors	0.01621622	1.0000000	16.818182
Interpretation: If Pune Warriors win the toss and elect to field, they always lose the match.			

4 Conclusion

We were able to relate many of the association rules to the trends that we derived in the descriptive analysis phase. Association rules are a useful tool in deriving trends from the available data and can be used for predictive analysis.