Clustering

Desarrollo de Sistemas Inteligentes

Alberto Velasco Mata
Alberto.Velasco1@alu.uclm.es
Escuela Superior de Informática
Ciudad Real

Antonio Pulido Hernández
Antonio.Pulido@alu.uclm.es
Escuela Superior de Informática
Ciudad Real

RESUMEN

In this document an analysis of the data related with the 2020 European Women's Handball Championship will be presented. The statistics about each player involved in the championship were used to perform a cluster analysis in order to understand which were the different types of players and teams that participated in the tournament.[1]

1. INTRODUCCIÓN

En este trabajo se pretende realizar un estudio exploratorio de datos usando análisis de conglomerados *(análisis cluster)*. Los datos que se van a analizar son los correspondientes a las estadísticas del Europeo Femenino de Balonmano del año 2020 en los que se pueden encontrar todas las estadisticas relevantes de cada jugadora en cada partido.

Para analizar esta información se seleccionarán las variables más relevantes y se definirán los diferentes grupos de jugadores y equipos por medio del algoritmo de clustering Expectation-Maximization [1]. Este algormitmo es un método iterativo que utiliza funciones Gaussianas para definir los diferentes grupos. Además, este algoritmo puede ser inicializado de manera aleatoria o mediante el algoritmo *K-means*.

2. HITO 1: AGRUPACIONES POR JUGADORAS

En este apartado se analizan los datos de cada jugadora en cada uno de los partidos en los que participa en el Europeo 2020 para determinar los diferentes tipos de jugadoras.

2.1. Análisis exploratorio

En primer lugar, se ha realizado un análisis exploratorio de los datos. Siguiendo el consejo del experto, se ha hecho un primer análisis cluster con los datos proporcionados por las variables **Goals** y **Shots**. Estas dos variables son agregadas, por lo que no serán utilizadas en el análisis cluster final, pero son ideales para explorar los datos teniendo en cuenta que pueden ser representadas en un gráfico.

Los datos de estas variables han sido agrupados (aplicando la suma) según los nombres de las jugadoras. De esta manera, obtenemos los goles y lanzamientos totales para cada jugadora en el campeonato. Posteriormente, se ha aplicado el algoritmo Expectation-Maximization (EM). Los parámetros se han seleccionado usando el BIC, obteniendo que lo más óptimo es agrupar en 7 componentes y con una covarianza esférica. El resultado del *clustering* puede verse en la Figura 1.

[1]El código se encuentra disponible en: https://github.com/AlbertoVelascoMata/course.dsi

Los datos parecen tener cierta correlación, teniendo en cuenta que los puntos se distribuyen prácticamente en una línea recta. Esto tiene sentido, ya que cuanto mayor es el número de lanzamientos, mayor es el número de goles. Las agrupaciones resultantes no proporcionan mucha información más allá de que hay jugadoras que lanzan más (y por tanto meten más goles) y viceversa, lo cuál **podría estar relacionado con su posición**.

Tras esta exploración inicial, se ha procedido a estudiar las demás variables para seleccionar aquellas de mayor relevancia de cara a la formación de diferentes grupos. En primer lugar, se han desechado las variables de carácter identificativo, las variables compuestas mediante la agregación de otros datos (como las usadas en la exploración anterior) y las variables que se consideran de poca relevancia según el experto (ej: tarjetas amarillas).

Además, se han extraído los diferentes valores que toman cada una de las variables restantes. Esto ha permitido detectar una columna adicional (**2+2**) que tomaba el mismo valor para todas las jugadoras, y por tanto ha sido descartada de la tarjeta de datos.

Se ha realizado también un estudio de la posible correlación entre las variables con el objetivo de descartar aquellas que estén muy relacionadas y puedan dar lugar a datos redundantes. En general, consideramos que no hay una correlación muy alta entre las variables seleccionadas por lo que no descartamos ninguna. Sí que podría entenderse que hay cierta relación entre el número de lanzamientos fallidos y el de goles metidos desde más de 9 metros (de manera similar al análisis exploratorio anterior), pero se ha decidido mantener ambas variables ya que podrían influir en las agrupaciones.

Este proceso de análisis ha dado lugar a 22 variables que se considerarán para el estudio.

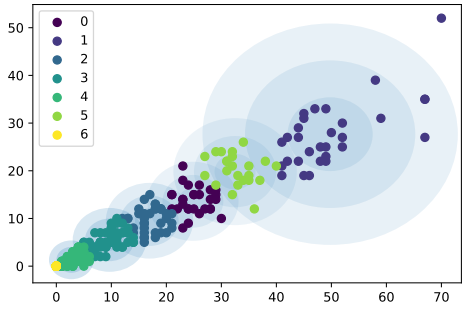

Figura 1: Clusters EM de jugadoras con las variables Goals y Shots

2.2. Determinación de los valores atípicos

Para que el cálculo de los diferentes grupos no sea engañoso y se ajuste lo máximo posible a la realidad se ha determinado que se eliminarán los posibles valores atípicos del conjunto de datos. Para ello, se ha utilizado el método **Jackknife**. Se ha implementado en Python a partir del código visto en clase, calculando la suma de los cuadrados de las distancias entre cada punto y su clúster. Una vez ejecutado, se han detectado cuatro valores atípicos que corresponderían a las jugadoras que se encuentran en los indices 157, 162, 166 y 176 en el conjunto de datos inicial como se puede observar en la Figura 2. Estos índices se corresponden con las jugadoras *Camila Micijevic, Nora Mork, Cristina Neagu y Stine Oftedal.*

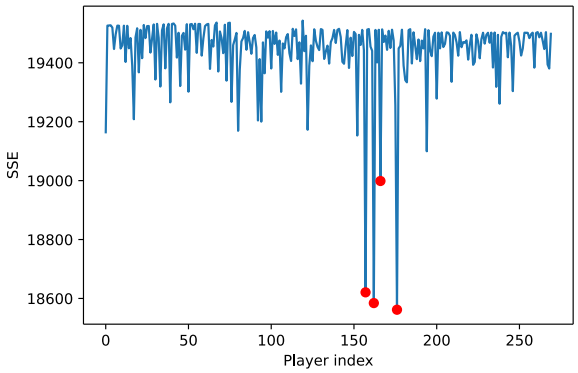

Figura 2: Detección de valores atípicos en el análisis de jugadoras.

Consultando los valores de las variables correspondientes a estas jugadoras se puede entender que sean considerados como valores atípicos. Por ejemplo, la jugadora *Nora Mork* marcó 29 goles desde los 7 metros en todo el torneo, un valor muy por encima de la segunda que más goles desde los 7 metros marcó con 19 goles. Por lo tanto, se han eliminado estos valores atípicos del conjunto de datos antes de realizar el análisis de conglomerados para evitar conclusiones erróneas.

2.3. Realización del análisis de conglomerados

Para reducir la dimensionalidad del problema se ha realizado un análisis de componentes principales previo al clustering. Esto nos permitirá no solo representar gráficamente los resultados sino también determinar qué variables son más relevantes en cuanto a la varianza de los datos. Para ello, se han normalizado los datos utilizando los escaladores proporcionados por la librería *scikit-learn* StandardScaler y MinMaxScaler, obteniendo los resultados que se pueden observar en la Tabla 1 en función del número de componentes principales y el proceso de normalización de datos seleccionado. Para poder dibujar de manera sencilla se han elegido dos componentes principales y el escalador MinMax, ya que proporciona un mayor porcentaje de varianza explicada. Sin embargo, un 45 % es relativamente bajo, por lo que habrá que tenerlo en cuenta al interpretar los resultados.

Scaler	# Components	Variance	Variance/Component
Standard	2	~40 %	[26.58 %, 13.09 %]
Standard	3	~50 %	[26.58 %, 13.09 %, 10.33 %]
MinMax	2	~45 %	[31.04 %, 13.87 %]
MinMax	3	~55 %	[31.04 %, 13.87 %, 10.33 %]

Cuadro 1: Varianza explicada con diferentes configuraciones del PCA

Habiendo analizado las variables, eliminado los valores atípicos y aplicado un análisis de componentes principales, se ha procedido a realizar el análisis de conglomerados utilizando el algoritmo Expectation-Maximization.

Primero, se ha calculado el mejor valor de k (número de grupos) y el mejor modelo de covarianza para este conjunto de datos utilizando el BIC. Se ha obtenido que el valor óptimo para el número de grupos es 5, con una covarianza "full"(es decir, las gausianas pueden adoptar independientemente cualquier posición y forma). La Figura 3 muestra los diferentes valores del BIC para cada tipo de covarianza y número de clústers.

Una vez calculado el número de grupos y el mejor tipo de covarianza, se puede proceder a calcular los grupos usando el algoritmo Expectation-Maximization. El resultado obtenido puede observarse en la Figura 4. Las conclusiones sobre el mismo se detallan más adelante.

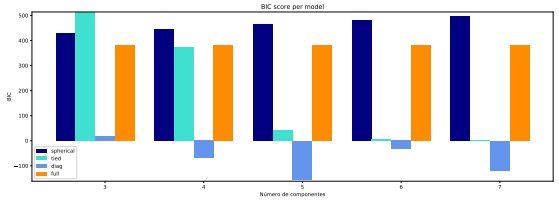

Figura 3: BIC según distintas configuraciones de K y tipo de covarianza

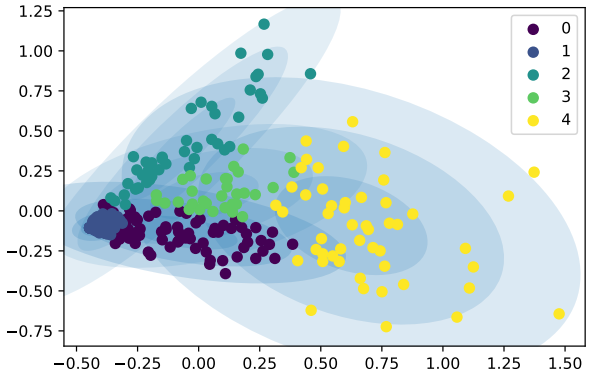

Figura 4: Clusters obtenidos al aplicar EM con las jugadoras

2.4. Estudio de la parametrización óptima del algoritmo

El algoritmo Expectation-Maximization se puede inicializar usando valores aleatorios o tomando como punto de partida el resultado de aplicar K-means. Se han probado ambas inicializaciones, pero los resultados obtenidos y mostrados se han realizado con inicialización aleatoria. Esto se debe a que este algoritmo tiende a quedarse estancado en mínimos locales, siendo difícil encontrar las mejores agrupaciones. Utilizando la inicialización aleatoria y ejecutándolo varias veces hemos obtenido resultados más diferenciados, pudiendo elegir aquel que consideramos que se ajusta mejor a los datos.

Tenemos que tener en cuenta que el estudio del mejor número de grupos y mejor tipo de covarianza utilizando el BIC también se ha llevado a cabo con inicialización aleatoria. Esto ha dado lugar a que, según la ejecución, el número óptimo de clusters variase entre 5 y 8, y el mejor tipo de covarianza fuese "full" o "diagonal" (ejes orientados a lo largo de los ejes de coordenadas) según la ejecución. Hemos decidido utilizar 5 como número de grupos y "fullçomo tipo de covarianza debido a los siguientes motivos:

- Es el resultado que aparecía en la mayoría de las ejecuciones.
- Según la opinión del experto, agrupar en 5 conglomerados parece lo más óptimo debido a las posiciones de las jugadoras.
- Cuando el resultado se calcula con más de 5 clústers, estos son subdivisiones de los que aparecen utilizando 5 grupos.

2.5. Caracterizacion de cada grupo

Se han obtenido las medias de cada variable para cada cluster resultante con el objetivo de caracterizar cada grupo. Observando los grupos obtenidos se puede detectar como curiosidad el *grupo 0* que es un pequeño grupo muy amontonado en el que se pueden encontrar las porteras de los equipos y las peores jugadoras del torneo. Esta situación se debe a que las porteras de los equipos no tienen goles marcados ni tiros, por lo que prácticamente todas sus estadísticas son 0. Sin embargo, hay que tener en cuenta que también se pueden encontrar jugadoras que no son porteras en este grupo.

También se puede observar el *grupo 2* que está en su mayoría formado por las extremos de los equipos y que sus datos de tiros y goles desde el extremo son altos.

El *grupo 4* está formado en su mayoría por laterales (y algunas centrales) que tienen una alta participación en el juego con datos bastante buenos de asistencias.

3. HITO 2: ANÁLISIS DE EQUIPOS

En este apartado se analizan los datos de cada equipo que ha participado en el Europeo de Balomano 2020. Debido a que los datos de los que parte este caso de estudio son los de las jugadoras en cada partido se han agrupado estos datos por equipo para poder analizar las estadísticas de cada uno.

3.1. Análisis exploratorio

Aplicando la misma estrategia que en el Hito 1, primero se ha realizado un análisis exploratorio de los datos utilizando las variables agregadas **Goals** y **Shots** que han permitido tener una visión inicial del conjunto de datos.

Cluster	Equipos
0	CRO, DEN, FRA, NED, RUS
1	CZE, POL, SLO, SRB
2	ESP, GER, HUN, MNE, ROU, SWE
3	NOR

Cuadro 2: Equipos agrupados usando las variables Shots y Goals

Aunque también se aprecia cierta correlación entre el número de lanzamientos y el de goles, sí que parecen formar grupos más o menos distinguidos. Aplicando de nuevo el algoritmo EM se han obtenido las agrupaciones que aparecen en la Tabla 2, y que se aprecian gráficamente en la Figura 5. Como se concluyó en el Hito 1, los equipos que más lanzan también son los que meten más goles. Además, podemos apreciar cómo Noruega, que quedó en el primer puesto, aparece en un grupo independiente. Por otra parte, Francia, Croacia y Dinamarca, todas ellas finalistas, quedan en otro grupo separado junto con Países Bajos y Rusia. Consideramos por tanto que esta agrupación, a pesar de ser una aproximación inicial, es bastante adecuada para los datos.

Es importante destacar que la parametrización del algoritmo EM ha tenido que ser realizada de forma manual, estableciendo el número de clústers en 4. La exploración de la parametrización utilizando el BIC indicaba que la K debería estar ser 8 o 9. Esto podría deberse a que el algoritmo que estamos utilizando tiende mucho al sobreaprendizaje, costándole mucho encontrar las mejores distribuciones. Estos factores, unidos a que hay muy pocos datos (16 equipos), dan lugar a que la parametrización supuestamente óptima tenga una granularidad muy alta.

Al igual que en el Hito 1, se ha llevado a cabo una exploración de las demás variables de manera similar. Por simplicidad, consideramos que no es relevante incluirla en este documento. Lo único destacable es que en este caso el análisis de componentes principales daba lugar a una mayor explicación de la varianza utilizando el escalador estándar, al contrario que en el Hito 1.

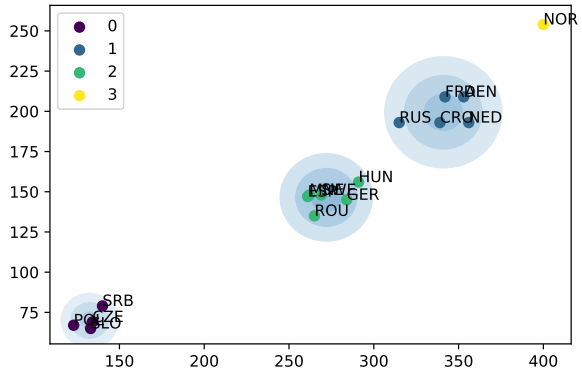

Figura 5: Agrupaciones de equipos con las variables Shots y Goals

Cluster	Equipos
0	CZE, POL, SLO, SRB
1	ESP, GER, NED, RUS, SWE
2	NOR
3	CRO, DEN, FRA
4	HUN, MNE, ROU

Cuadro 3: Equipos agrupados usando una parametrización manual (K=5, 'full)

3.2. Determinación de valores atípicos

De nuevo, de manera similar al Hito 1, se ha utilizado la técnica Jackknife para detectar valores atípicos. Sin embargo, y como era de esperar, no se han detectado outliers debido a que el número de equipos participantes en el Europeo de Balonmano Femenino 2020 es solamente de 16 equipos. Por tanto, no se ha eliminado ningún valor atípico.

3.3. Realización del análisis cluster

Se ha utilizado nuevamente el BIC como criterio de selección del número de clústers y del tipo de varianza a usar en el algoritmo EM. No obstante, ha ocurrido algo similar a la exploración inicial y el resultado supuestamente óptimo de la parametrización es demasiado granular, como puede verse en la Figura 6. Únicamente se agrupan aquellos puntos (equipos) que están muy cerca entre sí.

Esta vez se ha optado por establecer manualmente el número de clusters en 5, y utilizando una inicialización basada en K-means para evitar que los resultados fluctuasen tanto debido a la escasa cantidad de puntos. El resultado puede verse en la Figura 7 y en la Tabla 3.

3.4. Caracterización de cada grupo

Se han obtenido las medias de cada una de las variables para cada cluster resultante con el objetivo de caracterizar cada grupo. Es curioso el caso de Noruega, en el grupo 2, ya que tiene muchas asistencias y los goles están más o menos repartidos entre las distintas posiciones. Destaca también el hecho de que no hay una relación tan directa entre lanzamientos fallidos y goles, sino que los lanzamientos fallidos se mantienen a un nivel inferior.

En el caso del grupo 3, con las finalistas, hay también un elevado número de asistencias, pero en este caso los goles se realizan más desde los extremos.

El grupo 2 destaca por tener, de nuevo, muchos goles desde los extremos, pero en este caso muy pocas asistencias. Podría decirse que en estos equipos las jugadoras que más marcan son las que están en los extremos.

4. CONCLUSIONES

En general, podríamos decir que los resultados obtenidos con el clustering de jugadoras se basa mucho en la posición en la que juegan. En cuanto a los resultados del clustering de equipos, podría resultar destacable el hecho de que las agrupaciones coincidan con los resultados del campeonato en algunos casos. Esto podría indicar que se está modelando de cierta manera el estilo de juego de esos equipos.

Por último, respecto al algoritmo usado, consideramos que tiene la ventaja de proporcionar resultados probabilísticos. Esto ha sido de mayor utilidad en el caso de las jugadoras, ya que al modelar esas agrupaciones hay bastante solapamiento. Por otro lado, hemos notado mucho los inconvenientes del algoritmo EM, ya que los resultados variaban mucho según la ejecución y ha sido difícil encontrar agrupaciones que no estuviesen tan ajustadas a los datos. Se ha notado especialmente en el caso de los equipos, donde hemos tenido que recurrir a la inicialización usando K-means para evitar tanta aleatoriedad por la escasa cantidad de puntos que había.

REFERENCIAS

[1] T. K. Moon. 1996. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* 13, 6 (1996), 47–60. https://doi.org/10.1109/79.543975

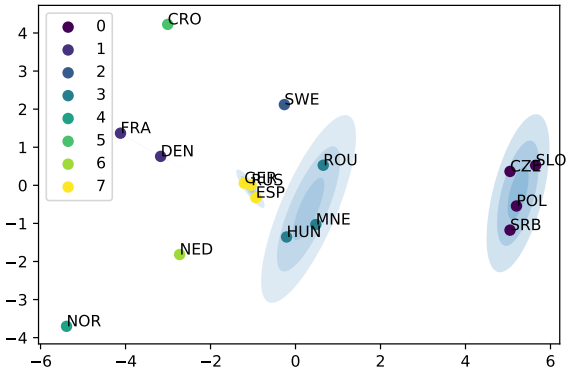

Figura 6: Agrupaciones de equipos usando la parametrización óptima según el BIC (K=8, 'full')

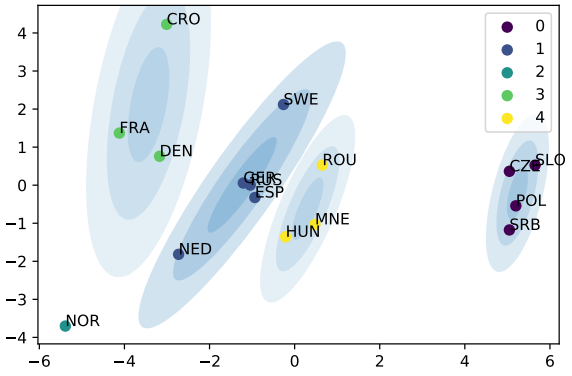

Figura 7: Agrupaciones de equipos usando una parametrización manual (K=5, 'full')