Проект по Машинно обучение и вероятностно моделиране

Георги Дежов

12 февруари 2020

1 Цел

Целта на проекта е да се създаде работещ модел за разпознаване на човешки дейности на база данните *WISDM SMARTPHONE AND SMARTWATCH ACTIVITY AND BIOMETRICS*.

2 Увод

"WISDM Smartphone and Smartwatch Activity and Biometrics" включва данни, събрани от 51 субекта, всеки от които е помолен да изпълни 18 дейности за по 3 минути. Всеки субект е имал "умен" часовник поставен върху доминиращата му ръка и смартфон в джоба си. Събирането на данните е било извършено чрез мобилно приложение. Събрани са данни от сензорът на акселерометъра и на жироскопа както от смартфона, така и от смарт часовника, получавайки данни от общо четири сензора. Данните са записвани със скорост 20 Hz (т.е. на всеки 50 ms). За смартфон е използван или Google Nexus 5 / 5X или Samsung Galaxy S5 с Android 6.0. За часовник е използван LG G Watch с инсталиран Android Wear 1.5. Общите характеристики на данни и процесът на събиране на данни са обобщени в Таблица 1.

В таблица 2 са изброени 18-те дейности, които са били извършвани.

Таблица 1: Обща характеристики на данните и процесът на събиране на данни

Брой субекта	51
Брой дейности	18
Минути за изпълнение на дейност	3
Скорост за запис	20Hz
Използван смартфон	Google Nexus 5/5x или Samsung Galaxy S5
Използван смарт часовник	LG G Watch
Брой измервания	15,630,426

Таблица 2: 18-те дейности представени в базата данни

Код	Дейност	Код	Дейност
ходене	A	ядене на паста	J
тичане	B	пиене от чаша	K
качване на стълби	C	ядене на сандвич	L
седене	D	ритане	M
стоене	E	хващане	O
писане на машина	F	дриблиране	P
миене на зъби	G	писане на ръка	Q
ядене на супа	H	ръкопляскане	R
ядене на чипс	I	сгъване	S

В текущото изследване дейностите ще бъдат разделени на две основни групи: съответно на дейности, в които не се използват главно ръце като: вървене, тичане, качване на стълби, седене, стоене и ритане на топка; и съответно такива, които се изпълняват основно с ръцете. Във вторите попаднат останалите дейности. Ще изберем един човек на случаен принцип и ще се опитаме да съставим алгоритъм за разпознаване на дейността, която извършва, т.е. дали е свързана с ръце или не, на база събраните данни от акселератора на телефона в джобът му.

3 Свързани публикации

Има доста публикации свързани с темата, някои от които са “K. Yoneda and G. M. Weiss (2017). Mobile Sensor based Biometrics using Common Daily Activities, Proceedings of the 8th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference , New York, NY, 584 590.”, “G.M. Weiss, J.L. Timko, C.M. Gallagher, K. Yoneda, and A. J. Schreiber (2016). Smartwatch based Activity Recogn i- tion: A Machine Learning Approach, Proceedings of the 2016 IEEE International Conference on Biomedical and Health Info r- matics (BHI 2016) 2016), Las Vegas, NV, 426 429.” и “J.R. Kwapisz, G.M. Weiss and S. A. Moore. Activity Reco g- nition using Cell Phone Accelerometers, ACM SIGKDD Explorations , 12(2):74 82”

4 Методология

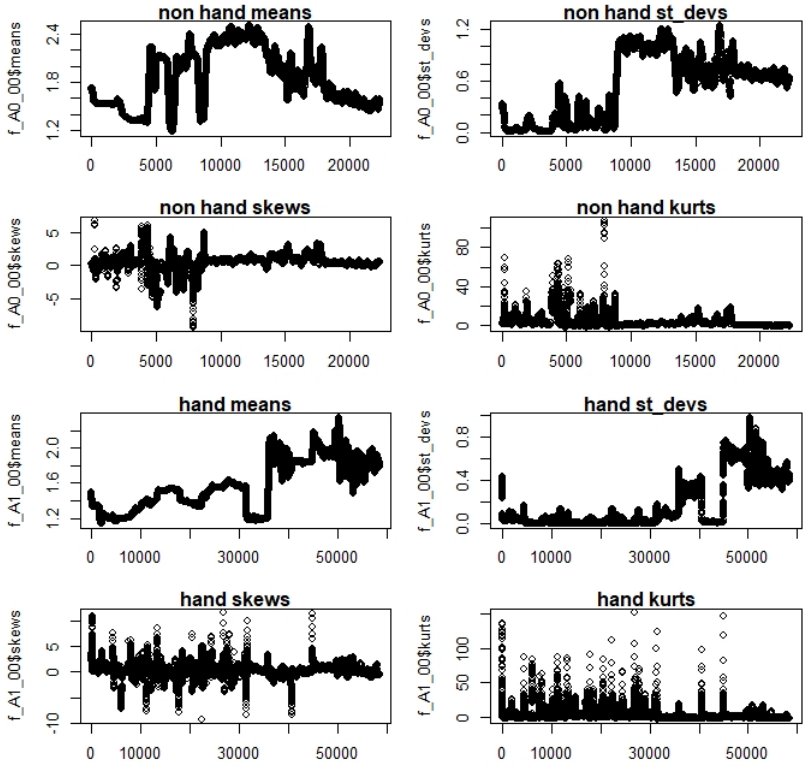
На случаен принцип избираме един от 51-та човека, които са изпълнявали въпросните дейноти. В конкретният пример се е паднал субект номер 5. Взимаме данните записани от акселераторът на телефонън, който е бил в джоба му и добавяме параметър, който да индикира разделението на дейностите на такива извършвани главно с ръце и такива, в които не се ползват ръцете активно. В първата група дейности попадат: писане на машина, миене на зъби, ядене на супа, ядене на чипс, ядене на паста, пиене от чаша,

ядене на сандвич, хващане, дриблиране, писане на ръка, ръкопляскане и сгъване. Във втората група попадат: вървене, тичане, качване на стълби, седене, стоене и ритане на топка.
Проверяваме за липсващи данни и форматите на съответните величини. В случая акселераторът на телефона записва линейното ускорението измерено по трите координата X, Y и Z като се използват мерни единици m/s^2. За по-лесна работа тези величини се нормализират като извади съответната средна стойност и се разделят на стандартното отклонение.
При така нормализирани величини съставяме един нов параметър, който е корен квадраен от сумата на квадратите на X, Y и Z . За удобство ще го кръстим "dist".
За да можем да идентифицираме по-добре дейностите използваме метода на плъзгащия се прозорец. В конкретния случай сме използвал прозорец с 200 точки като преди това сме разделили базато по дейности. Това ни дава запис на всеки ред за стойностите на новия параметър за интервал от 10 секунди, тъй като 200=10sec/(50sec/1000).
За определяне на разпределението на новият параметър може да използваме някой от четирите главни момента или всичките, а именно: средно, стандартно отклонение, асиметрия и ексцес.
На тази база създаваме две подбази данни, които ще използваме за трениране на модела и за проверка. Съотношениео в обема на двете подбази е 0.7 за трениране и 0.3 за тестване.
Така създаваме логистична регресия при което отклика ни е индикаторната променлива дали се използват ръцете, а като предиктор може да се използва средно, стандартно отклонение, асиметрия и ексцес или комбинация от тях.
Обученият модел го използваме, за да направи прогнози за това каква е дейнстта, която се извърша, в смисъл дали се използват активно ръцете или не. Моделът го изпробваме първо върху данните за трениране и сравняваме с истинският индикатор. В последствие се изпробва върху данните отделени за тестване, т.е. непознати данни. Така можем да провери дали има оверфитване или ъндерфитване. Ако моделът се справя много по-лошо от очекваното при нови данни, то имаме оверфитване.

5 Резултати

След като изберем субектът за когото ще се опитаме да прогнозираме дейностите, които извършва, в конкретния случай се е паднал човек номер 5, създаваме индикаторна променлива показваща дали се ползват ръце.
При проверката за качаството на данните се вижда, че няма липсващи данни. При първоначалното зареждане на данните виждаме, че стойностите на променливата Z са заредени като факторна променлива. Обръщаме я на числа, за да можем в последствие да я нормализираме.
Визуализацията на средното, стандартното отклонение, асиметрията и ексцеса съответно за дейности, които не се извършват активно с ръце и такива,

които се извършват активно с ръце са показани на графиките по-долу.



При създаване на логистична регресия с предиктор само средното получаваме статистически значима регресия, т.е. е по-добре отколкото да ползваме средната стойност на средните без да превим регресия. Също така получаваме AIC: 58105.

При логистична регресия с предиктор средното и стандартното отклонение също получаваме, че предикторите са статистически значими. При този модел получаваме AIC: 52674.

И така за модел със средно, стандартното отклонение и асиметрия също получаваме, че предикторите са статистически значими и AIC: 52574.

И накрая за модел със средно, стандартното отклонение, асиметрия и ексцес също получаваме, че предикторите са статистически значими и AIC: 52157

Като направим прогнози използвайки последния модел получаваме Таблица 3, която е сравнителна матрица на предсказани и истински стойности, кото са използвани данни за трениране: В таблица 4 е представено сравнение като се използват тестови, т.е. нови данни

Този модел познава средно в 79% от случаите при нови данни.

Таблица 3: Реални към предсказани стойности от данните за трениране

	Реални	
Предсказани	0	1
0	7721	3926
1	7923	37047

Таблица 4: Реални към предсказани стойности от данните за тестване

	Реални	
Предсказани	0	1
0	3358	1649
1	3347	15912

6 Заключение

Виждаме, че и четирите модела са статистически значими като стойността на "Akaike information criterion" намалява. Това е индикатор за по-молко тестови грешки, т.е. бихме избрали този модел, който има най-малко AIC или четвърти модел с случая.
Ако разгледаме Таблица 4 виждаме, че относително голям брой предсказания са верни, ако дейностите са с активно използване на ръце, т.е. познаваме в повече от 90% от случаите. Но когато дейностите реално не са с активно използване на ръце то познаваме само около 50% в случаите.