Coaching, Intelligence Artificielle et Science des Données

Cédric Damien, Omar El Euch, Théo Guillaumot,
Othmane Mounjid, Mathieu Rosenbaum
École Polytechnique
Contact: mathieu.rosenbaum@polytechnique.edu

19 avril 2018

Résumé

Nous présentons plusieurs outils d'aide à la décision pour un entraineur se basant sur des méthodes d'intelligence artificielle/machine learning et utilisant les données de matchs passés fournies par Opta. Nous proposons un replayer de match nous donnant la possibilité de mesurer les performances des joueurs et des équipes. Nous construisons aussi un simulateur de match nous permettant de prévoir les scores, noms des buteurs, possessions... Surtout, grâce à ce simulateur et à la théorie mathématique du contrôle stochastique, nous pouvons établir les compositions optimales d'une équipe selon son adversaire ainsi que les meilleures stratégies à utiliser pendant un match (quand faire des changements et quels remplacements effectuer par exemple). Nous illustrons ces résultats en proposant des compositions spécifiques de l'équipe de France en fonction de l'opposition et la liste des 23 joueurs la plus compétitive pour la coupe du monde 2018.

Mots-clés : Replayer de match, simulateur de match, prédiction, intelligence artificielle, machine learning, équipe de France, coupe du monde, liste des 23.

Attention nous nous contentons de faire parler les données. Les résultats présentés doivent être seulement considérés comme un outil d'aide à la décision !

1 On refait le match (comment se passer des données de tracking)

Opta fournit tous les événements **avec ballon** survenus pendant un match :

- Événements d'attaque : Passes, tirs, dribbles...
- Événements de défense : Tacles, récupérations de balle, duels aériens...
- Coups de pied arrêtés/remises en jeu : Fautes, touches, corners...
- Autres événements : Cartons, changements de joueur, changements de stratégie...

Les positions et identités des joueurs impliqués dans chaque événement sont données. Ceci permet de calculer plusieurs métriques, par exemple la probabilité qu'un joueur réussisse une passe ou un tir dans une certaine zone du terrain. Cependant ces données sont seulement centrées autour de la balle et nous n'avons pas d'information sur la position des joueurs ne touchant pas le ballon. On ne peut donc par exemple pas calculer d'indicateur mesurant les compétences défensives ou les qualités de placement d'un joueur et il est impossible d'utiliser le flux TV pour cela. **Comment donc inférer la position de tous les joueurs sur le terrain en ne connaissant que celle des joueurs touchant la balle ?** Notre replayer se base sur 3 méthodes.

- Méthode naïve : Pour un joueur donné, on collecte la suite des instants et des positions où il touche le ballon et on suppose qu'il se déplace linéairement entre ces instants avec une vitesse constante.
- Méthode quantile : Grâce aux données Opta, on connait la stratégie de chaque équipe (ex : 4-4-2, 4-3-3...) et le poste de chaque joueur. On estime alors la densité de positionnement sur le terrain d'un joueur évoluant à un poste dans une certaine stratégie. Quand un joueur donné possède la balle, on calcule sa déviation par rapport à la position moyenne de son poste. On suppose alors que les autres joueurs de l'équipe dévient de la même manière.
- Méthode Voronoï : Elle permet de modéliser les déplacements des joueurs pour effectuer un pressing/proposer des solutions à un coéquipier. Plus précisément :

– On suppose que l'équipe E_1 possède la balle et joue contre E_2.
– On définit la cellule de Voronoï de chaque joueur comme étant la région d'influence du joueur sur le terrain (il y a une définition mathématique rigoureuse).
– L'équipe E_2 qui défend cherche à réduire la surface d'influence de E_1.
– L'équipe E_1 qui attaque cherche à élargir sa surface d'influence.
– Cette surface d'influence est la surface de l'ensemble pondéré des régions d'influence des joueurs de E_1, avec un poids important sur les régions d'influence dangereuses (proches des buts) et la région d'influence du joueur qui possède la balle.

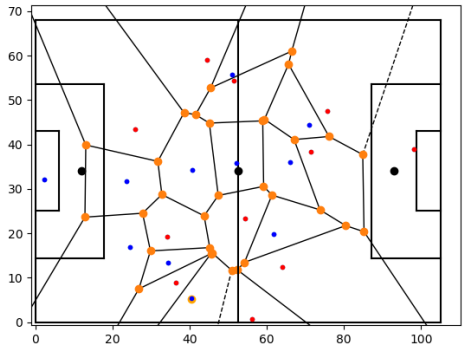

FIGURE 1 – Exemple de cellules de Voronoï appliquées à un screenshot du match Marseille-Toulouse (2016). Bleu : Marseille, Rouge : Toulouse.

La version finale du replayer est une combinaison optimisée de ces trois approches, où les positions des joueurs lorsqu'ils touchent la balle correspondent nécessairement aux positions observées en match. Il permet d'obtenir des résultats très proches de la réalité. L'illustration sur le match Real Madrid - Barcelone (17e journée de Liga 2017-2018) est présentée lors de la conférence.

2 On anticipe le match (modélisation)

Notre objectif est de modéliser un match entre deux équipes données afin de :

– Pouvoir faire des calculs prédictifs : score, possession, noms des buteurs...
– Permettre d'optimiser les compositions de départ selon l'adversaire et la stratégie en fonction de l'avancement du match.

Notre modèle est le suivant. Le terrain est divisé en 18 zones. Chaque joueur i_1 en zone j_1 peut :

– Transmettre le ballon à un coéquipier i_2 en zone j_2.
– Transmettre le ballon à un joueur de l'équipe adverse i_2 en zone j_2.
– Marquer un but.

La probabilité que l'un de ces événements se réalise dépend de :

– L'état actuel du match : positions des joueurs, score, temps...
– La capacité du joueur i_1 à transmettre la balle en zone j_2.
– La probabilité de présence du joueur i_2 en position j_2.
– La capacité du coéquipier (resp. adversaire) i_2 à recevoir (resp. intercepter) le ballon en zone j_2.
– La capacité de l'équipe adverse à intercepter le ballon en j_2.

Quand un joueur donné a la balle à un endroit donné, son action suivante et sa réussite ou son échec est décidée de manière aléatoire en fonction des capacités du joueur et de son environnement, selon un modèle mathématique calibré sur les données Opta des matchs passés. La méthode de calibration

se base sur des algorithmes d'intelligence artificielle/machine learning. À titre d'exemple on présente le classement obtenu sur 100 simulations pour la 2e partie du championnat de la saison 2016-2017 en Ligue 1, pour une calibration effectuée sur la première partie.

Équipe	Classement moyen	Écart-type du classement
Monaco	2.2	1.2
PSG	4.0	2.9
Nice	4.1	2.7
Bordeaux	4.6	2.2
Marseille	6.3	2.0
Angers	7.5	4.9
Lyon	8.7	3.9
Nantes	9.9	4.3
Lille	9.9	4.8
Dijon	11.0	4.5
Bastia	11.6	4.2
St-Etienne	12.6	4.7
Rennes	12.9	3.7
Montpellier	13.6	3.3
Guingamp	14.1	5.1
Lorient	14.5	3.7
Nancy	14.7	2.0
Metz	14.7	3.0
Toulouse	15.7	5.7
Caen	17.4	1.9

FIGURE 2 – Classement en simulation, phase retour 2016/2017.

3 On joue le match

Pour une composition de l'équipe adverse donnée, l'entraineur :

- Choisit sa tactique et composition de départ.
- Peut modifier sa stratégie tout au long du match (remplacements, changements de tactique).

Grâce à notre modèle et la théorie du contrôle stochastique, nous pouvons déterminer la suite de décisions optimale que doit prendre l'entraineur pour maximiser sa probabilité de gagner le match. Avant d'appliquer notre approche à la coupe du monde dans la section suivante, nous rappelons l'exemple simple présenté lors de la conférence :

- Match : huitième de finale retour de ligue des champions 2017/2018 : PSG-Real Madrid.
- Sortie de Benzema et rentrée de Bale à la $76^{\text{ème}}$.
- D'après notre modèle : changement pertinent, à l'instant quasi-optimal, permettant d'améliorer la probabilité de gagner d'environ 4%.

4 La coupe du monde

On applique ici notre modèle pour la prochaine coupe du monde.

4.1 Comment jouer à la coupe du monde ?

Notre méthodologie est la suivante :

- On considère 8 équipes (autres que la France) participant au mondial dont les 11 types sont constitués de joueurs de Bundesliga, Calcio, Liga , Ligue 1, Premier League : Allemagne, Angleterre, Argentine, Belgique, Brésil, Espagne, Portugal, Uruguay.
- On construit les onze types de ces équipes et estimons les paramètres du modèle pour ces équipes grâce aux données Opta (saison 2017/2018 jusqu'à fin mars).
- Pour près de 200 formations possibles pour l'équipe de France, on calcule les probabilités de victoire contre chacune de ces 8 équipes.
- On retient à chaque fois l'équipe pour laquelle la probabilité de victoire est la plus grande.
- Les compositions de l'équipe de France sont construites à partir des 17 joueurs suivants (joueurs quasi-certains de participer à la coupe du monde) :

  - Lloris, Mandanda ; Kimpembé, Koscielny, Mendy, Sidibé, Umtiti, Varane.
  - Kanté, Lemar, Matuidi, Pogba, Tolisso ; Dembélé, Giroud, Griezmann, Mbappé.

Nous obtenons alors différentes compositions optimales selon l'adversaire. Par exemple :

- Face au Brésil (Équipe type de la France la plus performante sur l'ensemble de la coupe du monde) : Lloris/Sidibé-**Varane-Koscielny**-Mendy/Tolisso-Kanté-Matuidi/Mbappé-Griezmann-Lemar.

- Face à l'Allemagne ou l'Espagne :
  Lloris/Sidibé-**Koscielny-Umtiti**-Mendy/Tolisso-Kanté-Matuidi/Mbappé-Griezmann-Lemar.

- Face au Portugal :
  Lloris/Sidibé-**Varane-Koscielny**-Mendy/Mbappé-Kanté-Matuidi-Lemar/Griezmann-**Giroud**.

- Face à l'Uruguay :
  Lloris/Sidibé-**Varane-Umtiti**-Mendy/**Pogba**-Kanté-Matuidi/**Dembélé-Mbappé-Lemar**.

On peut alors se demander quelle est la probabilité de victoire finale pour la France d'après notre modèle. Il est en fait difficile d'utiliser les valeurs brutes des probabilités obtenues car plusieurs facteurs de biais apparaissent : seulement 8 équipes, optimisation de l'équipe de France et pas des autres équipes... Après débiaisage, nous obtenons un ordre de grandeur de **14**% mais ce résultat est à prendre avec beaucoup de précautions.

4.2 La liste des 23 optimale

On se pose maintenant la question de la présence des joueurs suivants dans la liste des 23 pour la coupe du monde :

- Areola, Costil, Maignan, Ruffier.
- Debuchy, Pavard.
- Amavi, Digne, Hernandez, Kurzawa.
- Aouar, Bakayoko, Doucouré, Kondogbia, Nzonzi, Rabiot, Sissoko.
- Ben Yedder, Benzema, Lacazette.
- Coman, Fekir, Martial, Payet, Ribery, Thauvin.

Pour y répondre, on remplace (poste par poste) les joueurs de l'équipe de France type (celle qui donne la plus grande probabilité de victoire finale) par ceux-ci et on conserve ceux dégradant le moins la probabilité de victoire finale. On obtient la liste optimale suivante où le modèle ne permet pas de distinguer significativement Aouar, Doucouré et Kondogbia).

- Lloris, **Maignan**, Mandanda.
- **Amavi**, **Debuchy**, Kimpembé, Koscielny, Mendy, Sidibé, Umtiti, Varane.
- **Aouar ou Doucouré ou Kondogbia**, Kanté, Lemar, Matuidi, Pogba, Tolisso.
- Dembélé, **Fekir**, Giroud, Griezmann, **Lacazette**, Mbappé.