UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Clasificador de Aprendizaje Automático para determinar Acciones Clave en el Desempeño Futbolístico de Jugadores

PROPUESTA DE TEMA DE MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

CRISTIAN LILLO CIERO

MODALIDAD:
Doble Titulación con Magíster

PROFESOR GUÍA:
Javier Bustos

SANTIAGO DE CHILE
2024

1. Introducción

En la actualidad, el fútbol es un deporte que ha evolucionado en términos de tecnología y análisis de datos. Los clubes de fútbol han comenzado a utilizar herramientas tecnológicas para mejorar el rendimiento de sus jugadores y equipos. Sin embargo, en las etapas formativas, no se dispone de un modelo que permita evaluar el rendimiento de los futbolistas de manera objetiva. Esto conlleva a que los entrenadores deban evaluar el rendimiento de los jugadores en base a su experiencia y conocimiento del deporte, lo que puede llevar a sesgos y subjetividad.

En este contexto, el presente proyecto busca desarrollar un clasificador de aprendizaje automático que permita determinar las acciones clave en el desempeño futbolístico de los jugadores. Este clasificador permitirá a los entrenadores evaluar el rendimiento de manera objetiva y precisa. Para ello, se utilizarán técnicas de Machine Learning o Deep Learning para analizar los datos de los jugadores y determinar las acciones clave que influyen en su desempeño.

El desarrollo de este clasificador ayudará a identificar áreas de mejora y a tomar decisiones informadas en base a los datos. Además, se podrá utilizar en las etapas formativas de los clubes de fútbol para identificar y potenciar el talento desde edades tempranas, contribuyendo al desarrollo y formación de futbolistas.

2. Situación Actual

Actualmente, existen modelos como PlayeRank [1] que categorizan las acciones de los jugadores en el campo de juego, asignando un valor a cada acción. Sin embargo, estos modelos presentan limitaciones en la asignación de valores, ya que la gran mayoría de las acciones son calificadas con valores similares, lo que dificulta la identificación de las acciones clave en el desempeño de los jugadores. Este problema es algo que, según Guo et. al. [2], ocurre en modelos del tipo Support Vector Machine (SVM) al tener una alta cantidad de dimensensiones, pues se produce una reducción en la precisión y efectividad.

También han sido realizados otros modelos que buscan una manera más simple de determinar si los jugadores realizaron acciones clave. Uno de ellos es propuesto por Duch et. al. [3], donde se define una métrica relacionada a la fracción de veces que el jugador realiza un pase en una jugada que termina en gol, lo que tiene sentido para delanteros y mediocampistas, pero no permite analizar correctamente las intervenciones de defensas y porteros en el juego. Existe también el trabajo de Brooks et. al. [4], el cual tiene un enfoque en la cantidad de pases completados, y determina que su sistema de ranking termina favoreciendo a los jugadores más ofensivos, lo que no necesariamente se traduce en un mejor rendimiento del equipo.

3. Objetivos

Objetivo General

Se propone desarrollar un clasificador de aprendizaje automático que permita determinar las acciones clave en el desempeño futbolístico de los jugadores. Se utilizarán datos provenientes de StatsBomb [5], que han registrado eventos ocurridos en miles de partidos de fútbol, para entrenar el clasificador y se evaluará comparándolo con el modelo de PlayeRank [1], que también categoriza las acciones de los jugadores en el campo de juego. El objetivo es lograr un clasificador que asigne correctamente la importancia de las acciones y que no sean calificadas la gran mayoría con valores similares, como ocurre en el modelo de PlayeRank.

Objetivos Específicos

1. Obtener acciones realizadas por jugadores profesionales de fútbol. Estas deben estar en orden cronológico según la fecha del partido y el minuto de juego.
2. Entrenar distintos tipos de modelos, tanto de Machine Learning como de Deep Learning, para encontrar el que mejor se ajuste al problema y obtenga los mejores resultados.
3. Evaluar el clasificador con un conjunto de datos de prueba y comparar resultados con PlayeRank.

Evaluación

Para evaluar el trabajo, se utilizará un conjunto de datos de prueba que no haya sido utilizado en el entrenamiento del clasificador. Se compararán los pesos asignados por el clasificador para las acciones de los jugadores con los pesos asignados por PlayeRank, donde se espera que el clasificador sea capaz de asignar valores más precisos y diferenciados entre las acciones.

4. Solución Propuesta

Se planea desarrollar el modelo en el lenguaje de Python, haciendo uso de librerías como scikit-learn, TensorFlow y Keras para el entrenamiento de los modelos de Machine Learning y Deep Learning. Como se mencionó previamente, una parte importante del trabajo consiste en determinar cuál tipo de modelo es el más óptimo para la resolución de la problemática. Para ello, se analizarán distintos modelos, como Random Forest (RF), Support Vector Machine (SVM), Redes Neuronales Convolucionales (CNN), entre otros, y se compararán los resultados obtenidos.

En relación a los datos, se dispone de una gran cantidad de archivos en formato JSON. Estos archivos están disponibles en el repositorio de GitHub de StatsBomb [5] y contienen información detallada de más de 3000 partidos. Los archivos se dividen en cuatro categorías: eventos, alineaciones, partidos y competiciones por temporada. Aunque el enfoque principal se centra en los eventos, se utilizarán los datos de las otras categorías para complementar la información y mejorar la calidad del clasificador.

Mediante el uso de librerías como Pandas y Matplotlib, se realizará un análisis exploratorio de los datos para identificar patrones y características relevantes. Se buscará determinar objetivamente cuáles son las acciones que modifican el resultado de un partido de fútbol y que, por lo tanto, deben ser consideradas como acciones clave por parte del modelo. Este análisis permitirá seleccionar las variables más relevantes para el entrenamiento del clasificador y descartar aquellas que no aporten información significativa.

Se debe determinar además la mejor manera de representar los datos para el entrenamiento del modelo, pues debe ser capaz de identificar patrones y relaciones entre las acciones realizadas por los jugadores. Se considera que la representación de los datos es un aspecto fundamental para el éxito del clasificador, por lo que se dedicará especial atención a este aspecto.

5. Plan de Trabajo (Preliminar)

1. Definir la mejor alternativa para almacenar la masiva cantidad de datos de StatsBomb.
2. Realizar un procesamiento a los datos para obtener las acciones realizadas por cada jugador en todos sus partidos disputados.
3. Realizar un análisis exploratorio de los datos para identificar patrones y características relevantes.
4. Seleccionar las variables más relevantes para el entrenamiento del clasificador.
5. Definir la mejor manera de representar los datos para el entrenamiento del modelo.
6. Entrenar distintos modelos de Machine Learning y Deep Learning para determinar cuál es el más óptimo para el problema.
7. Evaluar el clasificador con un conjunto de datos de prueba y verificar si resultados son más precisos que en PlayeRank.
8. Escribir el informe final de la tesis.

Referencias

[1] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, y Fosca Giannotti, «PlayeRank: Data-Driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach», *ACM Transactions on Intelligent Systems and Technology*, vol. 10, n.º 5, pp. 1-27, sep. 2019, [En línea]. Disponible en: https://doi.org/10.1145/3343172

[2] Baofeng Guo, R.I. Damper, Steve R. Gunn, y J.D.B. Nelson, «A fast separabilty-based feature selection method for high-dimensional remotly sensed image classification», *Pattern Recognition*, vol. 41, n.º 5, pp. 1653-1662, may 2008, [En línea]. Disponible en: https://doi.org/10.1016/j.patcog.2007.11.007

[3] Jordi Duch, Joshua S. Waitzman, y Luís A. Nunes Amaral, «Quantifying the Performance of Individual Players in a Team Activity», *PLoS ONE*, vol. 5, n.º 6, jun. 2010, [En línea]. Disponible en: https://doi.org/10.1371/journal.pone.0010937

[4] Joel Brooks, Matthew Kerr, y John Guttag, «Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights», *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 49-55, ago. 2016, [En línea]. Disponible en: https://doi.org/10.1145/2939672.2939695

[5] «StatsBomb». [En línea]. Disponible en: https://github.com/statsbomb/open-data