

RAPPORT INTERMÉDIAIRE

INF14 : Prédiction de la glycémie du diabétique à partir de ses activités

24 janvier 2018

—
Antonin Boniteau, Noranne Gabouge, Ayman Idrissi Kaïtouni,
Arthur Loison, Théo Mondou, Alexis Prodel, Nicolas Zucchet



TABLE DES MATIÈRES

I Résultats intermédiaires	3
I.1 Récupération et traitement des données	3
I.1.1 Structures et résultats	3
I.1.2 Difficultés face aux données récupérées	4
I.2 Machine learning	5
I.2.1 Spécificités des bases de données utilisées	5
I.2.2 Structure du code implémenté	6
I.3 Perspectives d'évolution en machine learning	7
II État des lieux de notre avancée et bilan intermédiaire	9
II.1 Comparaison avec notre planning	9
II.2 Remaniement de l'organisation et de la répartition du travail	10
II.3 Évolution de la répartition des tâches et du travail	11

RÉSULTATS INTERMÉDIAIRES

Comme cela était prévu, le travail s'est focalisé sur les deux aspects les plus importants du projet, à savoir la récupération ainsi que leur traitement et la phase d'apprentissage. La phase de récupération des données s'est échelonnée sur plusieurs semaines fin septembre et début octobre grâce à des montres connectées et un système informatique gracieusement fournis par Healsy. Toutefois, nous nous sommes heurtés à de nombreux obstacles majeurs durant le déroulement du projet concernant ces mêmes données. En effet, si la partie concernant le traitement des données était théoriquement finie, les données que nous récupérions nous ont déconcertés à plus d'un titre – nous y reviendrons dans la partie consacrée. Peut-être en raison d'une conception du travail trop linéaire, nous avons longuement essayé de déchiffrer certaines des données fournies par les montres, en vain, retardant ainsi la phase de commencement de l'apprentissage. En conséquence, nous avons pris du retard par rapport à l'échéancier exigeant figurant dans notre proposition détaillée.

Toutefois, nous avons pallié ce problème en débutant le machine learning sur des bases de données similaires à ce que nous les montres connectées sont censées nous fournir. Les différentes expérimentations et implémentations de machine learning que nous avons effectué sur ces bases de données sont très prometteuses, et constituent sans nul doute une très forte preuve de concept.

I.1 RÉCUPÉRATION ET TRAITEMENT DES DONNÉES

I.1.1 • STRUCTURES ET RÉSULTATS

Nous disposons en l'état d'un algorithme efficace effectuant ces différentes tâches :

- il prend en entrée des fichiers du type `heart-rate.csv` correspondant aux données brutes fournies par la montre connecté pour une seule donnée physiologique (e.g. le pouls) et sort un tableau épuré avec date et mesure sur deux colonnes ;
- il effectue cela pour le pouls, l'accélération linéaire et le gyromètre, il corrèle ensuite les trois tableaux de données pour en fournir un seul à quatre colonnes : temps (en timestamp), pouls, accélération linéaire, gyromètre. Il dédouble les données si besoin (le pouls est pris toutes les 5 secondes, les accélérations trois fois par seconde) ;
- il lit un fichier texte rempli par l'utilisateur contenant ses activités avec heure de début et heure de fin, associe un identifiant à chaque activité et ajoute le numéro de cette activité à chaque ligne correspondant à la bonne activité ;
- il effectue un découpage en fenêtres, avec des blocs de données d'une même activité pour augmenter artificiellement le nombre de données afin de faciliter l'apprentissage.

Ainsi, en l'état, l'algorithme prend le nom du dossier dans lequel se trouvent les .csv correspondant aux données brutes, ainsi que le fichier contenant les activités renseignées par la personne, et sort plusieurs fichiers .txt chacun associé à une grandeur mesurée :

- bpm.txt qui contient la liste chronologique des pouls ;
- gyrx.txt, gyry.txt, gyrz.txt qui contiennent les données du gyromètre respectivement selon x , y et z ;
- accx.txt, accy.txt, accz.txt qui contiennent l'accélération linéaire respectivement selon x , y et z .

Des données de même rang correspondent à une même mesure. Par exemple, si l'on a les fichiers suivants :

Fichier	Première valeur	Deuxième valeur	...
bpm.txt	79	81	...
gyrx.txt	3e-2	4e-2	...
gyry.txt	6e-2	4e-2	...
gyrz.txt	2e-2	1e-2	...
accx.txt	2e-2	1e-2	...
accy.txt	1e-2	7e-2	...
accz.txt	5e-2	5e-2	...

Alors [79, 3e-2, 6e-2, 2e-2, 2e-2, 1e-2, 5e-2] (i.e. la première colonne) correspond à la première mesure, et [81, 4e-2, 4e-2, 1e-2, 1e-2, 7e-2, 5e-2] (i.e. la deuxième colonne) correspond à la deuxième mesure. Ce format est particulièrement utile pour notre implémentation en machine learning.

I.1.2 • DIFFICULTÉS FACE AUX DONNÉES RÉCUPÉRÉES

Malheureusement, de nombreuses difficultés se sont profilées sur les données fournies par les montres, principalement autour des timestamps. Afin de mieux appréhender la situation, voici un extrait typique des données :

Username	heart_rate	end_time-date	create_time-date
'Alexis'	'96'	'2017-12-14 - 12:33:37'	'2017-12-14 - 11:05:07'
'Alexis'	'90'	'2017-12-14 - 07:00:32'	'2017-12-14 - 06:55:48'
'Alexis'	'79'	'2017-12-16 - 03:18:32'	'2017-12-15 - 22:22:00'
'Alexis'	'81'	'2017-12-15 - 16:51:56'	'2017-12-15 - 14:23:01'

On y voit plusieurs problèmes :

- à une mesure correspondent deux timestamps, deux temps : quel est le timestamp qui correspond effectivement à la mesure ? Nous pensons que l'un correspond au temps de la mesure, et que l'autre correspond au temps d'envoi de la donnée sur le serveur, ce que Nicolas Caleca nous a confirmé. Cela a été d'autant plus problématique que ces temps ne sont pas espacées de la même durée ;
- les données ne sont pas rangées dans l'ordre chronologique, que ce soit pour le temps de mesure ou le temps d'envoi. Nous nous attendions au moins à ce que les temps

d'envoi soient triés, mais ce n'est pas le cas, ce qui a rendu les tableaux plus difficiles à interpréter. Si cela semble anodin sur une poignée de données, c'est tout à fait différent sur plusieurs dizaines de milliers ;

- même après avoir trié et épluché les tableaux et les activités, on se rend compte que les temps sont souvent peu fiables. Parfois, un pouls de 120 est donné pour un individu censé faire une sieste. D'autres fois, on lit un pouls de 80 pour une séance de sport. Nous avons remarqué que les données étaient la plupart du temps décalées (quelques minutes, quelques heures), et qu'en plus ce décalage était variable.

Ce sont des obstacles considérables qui nous ont longtemps bloqués, d'autant plus qu'il a fallu effectuer de nombreux tests et examiner minutieusement les .csv afin de repérer les incohérences. De fait, l'ensemble des données que nous avions récupéré en septembre et octobre sont inutilisables.

Nous avons discuté récemment de ces problèmes avec notre tuteur, Nicolas Caleca. Healsy avait mis à notre disposition un système fait exprès pour nous et dont nous étions les seuls utilisateurs ; le système de récupération des données qu'ils utilisent durant leurs campagnes étant différent. Face à ces difficultés, Nicolas Caleca nous a affirmé que nous allions basculer sur ce deuxième système, qui, lui, a beaucoup été utilisé sans que quiconque n'ait repéré d'anomalies semblables à nos observations ; nous l'en remercions vivement.

De fait, le traitement des données que nous avons mis en place devra éventuellement être remanié pour prendre en compte des changements de format dans les données brutes, mais ce seront des changements mineurs (format d'entrée différent, par exemple), le fond du traitement des données étant fini.

I.2 MACHINE LEARNING

I.2.1 • SPÉCIFICITÉS DES BASES DE DONNÉES UTILISÉES

Compte tenu des difficultés que nous éprouvions quant aux données inexploitables, nous avons décidé d'utiliser deux bases de données disponibles en ligne, similaires à ce que nous pourrions récupérer des montres connectées. Ces bdd sont HARUS¹ et PAMAP2².

Il est sans aucun doute judicieux de s'intéresser aux méthodologies et protocoles mis en place pour ces bases de données afin d'avoir plus de recul sur les résultats que nous avons obtenus.

Concernant PAMAP2, il y avait 9 testeurs, pour 19 activités à reconnaître au total. Nous pouvons déjà émettre quelques réserves sur les activités : elles sont nombreuses, ce qui rend

1. "Human Activity Recognition Using Smartphones" Data Set, base de données disponible sur <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

2. "PAMAP2 Physical Activity Monitoring" Data Set, base de données disponible sur <http://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring>

la distinction parfois difficile, parfois trop spécifiques (on y trouve par exemple : marche nordique, repassage, corde à sauter) et parfois se recoupent (par exemple, être assis et regarder la télévision). De plus, le protocole expérimental demandait aux sujets d'effectuer les 19 activités pendant quelques minutes sur une durée totale d'une heure, cela diffère grandement de notre objectif et de nos méthodes. PAMAP2 recense environ 260 000 données.

Le nombre d'activités que les 30 sujets d'HARUS ont effectuées est plus restreint : elles sont au nombre de 6 et sont moins spécialisées – marcher, monter les escaliers, descendre les escaliers, être assis, être debout, être allongé. Les données obtenues ont également subi des transformations, à savoir un filtre passe-bas du troisième ordre à fréquence de coupure 20 Hz pour enlever le bruit, ainsi qu'une transformation de Fourier rapide sur certaines des données. Il est également à noter que le protocole d'HARUS ne mesurait que des données d'accéléromètre. HARUS recense environ 1 300 000 mesures.

Par ailleurs, la fréquence de mesure pour PAMAP2 était de 100 Hz, et la fréquence de mesure pour HARUS était de 50 Hz. C'est bien loin de ce que nous possédons avec les montres connectées, qui ont une fréquence de 3 Hz pour les données d'accélération, et de 0,2 Hz pour le pouls.

I.2.2 • STRUCTURE DU CODE IMPLÉMENTÉ

L'approche concernant l'apprentissage a été la suivante :

— création d'une classe DataReader, qui permet l'extraction et la mise en forme des données d'une base de données, ses arguments sont :

- dataset, le nom de la base de données ;
- n (facultatif), nombre de données maximal.

et ses fonctions sont :

- data_extraction qui extrait les données sous forme brute ;
- split qui sépare les données en un échantillon train et un échantillon test ;
- standardize qui normalise les données ;
- windowing qui sépare les données en différentes fenêtres.

Cette classe prendra en entrée les fichiers .txt épurés fournis par la partie de traitement des données.

— création d'une classe abstraite Model pour implémenter plus facilement les différents algorithmes de machine learning, ses attributs sont :

- data qui est une instance de la classe DataReader

et ses fonctions sont :

- train, pour entraîner le modèle sur les données d'entraînement du DataReader ;
- predict qui, une fois l'entraînement fait, prédit le résultat de ce qu'on lui donne ;
- efficiency qui permet de donner la précision du modèle et le tableau des erreurs.

- implémentation d'un réseau de neurones convolutif (CNN), avec tensorflow;
- implémentation d'une descente de gradient stochastique (SGD), avec scikit-learn.

Le CNN a été testé sur la base de données HARUS avec succès. Nous obtenons jusqu'à 97,6% de réussite sur cette base de données à 6 activités. Un graphique figure ci dessous afin d'y représenter les résultats que nous obtenons après variation des différents paramètres.

TODO Zucchet, pls

Le SGD a quant à lui été testé sur la base de données PAMAP2. Il a été choisi en raison de sa facilité d'utilisation et du nombre de mesures dans PAMAP2 (260 000). En y effectuant un découpage de fenêtres de 8 mesures, par exemple, on obtient une précision moyenne de 77,3%. C'est très encourageant. Ci-dessous figurent les différents tests que nous avons effectués, en faisant varier par exemple la taille du découpage en fenêtres.

TODO Théo pls

I.3 PERSPECTIVES D'ÉVOLUTION EN MACHINE LEARNING

Il y a un grand nombre de possibles améliorations pour l'apprentissage que nous mettons en place.

En premier lieu, l'un des premiers enjeux sera d'implémenter un grand nombre de modèles pour en tester les limites et les résultats. Par exemple, nous envisageons la mise en place d'algorithme de clustering comme le k -nearest neighbours afin de le confronter avec que que nous possédons d'ores et déjà, et pouvoir utiliser de manière intelligente l'algorithme le plus efficace. Par ailleurs, il est certain que les performances des différents algorithmes dépendront des tailles des échantillons. Tester ces différentes limites nous permettra de mettre en place des valeurs seuils pour la taille du data set : par exemple, en dessous d'un certain n , tel méthode sera utilisée, et au dessus, ce sera une autre.

La documentation de la bdd HARUS suggère également d'effectuer un raffinement préliminaire des données : en effet, les données brutes d'Harus subissent un filtrage passe-bas et une transformation de Fourier rapide. L'un de nos objectifs sera de mesurer l'impact qu'a l'ajout d'une variable non linéaire sur la performance du modèle. Un premier exemple peut être la norme euclidienne de l'accélération, qui se déduit non linéairement des 3 données d'accélération linéaire que nous fournissent dès à présent les montres connectées. Une batterie de tests est prévue pour déterminer la meilleure approche.

De même, la documentation de la bdd PAMAP2 suggère également une autre idée. Nous avons vu que certaines des activités se regroupaient, par exemple « être assis » et « regarder la télévision ». Par conséquent, introduire une notion de *distance* entre les différentes activités pour proposer des activités connexes en cas de mauvaise prédiction semble être judicieux.

Finalement, nous avons également pensé à une approche transversale des différents labels/activités que nous devons retenir pour le machine learning. Il s'agirait là de minimiser le

nombre d'activités à prédire et donc de regrouper toutes les activités qui, physiologiquement, auront les mêmes conséquences pour l'individu du point de vue de la glycémie.

II

ÉTAT DES LIEUX DE NOTRE AVANCÉE ET BILAN INTERMÉDIAIRE

II.1 COMPARAISON AVEC NOTRE PLANNING

Les échéances que nous avions fixées n'ont été globalement pas respectées. D'une part, certains des objectifs étaient peu réalistes, comme celui d'obtenir un algorithme efficace le 15 novembre. En outre, ces objectifs n'ont pas pu être tenus car la collecte de nos données personnelles a posé d'énormes difficultés qui ont effectivement rendu ces échéances impossibles : comment disposer d'un algorithme efficace avec une durée d'apprentissage inférieure à 10 jours de mesure alors même que les données que nous récupérions étaient inexploitables ? Ces problématiques ont considérablement réduit l'avancée de notre projet. Nous avons par conséquent dû modifier nos objectifs et entamer la phase de machine learning sur des bases de données différentes. Voici les différentes phases de travail que nous avions proposées :

5 Sept. - 20 Sept.	Tester la collecte de données et la compatibilité de l'application de Healsy.
20 Sept. - 11 Oct.	Collecter et traiter les données brutes.
20 Sept. - 4 Oct.	Déterminer le type d'algorithme à utiliser et le mettre en œuvre sur des cas simples ; identifier les grandeurs statistiques pertinentes ; affiner le choix des critères de classification.
11 Oct. - 15 Nov.	Programmer l'algorithme et le mettre en œuvre sur les données à disposition.
15 Nov. - 20 Déc.	Améliorer les performances du programme, notamment en termes de durée d'apprentissage.

La première phase a été respectée. Pour toutes les raisons que nous avons exposées précédemment, le traitement des données, censé se terminer le 11 octobre, s'est doublé de leur analyse et a duré jusque fin novembre – début décembre et a concentré la plupart des efforts du groupe. La détermination du type d'algorithme utilisé et sa mise en place ont donc été retardées, commençant pour la première fin novembre, et pour la deuxième durant les premières semaines de décembre.

II.2 REMANIEMENT DE L'ORGANISATION ET DE LA RÉPARTITION DU TRAVAIL

Une des principales conséquences de ces retards et obstacles a été une remaniement important de la répartition du travail et des différentes tâches. Afin de mieux la comprendre, voici un tableau actualisé de nos tâches (les nouvelles figurent en gras) :

<i>Antonin</i>	Testeur pour la collecte de données Chargé des relations avec la start-up Healsy et le tuteur
<i>Alexis</i>	Testeur pour la collecte de données Analyse des données et de leur cohérence, conduite de tests
<i>Arthur</i>	Testeur pour la collecte de données Centralise et gère les données récupérées via l'application Analyse des données et de leur cohérence, conduite de tests Réalisation du traitement des données
<i>Ayman</i>	Chargé de rédaction des différent rapports Réalisation du traitement des données
<i>Nicolas</i>	Prépare, conduit et rédige les comptes rendus des réunions hebdomadaires Chargé de l'aspect programmation Mise en place du machine learning, réseau de neurones
<i>Noranne</i>	Prépare, conduit et rédige les comptes rendus des réunions hebdomadaires Rappelle les échéances et les objectifs intermédiaires Machine learning, descente de gradient stochastique
<i>Théo</i>	Chargé de l'aspect programmation Machine learning, descente de gradient stochastique

Une différence notable concerne la création de plusieurs pôles de travail. Notre organisation originelle ne prévoyait pas ces derniers et était moins sophistiquée : tout le groupe était censé travailler sur les mêmes problématiques. Nous nous sommes peu à peu rendu compte que l'approche optimale n'était non pas de concentrer nos efforts sur un seul enjeu afin de le terminer plus rapidement, mais plutôt de traiter en parallèle (et donc, plus lentement) les différentes composantes du projet.

II.3 ÉVOLUTION DE LA RÉPARTITION DES TÂCHES ET DU TRAVAIL

L'organisation précédente devrait être stable pour les prochaines semaines. En effet, une fois le nouveau type de données mis en place, nous achèverons leur traitement et nous effectuerons une batterie de tests pour mettre un terme à cette phase de travail. Ce travail devrait prendre une voire deux semaines, si aucun obstacle majeur ne se présente. Ceci fait, Antonin, Alexis et Arthur entameront de nouveau la collecte des données. Pendant ce temps, le pôle traitement des données basculera en renfort sur la partie machine learning, qui sera affinée jusqu'à ce que l'on puisse mettre à l'épreuve nos propres données.