MLC

Contents

[Introduction 1](#_Toc361662055)

[Objective 1](#_Toc361662056)

[Requirements 1](#_Toc361662057)

[Library Folder 2](#_Toc361662058)

[How Do I Use MLC? 4](#_Toc361662059)

# Introduction

MLC is a windows application. It is a consolidation of machine learning algorithms. Its primary purpose is to predict the outcome of soccer matches (primarily at the moment the English premiership).

At the time of writing MLC is a prototype. I’m still investigating how to improve the classification.

# Objective

Fine tune a classifier, such that, it that can accurately predict the outcome (home win, draw or away win) and make a fortune with the bookies x-match accumulators, where “x” is the number of games that are predicted.

Accumulators are low-risk high-gain betting strategies offered by many bookies. For example, for a bet of €5 on a 10 game outcome the potential returns can be very high.

# Requirements

Visual studio 2010. Also like any project MLC requires some external dlls to help it run. The are shown in screen shot Figure 1:

Nunit v2.6.1 (<http://www.nunit.org/>) <-- Requires .NET runtime v2.050727

Weka v3.6 (<http://www.cs.waikato.ac.nz/ml/weka/>) <-- Requires JRE v1.5 or v1.6 or v1.7

IKVM (<http://www.ikvm.net/> and also <http://sourceforge.net/apps/mediawiki/ikvm/index.php?title=Installation>)

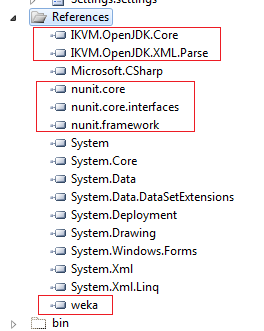


Figure MLC references

# Library Folder

The library folder is important. It contains historical data files for premiership games (essentially the .txt files).

It also contains .arff files which are files that the Weka windows application can load and evaluate. The MLC application can converts the .txt file into .arff file. Figure 2 below…

The format of these files is a bit important but readable. Figure 2 below shows the attributes that make up this particular file. In this case the fixture is Blackburn at home versus Wigan. Next come 16 in-game statistics like: full time home team goals, full time away team goals, half time home team goals, half time away team goals, home team shots, away team shots etc…After these come the bookies odds from a home win, draw or an away win. In this particular case I have included 3 bookies – Bet365, Ladbrooks and finally William Hill. Finally the last column is the actual outcome, which in this case is an away win.

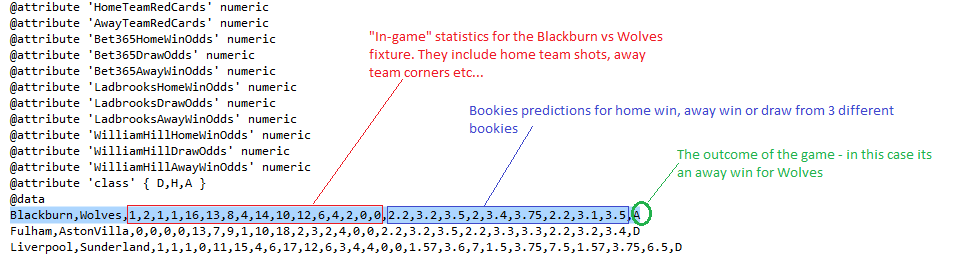


Figure ARFF file format showing the types of game attributes

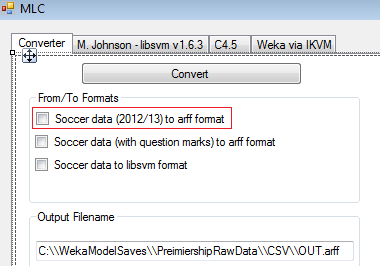
****

Figure MLC showing how to convert .txt to .arff format

# How Do I Use MLC?

As MLC is a prototype experimental area at the moment it can be tricky to get you head around it. Probably the best area to see it running is use the “Run Weka” button on the 4 tab called: “Weka via IKVM”. Screen shot below.

This is fairly typical output of classified data. In the example below the training data in this case is the file: “Prem12to08\_Train\_3Bookies.arff”. This means that the classifier is learning from premiership data from the years 2012 to 2008 and this data contains 16 in-game attributes (stuff like home team corners etc...) and also the bookies predictions form that particular fixture from 3 different bookies (and of course the outcome of the fixture). Again, in the output below the test data i.e. the data requiring prediction

I have added some extra output parameters to reveal more information concerning misclassifications.

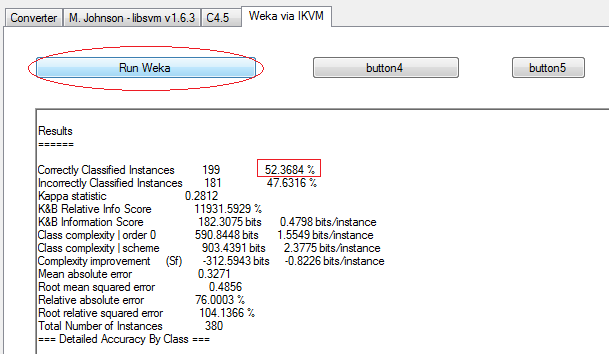


Figure MLC calling the Weka API and classifying some data. Typical output...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 3 yrs (2012-09)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 51.58 |
| 3 yrs (2012-09)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 51.05 |
|  |  |  |  |  |
| 4 yrs (2012-09)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 51.32 |
| 4 yrs (2012-09)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 52.37 |
|  |  |  |  |  |
| 5 yrs (2012-08)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 53.42 |
| 5 yrs (2012-08)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 52.36 |
|  |  |  |  |  |
| 6 yrs (2012-07)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 52.63 |
| 6 yrs (2012-07)  21 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 52.63 |
|  |  |  |  |  |
| 12 yrs (2012-01) 0 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 51.58 |
| 12 yrs (2012-01)  0 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 52.89 |
|  |  |  |  |  |
| 12 yrs (2012-01) 15 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Supervised descritization | 51.58 |
| 12 yrs (2012-01)  15 macro-game  3 bookies | 1 yr (2013)  3 bookies only | NaiveBayes | Kernel estimator | 52.89 |

With “Match Rating” data – calculated for each individual season...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  15 macro-game  3 bookies  1 match rating | 1 yr (2013)  3 bookies only  1 match rating | NaiveBayes | Supervised descritization | 51.72 |
| 5 yrs (2012-08)  15 macro-game  3 bookies  1 match rating | 1 yr (2013)  3 bookies only  1 match rating | NaiveBayes | Kernel estimator | 51.978 |

**Training Data:** Prem12to08With3BookiesPlusMatchRating\_Training.arff

**Test Data:** Prem13With3BookiesPlusMatchRating\_Test.arff

I wanted to investigate what happens when the 15 macro-game statistics are included as questions marks. That is, running with the 3 bookies and the match rating statistic. Results are below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  3 bookies  1 match rating | 1 yr (2013)  3 bookies only  1 match rating | NaiveBayes | Supervised descritization | 51.72 |
| 5 yrs (2012-08)  3 bookies  1 match rating | 1 yr (2013)  3 bookies only  1 match rating | NaiveBayes | Kernel estimator | 51.978 |

**Training Data:**Prem12to08WithQuestionsWith3BookiesPlusMatchRating\_Training.arff

**Test Data:** Prem13With3BookiesPlusMatchRating\_Test.arff

Conclusion: The presence of the 16 macro-game statistics do not improve the classifier!

Now running with that these I wanted to investigate what happens when the 15 macro-game statistics are included as questions marks and also the match rating. That is, running with the 3 bookies data only...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  3 bookies  ? match rating | 1 yr (2013)  3 bookies only  ? match rating | NaiveBayes | Supervised descritization | 50.53 |
| 5 yrs (2012-08)  3 bookies  ? match rating | 1 yr (2013)  3 bookies only  ? match rating | NaiveBayes | Kernel estimator | 51.84 |

**Training Data:**Prem12to08WithQuestionsWith3BookiesNoMatchRating\_Training.arff

**Test Data:** Prem13With3BookiesNoMatchRating\_Test.arff

Interesting. The presence of the match rating did increase the accuracy of the classifier *slightly*.

This match rating was applied to “recent form” – that is the last 6 games regardless if there was 6 games actually played. I was reading the documentation that I found it in and it implied that it should only be applied to 6 games i.e. 5,4,3,2,1 are just not valid.

TODO: Refactor the relevant method to take this into account:

“...provided at least 6 games have been played in the season to describe the recent form”.

This will mean that this statistic will only kick in after a certain time-period, that is, both teams for the fixture have played a minimum of 6 games. Otherwise it’s a question mark I’m afraid.

With “strict” Match Rating – by “strict” I mean 6 games minimum...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  3 bookies  Strict match rating | 1 yr (2013)  3 bookies only  Strict match rating | NaiveBayes | Supervised descritization | 51.05 |
| 5 yrs (2012-08)  3 bookies  Strict match rating | 1 yr (2013)  3 bookies only  Strict match rating | NaiveBayes | Kernel estimator | 52.37 |

**Training Data:**Prem12to08WithQuestionsWith3BookiesStrictMatchRating\_Training.arff

**Test Data:** Prem13With3BookiesWithStrictMatchRating\_Test.arff

**Conclusion:** Classifier went up a bit across the board.

I wanted to add the “fair-odds” calculation. So what I have now is the 3 bookies data, strict match rating and the fair-odds calc. The results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  3 bookies  Strict match rating  3 fair-odds | 1 yr (2013)  3 bookies only  Strict match rating  3 fair-odds | NaiveBayes | Supervised descritization | 52.11 |
| 5 yrs (2012-08)  3 bookies  Strict match rating  3 fair-odds | 1 yr (2013)  3 bookies only  Strict match rating  3 fair-odds | NaiveBayes | Kernel estimator | 50.26 |

**Training Data:**Prem12to08WithQuestionsWith3BookiesStrictMatchRatingPlusFairOdds\_Training.arff

**Test Data:** Prem13With3BookiesWithStrictMatchRatingPlusFairOdds\_Test.arff

**Conclusion:** Results are mixed. Slightly up with supervised descritization and slightly down with kernel estimator.

**TODO:** Run with “strict” match-rating and the 3 fair odds only.

**TODO:** Incorporate the actual league score that associated with the teams....

Running with the strict match-rating & 3 fair odds calculation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds | 1 yr (2013)  Strict match rating  3 fair-odds | NaiveBayes | Supervised descritization | 45 |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds | 1 yr (2013)Strict match rating  3 fair-odds | NaiveBayes | Kernel estimator | 47.63 |

**Training Data:**Prem12to08WithQuestionsNoBookiesStrictMatchRatingPlusFairOdds\_Training.arff

**Test Data:** Prem13NoBookiesWithStrictMatchRatingPlusFairOdds\_Test.arff

**Conclusion:** All results are down...

Now I want to add the league score to both teams and see if things get better...

So the training data is: no bookies & strict match rating & league scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score | 1 yr (2013)  Strict match rating  3 fair-odds  Current league score | NaiveBayes | Supervised descritization | 53.68 |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score | 1 yr (2013)  Strict match rating  3 fair-odds  Current league score | NaiveBayes | Kernel estimator | 56.58 |

**Training Data:**Prem12to08WithQuestionsNoBookiesStrictMatchRatingPlusFairOddsPlusLeagueScores\_Training.arff

**Test Data:** Prem13NoBookiesWithStrictMatchRatingPlusFairOddsPlusLeagueScores\_Test.arff

**Conclusion:** Results are nicely up...

So, now I want to investigate training data years 12 to 07 with test-data from 2012.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2011-07)  Strict match rating  3 fair-odds  Current league score | 1 yr (2012)  Strict match rating  3 fair-odds  Current league scores | NaiveBayes | Supervised descritization | 46.31 |
| 5 yrs (2011-07)  Strict match rating  3 fair-odds  Current league score | 1 yr (2013)  Strict match rating  3 fair-odds  Current league scoresscore | NaiveBayes | Kernel estimator | 48.16 |

**Training Data:**Prem11to07WithQuestionsNoBookiesStrictMatchRatingPlusFairOddsPlusLeagueScores\_Training.arff

**Test Data:** Prem12NoBookiesWithStrictMatchRatingPlusFairOddsPlusLeagueScores\_Test.arff

**Conclusion:** Worrying – big fall in accuracy. Worrying...

Now I want to add in the league score delta. So the training data is:

No bookies, 3 fair-odds, strict match rating plus league scores and also now league score delta...

League score delta is defined as the home team league score minus the away team league score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score | 1 yr (2013)  Strict match rating  3 fair-odds | NaiveBayes | Supervised descritization | 45 |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score | 1 yr (2013)  Strict match rating  3 fair-odds  Current league score | NaiveBayes | Kernel estimator | 47.11 |

**Training Data:**Prem12to08WithQuestionsNoBookiesStrictMatchRatingPlusFairOddsPlusLeagueScores\_Training.arff

**Test Data:** Prem12NoBookiesWithStrictMatchRatingPlusFairOddsPlusLeagueScores\_Test.arff

**Conclusion:** Results are well down for this season...

Lets keep going...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2010-06)  Strict match rating  3 fair-odds  Current league score | 1 yr (2011)  Strict match rating  3 fair-odds  Current league scores | NaiveBayes | Supervised descritization | 41.05 |
| 5 yrs (2010-06)  Strict match rating  3 fair-odds  Current league score | 1 yr (2011)  Strict match rating  3 fair-odds  Current league score | NaiveBayes | Kernel estimator | 44.21 |

**Training Data:**Prem10to06WithQuestionsNoBookiesStrictMatchRatingPlusFairOddsPlusLeagueScores\_Training.arff

**Test Data:** Prem11NoBookiesWithStrictMatchRatingPlusFairOddsPlusLeagueScores\_Tes t.arff

**Conclusion:** Again, another, big fall in accuracy...

Recently I have noted that the J48 decision tree is returning some excellent results for the most recent season but not necessarily for the others...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score | 1 yrs (2013)  Strict match rating  3 fair-odds  Current league score | J48 | -C 0.25 –M 2 | 73.442 |
|  |  |  |  |  |
| 5 yrs (2011-07)  Strict match rating  3 fair-odds  Current league score | 1 yrs (2012)  Strict match rating  3 fair-odds  Current league score | J48 | -C 0.25 –M 2 | 46.84 |
|  |  |  |  |  |
| 5 yrs (2010-06)  Strict match rating  3 fair-odds  Current league score | 1 yrs (2011)  Strict match rating  3 fair-odds  Current league score | J48 | -C 0.25 –M 2 | 44.74 |
|  |  |  |  |  |

**Ques:** Why is J48 so accurate with the most recent premiership data?

General idea: Apply the classifier to the forthcoming premiership season. Use it “in anger” when the first fixture experiences the full 5-attribute experience of:

* Strict match rating
* 3 fair-odds statistics
* Current league scores (for both)

Error detected. The very high 2013 test with 2012-08 training data was wrong. There was not enough data in the training file – which caused me to rebuild it with greater care.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score (for both) | 1 yrs (2013)  Strict match rating  3 fair-odds  Current league score (for both) | Naive Bayes | Kernel Estimator  SupervizedDescritization | 46.84  45.79 |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score (for both) | 1 yrs (2013)  Strict match rating  3 fair-odds  Current league score (for both) | Random Forest | Default | 47.63 |
| 5 yrs (2012-08)  Strict match rating  3 fair-odds  Current league score (for both) | 1 yrs (2013)  Strict match rating  3 fair-odds  Current league score (for both) | LMT | LMT –I -1 –M 15 –W 0.0  +Split on residuals  +useAIC  -FastRegression  minNoInstances=50  minNoInstances=2 | 51.32  40.26  51.05  51.32  51.32  51.32 |
|  |  |  |  |  |
| 5 years(2012-08)  3 bookies  Strict match rating  3 fair-odds | 1 yrs (2013)  3 bookies  Strict match rating  3 fair-odds | LMT  Naive-Bayes  Naive-Bayes  Logistic | LMT –I -1 –M 15 –W 0.0  + Kernel Estimator  +SupervizedDescritization  -R 1.0E-8 –M -1 | 51.57  50.26  52.11  52.63 |
|  |  |  |  |  |
| 5 years(2012-08)  3 bookies  Strict match rating  3 fair-odds  League delta | 1 yrs (2013)  3 bookies  Strict match rating  3 fair-odds  League delta | LMT  Naive-Bayes  Naive-Bayes  Logistic | LMT –I -1 –M 15 –W 0.0  Kernel Estimator  SupervizedDescritization  -R 1.0E-8 –M -1 | 50.52  50.26  52.37  51.57 |
|  |  |  |  |  |
| 5 yrs (2012-08)  Strict match rating  3 fair odds | 1 yrs (2013)  Strict match rating  3 fair odds | LMT  Naive-Bayes  Naive-Bayes  Logistic  Simple Logistic | LMT –I -1 –M 15 –W 0.0  Kernel Estimator  Supervized Descriization  -R 1.0E-8 –M -1  -I 0 –M 500 –H 50 –W 0.0 | 52.37  47.63  45  48.68  52.37 |
|  |  |  |  |  |

I want to take the data from the first part of the 2013 season (until match-rating kicks in) for the first time and append it to the training data...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 5 yrs (2012-08) + First part of 13 season\*  Strict match rating  3 fair-odds | 1 yrs (2013) – First part of season\*  Strict match rating  3 fair-odds | LMT  Naive Bayes  Naive-Bayes  Logistic  Simple Logistic | LMT –I -1 –M 15 –W 0.0  Kernel Estimator  SupervizedDescritization  -R 1.0E-8 –M -1  -I 0 –M 500 –H 50 –W 0.0 | 44.55  49.84  47.35  51.71  51.09 |

\* By first part of the season I am referring to the part until the full match rating kicks in...

**Conclusion:** Mixed. LMT shot down. Naive-Bayes/Logistic went up. Simple logistic down slightly.

I want to drop this now and go for 6 years on training data and see what happens...

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Train** | **Test** | **Model** | **Toggles** | **% Correctly Classified** |
|  |  |  |  |  |
| 6 yrs (2012-07) +  Strict match rating  3 fair-odds | 1 yrs (2013) –  Strict match rating  3 fair-odds | LMT  Naive Bayes  Naive-Bayes  Logistic  Simple Logistic | LMT –I -1 –M 15 –W 0.0  Kernel Estimator  SupervizedDescritization  -R 1.0E-8 –M -1  -I 0 –M 500 –H 50 –W 0.0 | 44.55  48.59  47.35  51.71  51.09 |

Read an excellent paper of soccer predictions...