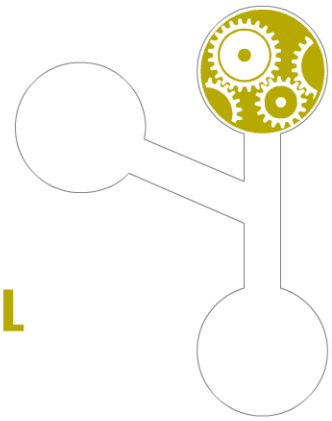
Brígida Mónica Faria · btf@estsp.ipp.pt[1,2]
Gladys Castillo · gladys@ua.pt[3]
Nuno Lau · nunolau@ua.pt[3]
Luís Paulo Reis · lpreis@fe.up.pt[4]



CLASSIFICATION OF FC PORTUGAL ROBOTIC SOCCER FORMATIONS: A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS

ABSTRACT
Mobile robots have the potential to become the ideal tool to teach a broad range of engineering disciplines. Indeed, mobile robots are getting increasingly complex and accessible. They embed elements from diverse fields such as mechanics, digital electronics, automatic control, signal processing, embedded programming, and energy management. Moreover, they are attractive for students which increases their motivation to learn. However, the requirements of an effective education tool bring new constraints to robotics. This article presents the e-puck robot design, which specifically targets engineering education at university level. Thanks to its particular design, the e-puck can be used in a large spectrum of teaching activities, not strictly related to robotics. Through a systematic evaluation by the students, we show that the epuck fits this purpose and is appreciated by 90 percent of a large sample of students.

KEYWORDS
Knowledge Discovery, Data Mining; Support Vector Machines; RoboCup Soccer; Simulation; Formations.

I. INTRODUCTION
RoboCup Simulation League has been one of the first competitions integrated on the RoboCup international project. The main goals of this league are concerned with developing the high-level decision and coordination modules of teams of robots [1]. The 2D simulation league has evolved over the years, but the principal architecture of the simulator is the same at it was firstly used in 1997 [1]. The Soccerserver is the simulator that creates a 2D virtual soccer field and the virtual players that are modeled as circles. This simulator implements the movement, stamina, kicking and refereeing models of the virtual world [2]. Another aspect that brings realism is the fact that the models in the simulator are taken both from real robots and from human like characteristics. The team of FC Portugal [2,3] has demonstrated very good results since its creation in 2000 and has won several European and World competitions [1]. The research focused development of the team is one of its main assets and still continues as every year new challenges are introduced. One concept that has been studied it is the usability of formations [4]. One important aspect is to be able to classify and predict the formations that are being used on games. Another important aspect is to identify the opponent team that FC Portugal is playing with and its characteristics in the first moments of the game. In this paper a comparative study of three techniques for classification is presented. The following techniques have been used: Support Vector Machines (SVM) [5]; Artificial Neural Network (ANN) and k-Nearest-Neighbor.The environment tool used for machine learning and data mining experiments was RapidMiner [6].
This paper is organized with an initial explanation of the RoboCup Competition with special relevance for the simulation leagues. Next, an explanation and description of the three algorithms is presented. After that the kind of measures used to compare the classifiers and the statistical hypothesis to compare the average performance of the classifiers are presented. Finally the experimental results are presented along with some conclusions and future work.

II. ROBOCUP SOCCER
RoboCup is an international cooperative project to promote Artificial Intelligence, Robotics and related fields. It is an attempt to promote artificial intelligence and robotics research by providing a standard problem where a wide range of technologies can be integrated and examined. The known goal of the RoboCup project is to create a soccer team with humanoid robots that can play and win to the world champion soccer team, by the year of 2050 [7]. In this project there are different leagues divided in two main groups: robotics and simulation. The first group involves physical robots with different sizes and different rules based on the competition that they integrate. The second one has the goal of, without the necessity to maintain any robot hardware, being able to research on artificial intelligence, coordination methodologies and team strategy. There is plenty of work performed and

[1] **DETI/UA**: Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro, IEETA - Instituto de Engenharia Electrónica e Telemática de Aveiro, Aveiro, Portugal.
[2] **ESTSP/IPP**: Escola Superior de Tecnologia de Saúde do Porto, Instituto Politécnico do Porto, Porto, Portugal.
[3] **DMAT/UA**: Departamento de Matemática, Universidade de Aveiro, Aveiro, Portugal.
[4] **DEI/FEUP**: Departamento de Engenharia Informática, Faculdade de Engenharia da Universidade do Porto, LIACC: Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, Porto Portugal.

The neurons are typically identical units that are connected by links. The interconnections are used to send signals from one neuron to the other [13]. The concept of weights between nodes is also present since it is used for establishes the importance from one connection to the other. The network may contain several intermediary layers between its input and outputs layers. The intermediary layers called hidden layers and the nodes embedded in these layers are called hidden nodes. In a feed-forward neural network the nodes in one layer are connected only to the nodes in the next layer. The Perceptron is the simplest model since do not use any hidden layers. One of the most used models for classification using ANNs is the Multilayer Perceptron using the backpropagation algorithm in which ANNs have 3 or 4 layers. The latter was the model used in this study.The ANN model has several characteristics like the capability of handling redundant features since the weights are automatically learned during the learning phase. The weights for redundant features tend to be very small. The method called gradient descent [9] is used for learning the weights which often converge to some local minimum; however one way to overpass the local minimum is to add a momentum term [9] to the weight update formula. Another known characteristic is the consuming time for training an ANN, especially when the number of hidden nodes is large.

C. K-Nearest Neighbor

A Nearest Neighbor classifier represents each example as a data in a d-dimensional space, where d is the number of attributes. Given a test example it is computed the proximity to the rest data points in the training set, using a measure of similarity or dissimilarity, such as Euclidian measure or its generalization, the Minkowski distance metric, the Jaccard Coefficient or Cosine Similarity [9]. The k-nearest neighbor (k-NN) of a given example refers to the k points that are closest to the example. Some of the main points that characterized k-NN are the insertion in the category of lazy learners, since they do not require building a model and only make their predictions on local information. However classifying a test example is an expensive task because it is necessary to compute individually the proximity values between the test and training examples. An important decision about the proximity measure it is also necessary since the wrong choice can produce wrong predictions [9].

D. RapidMiner Environment

The RapidMiner is a software for all stages in Knowledge Discovery in Databases. It runs on every platform and operating system with the language Java, the KDD projects are modeled as trees operator which is extremely intuitive and can be saved as building blocks for later re-use. The internal XML representation ensures standardized interchange format of data mining experiments. Other interesting characteristics of RapidMiner are: simple scripting language allowing for automatic large-scale experiments; multi-layered data view concept ensuring efficient and transparent data handling. An additional property is that the machine learning library WEKA is fully integrated in RapidMiner [6].

The flexibility in using RapidMiner is another characteristic, since it has graphical user interface (GUI) for interactive prototyping; a command line mode (batch mode) for automated large-scale applications and Java application programming interface (API) to produce more programs. The initial version known as YALE (Yet Another Learning Environment) has been developed by the Artificial Intelligence Unit of University of Dortmund [6]. Today the core of RapidMiner is Open-Source and an edition for the Community is free of charge, however the Enterprise Edition needs a proprietary license. The recent version 4.5 brings more facilities like a new operator called "Script" [6] for professional analysis process design where built-in operators are not sufficient to achieve a desired task. The RapidMiner project is also characterized for giving quick responses to developer questions posted in its forum (rapid-forum [6]), since it is maintained by several full members. This reveals the activity and growing of this software allied to the attention given by the users and researchers on Data Mining.

IV. EXPERIMENTAL DEVELOPMENT

The comparative study of the three above mentioned algorithms involves the dataset produced by the positions of the players of the FC Portugal in 2D Simulation league. The performance measures are briefly described in this section together with the experimental settings and results.

A. Data set description

The dataset was produced with the x, y positions of eleven players of FC Portugal in 2D Simulation League in six distinct games with dynamic positioning and role exchange for the players. FC Portugal played two games against some known robotic soccer teams: Hellios, Brainstormers and NCL [7]. The attributes used for this study are the ball and players' positions and the class is the formation that the team was playing with. The classification became a multi-class problem since the FC Portugal could play with ten different formations.

Table I displays the possible formations that the team could play and Fig. 3 presents an example of a formation (325).

Table I · Formations of FC Portugal - Multi-class problem

Classes	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten
Formation	433	442	343	352	541	532	361	451	334	325

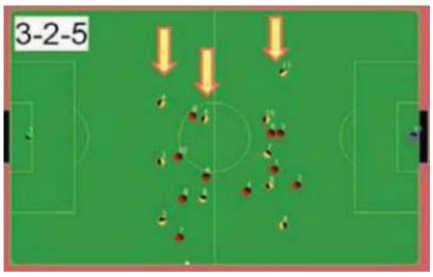


Figure 3 · FC Portugal team playing in 325

The coordinate x has the range of -52,5 and 52,5 and the coordinate y varies between -34,0 and 34,0 (corresponding to a typical real soccer field of 105x68m), where the center of the field is the origin of the referential [3]. The games were executed in Linux and the logs files are converted in text files with a simple application getWState [4] written in C++ for this purpose. The information that can be extracted from the games are the position and velocity of the ball and the eleven players of the two teams and other particular characteristics like stamina, kicks, head and body angles. In a previous work [4] it was discovered that the database with the center of mass of the FC Portugal team produces better results. Since the primordial objective of this work is to compare three different classifiers and obtain the best model, the variables corresponding to the center of mass were included on the databases. Therefore the final data set had the positions of the players, the position of the ball, the center of mass and the formation that FC Portugal was playing. Thus, the data base has 26 numerical and continuous attributes (R^{26}) and one nominal attribute (10 formations options of FC Portugal). The first dataset (Database A) has 37943 examples with approximately 6000 cycles by game. There are differences on the number of examples since there is the possibility to have periods in the game that are stopped or others that are not counted but in which players are still moving and thus are included in the database.