Predicting soccer match outcome using machine learning algorithms

Authors: C.Liti*, V. Piccialli* and M. Sciandrone**

Speaker: Chiara Liti

*DICII University of Rome “Tor Vergata”

** DINFO University of Florence

MathSport International 2017

Agenda

- Problem description
- Purpose
- Statement of the problem
- Theoretical background
- Methodology
- Numerical Results
- Conclusions and Future Developments

Problem Description

We have considered the problem of predicting the outcome of a soccer match finished with a draw at the end of the first half using mainly the information stored during the first part of the match.

Purpose

- Use machine learning algorithms to predict the results of soccer matches finished with a draw at the end of the first half.

- Show the usefulness of the match-statistics used as features to train our models.

Statement of the problem

- Randomness of the data
- Existence of complex interacting factors
- Limited numbers of available match statistics
- Unbalanced sample of data

the prediction of soccer match outcome could be translated into a hard three-class classification problem.

Classification Problem

The classification consists of a learning step where given a training set

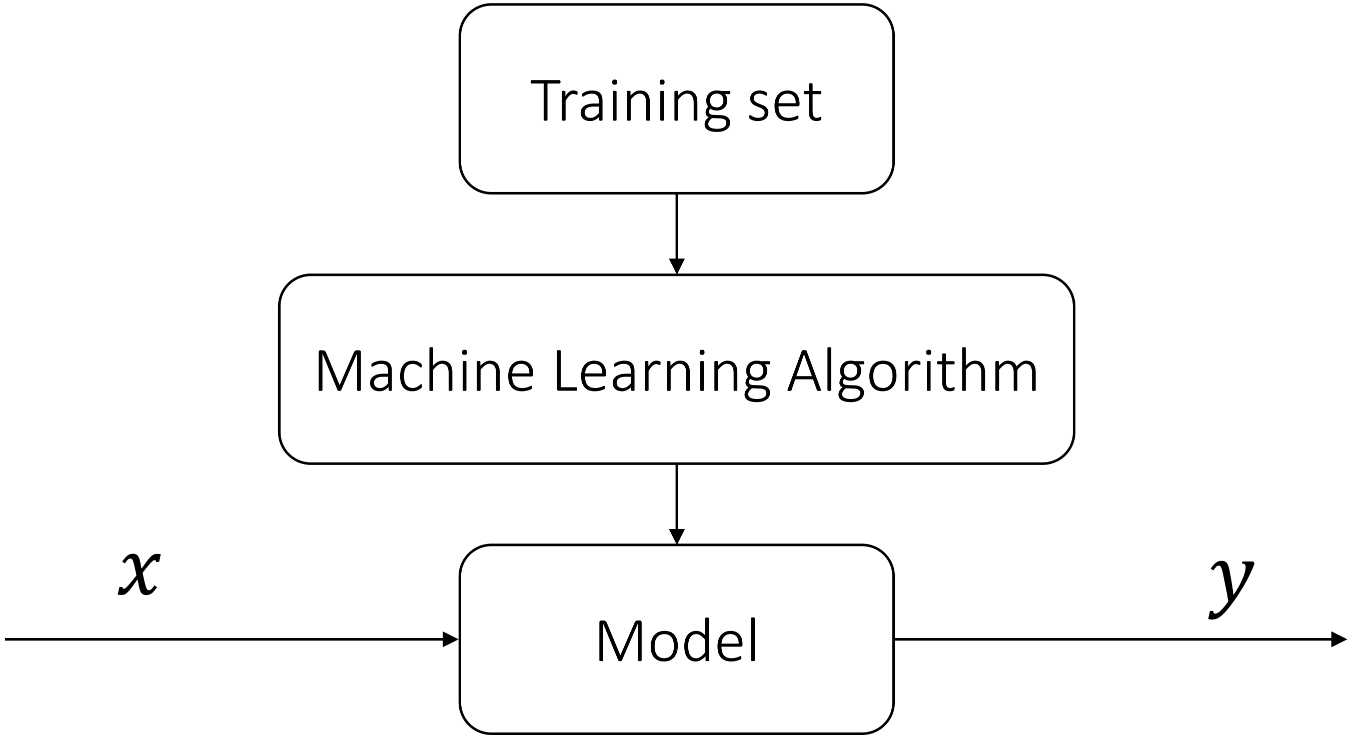
$$T = \{(x^p, y^p): x \ \in R^n, y^p \ \in S \subset N, p = 1, \dots, P\}$$

a classification model (or Target Function)

$$f: R^n \mapsto S \subset N$$

is trained in order to predict class labels for new data.

Classification Problem



The analysed sample of data

The analysed dataset contains the results of 166 matches finished with a draw at the end of the first half.

The initial set of feature is composed of 50 attributes representing the most relevant match-statistics (e.g., number of crosses, number of shots on target, etc.)

Feature Reduction

- Calculating the difference between the Home and Away descriptive match-statistics.
- Removing the attributes containing few values different from zero.

The final set of feature is composed of 27 attributes, the latest of which represents the class (i.e., Home win, Away win and Draw).

Data Preparation

Ten different pair of training and test were built in order to train and to evaluate the classification model.

Every training set contains the 70% of instances of each class, while each test set includes the remaining observations.

Model Training

We have used four different classifiers. The algorithms applied are the following:

- Naïve Bayes;
- C-SVM;
- ν-SVM;
- RBF Network.

Model Evaluation

Classifier	Accuracy	TPR Home win	TPR Away win	TPR Draw
Naïve Bayes	0.4340	0.6278	0.2167	0.3900
C-SVM	0.3920	0.5222	0.0333	0.4900
ν-SVM	0.3640	0.4167	0.2083	0.4100
RBF Network	0.4360	0.4833	0.2501	0.5200

Average Accuracy and True Positive Rate (TPR) over Test Sets employing the classifiers in their default version

Feature Selection

We have defined four different feature selection strategies based on the use of the filter ReliefF applied over each training set.

First Strategy: we have deleted those attributes having a negative score in all training sets.

Feature Selection

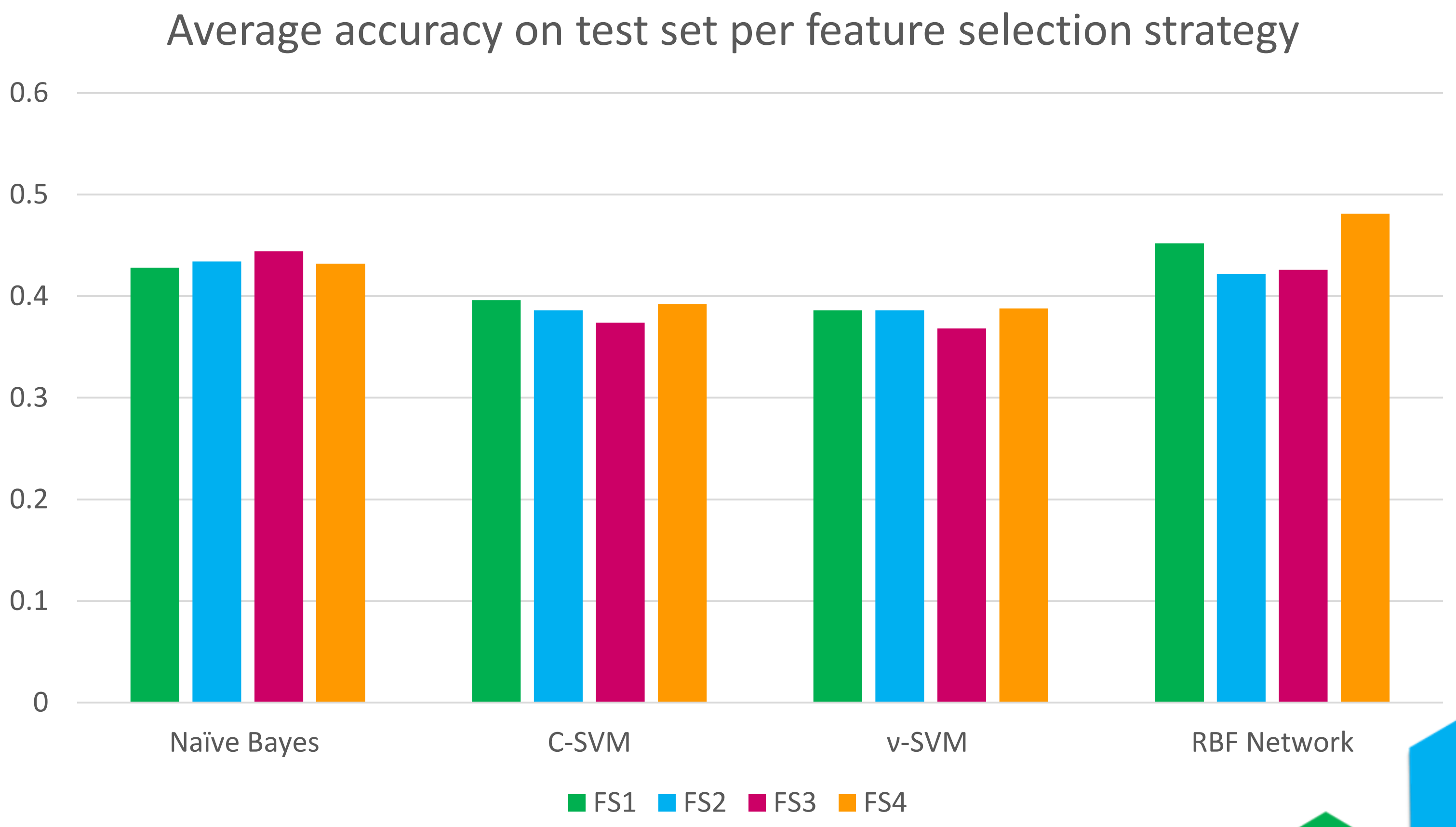
Second Strategy: we have deleted those attributes having a negative mean score.

Third Strategy: we have applied again the filter over the sets of feature identified using the first strategy and we have maintained those attributes having a positive score in at least six training sets.

Feature Selection

Fourth Strategy: we have applied again the filter over the sets of feature identified using the first strategy and we have removed those attributes having a negative score in at least eight training sets.

Model evaluation after the feature selection


Average accuracy on test set per feature selection strategy

The Baseline

Given ds the different between the score of Home and Away team before a match. Given $y = \{H, A, D\}$ the outcome of the match. Using the baseline we obtain the following classification

$$y = \begin{cases} H & if\ ds > 5, \\ A & if\ ds < -5, \\ D & otherwise. \end{cases}$$

Numerical Results

Classifier	Accuracy	TPR Home win	TPR Away win	TPR Draw
Baseline		0.3220	0.2000	0.5100
Naïve Bayes	0.4440	0.5945	0.3333	0.3750
C-SVM	0.3860	0.5000	0.0667	0.4750
ν-SVM	0.3880	0.4556	0.2250	0.4250
RBF Network	0.4520	0.5000	0.2582	0.5256

Comparison between the results obtained using our models and those predicted using the baseline

Conclusions

- Using a RBF Network combined with the first feature selection strategy we get results better than those predicted using the baseline.

- The obtained results to show the usefulness of the match-statistics collected during the first half to predict the outcome of a soccer match.

Future Developments

- Increase the dimension of the training set;
- Define an ad hoc feature selection strategy;
- Evaluate the machine learning algorithms over the new set of data.