Universitat de Girona
Escola Politècnica Superior

Treball final de grau

Estudi: Grau en Enginyeria Informàtica

Títol: Hyper and hypo glycaemia detection based on bag of words

Document: Resum

Alumne: Eduard Berloso Clarà

Tutor: Beatriz Lopez Ibañez

Departament: Enginyeria Elèctrica, Electrònica i Automàtica

Àrea: ESA

Convocatòria (mes/any): juny / 2017

Hyper and hypo glycaemia detection based on bag of words

1 Introducció

Els pacients amb diabetis han d'estar controlant contínuament els seus nivells de glucosa i prenent insulina per tal de mantenir-se en els nivells normals i evitar així episodis de hiperglucèmia i hipoglucèmia. En aquest treball, s'estudia la utilització de tècniques de Machine Learning sobre dades històriques de pacients junt amb les dades que es tenen abans d'un àpat (carbohidrats, glucosa, insulina, etc.) per tal de predir el futur estat del pacient: hiperglucèmia, normoglucèmia o hipoglucèmia. D'aquesta manera un pacient podria saber el risc que té de patir un d'aquests episodis abans que es produeixi i actuar en conseqüència, per exemple variant la quantitat de insulina a prendre.

2 Dataset

El treball es basa amb les dades proporcionades per l'Imperial College of London al grup de recerca eXiT. El data set està format per 10 pacients amb diabetis de tipus 1. Aquests pacients portaven un sensor que proporcionava lectures de glucosa (mmol/l) cada 5 minuts.

A més de les lectures de glucosa, es disposa de un seguit d'informació addicional sobre els pacients: edat i sexe; i sobre l'estat abans de l'àpat: carbohidrats, insulina, si s'ha pres alcohol i si s'ha realitzat exercici abans o després d'un àpat. Després d'un procés de neteja (eliminant àpats en que no es tenien lectures suficients de glucosa) el dataset resultant va ser de 1099 àpats (a partir d'ara anomenats "sessions"). Una sessió té: edat, sexe, carbohidrats, insulina, glucosa, exercici abans, exercici després, alcohol i la corba de glucosa fins al següent àpat.

3 Metodología

La metodologia presentada en aquest treball combina una aproximació a la *tècnica bag of words* [1] amb les tècniques de classificació KNN [2] i SVM [3]: es genera un vocabulari de les corbes postprandials dels pacients, es generen seqüències de les corbes de les sessions i finalment, junt amb la resta de dades de les sessions (carbohidrats, insulina, etc.) s'hi aplica una tècnica de classificació (SVM o KNN).

S'ha inclòs un disseny d'interfícies web amb finalitats demostratives de la recerca. Tot el desenvolupament s'ha fet seguint la metodologia RUP.

3.1 Generació de Histogrames i Vocabulari

Cada sessió es classifica com a hiperglucèmia, normoglucèmia o hipoglucèmia segons la següent definició:

- Una lectura de glucosa per sota de 3.889 mmol/l es considera hipoglucèmia.
- Si no hi ha hipoglucèmia, una lectura per sobre de 10 mmol/l es considera hiperglucèmia.
- Si no es produeix cap de les dos situacions es classifica com a normoglucèmia.

Es generen els histogrames pertanyents a les corbes de glucosa de les sessions. Abans de generar aquests histogrames, s'aplica una sèrie de preprocessos a les corbes de glucosa:

- Filtratge: només es treballa amb les lectures de 1.5 a 6 hores, ja que es considera que és en aquest període de temps en que la insulina afecta al pacient i es poden produir els episodis de hiperglucèmia i hipoglucèmia.
- "Smooth": es tracte d'un procés en que es 'suavitzen' els valors de les lectures de cada sessió de manera que una lectura comparada amb l'anterior i la següent els canvis siguin més petits.
- Normalitzar: implica canviar l'escala de la distribució de valors de manera que la mitjana observada és 0 i la desviació estàndard és 1.
- Reescalar: canviar l'escala de les dades de la gamma original de manera que tots els valors observats estan dins l'interval 0 i 1.

S'aplica la tècnica de clustering K-Means [4] als histogrames resultants de les corbes de glucosa ja preprocessades. Els centroides generats amb aquesta tècnica formen el vocabulari amb el que es treballarà. A cada sessió se li assigna el centroide (o paraula) més proper a l'histograma de la seva corba de glucosa.

3.2 Bag of Words: Generació de Seqüències

El model *bag of words* [1] és un mètode utilitzat en el processament de llenguatge per representar documents ignorant l'ordre de les paraules. La idea bàsica és representar un document com un histograma amb el número de cops que apareix cada paraula.

Aquesta idea es porta sovint al camp del reconeixement d'objectes per representar imatges. Una imatge pot ser tractada com un document i les característiques extretes de certs punts de la imatge són considerades com paraules.

El seu funcionament és el següent:

- Identificar les paraules rellevants
- Definir un vocabulari amb les paraules més freqüents
- Es compta les ocurrències de cada paraula del vocabulari dins el document
- Cada document es representa com un histograma del vocabulari

En aquest treball s'ha aproximat aquesta tècnica i s'ha utilitzat per representar una sessió. Es considera que cada corba de glucosa entre dos àpats és una paraula i els documents són les seqüències de sessions. Per tal de reduir la mida del vocabulari i no tenir una paraula diferent per cada corba de glucosa que hi ha en el sistema s'ha aplicat clustering (k-means) per generar un vocabulari de K paraules seguint el procés descrit en l'anterior apartat.

Per generar les seqüències, per cada sessió s'obtenen les N sessions anteriors pertanyents al mateix pacient (N és un nombre fixat). Per tant, cada sessió estarà representada per les paraules pertanyents a les N sessions anteriors a ella a més de les dades pre-àpat referents a la pròpia sessió (carbohidrats, glucosa inicial, insulina, edat, sexe, etc.).

3.3 Model Predictiu

S'ha utilitzat dues tècniques de classificació diferents KNN [2] i SVM [3]. Per cadascuna d'elles, s'ha realitzat una *grid-search* per tal de obtenir-ne la millor configuració de paràmetres, d'acord amb els mètodes propis de la llibreria sklearn.

4. Servidor Web i Conceptes de Disseny

S'ha implementat un servidor que proporciona una interfície web per a la visualització de dades i interacció amb el sistema que realitza prediccions.

Durant el disseny, s'ha seguit la metodologia de desenvolupament RUP utilitzant diferents eines de l'enginyeria del software: diagrames de classes, fitxes de cas d'ús, etc. També s'ha aplicat patrons de disseny com el patró decorator a l'hora de resoldre certs problemes de implementació.

Per tractar amb les dades del dataset amb el que s'ha treballat s'ha dissenyat i construït una base de dades utilitzant el sistema de gestió de bases de dades SqLitle3.

5. Avaluació

Per tal d'avaluar el sistema s'ha utilitzat la tècnica stratified k-folds cross-validation [5] amb k=5. S'han avaluat diferents escenaris en els que es variava:

- Mida del vocabulari
- Llargada de les seqüències
- Representació de les sessions: s'ha implementat diferents maneres de representar sessions amb petites diferencies entre elles. Per exemple: representacions alternatives dels histogrames a partir de funcions gaussianes, normalitzacions dels atributs pre-àpat, etc.
- Model predictiu utilitzat: KNN o SVM

La mètrica escollida per puntuar el sistema ha estat la AUC. S'ha escollit aquesta contra altres més comunes com la accuracy degut a que el dataset amb el que es treballa està desbalancejat: hi ha moltes més hiperglucèmies que normoglucèmies i hipoglucèmies. Això implica que les puntuacions que dona la accuracy no siguin representatives de la qualitat del sistema.

6. Resultats

La millor puntuació obtinguda ha estat una AUC de 0.65 per hiperglucèmies utilitzant SVM. Correspon al escenari en que es normalitzen tots els atributs, el vocabulari te mida 15, les seqüències son de llargada 3 i el model predictiu utilitzat és el SVM.

Els resultats mostren que el SVM és lleugerament superior al KNN a l'hora de fer les classificacions. Tot i això, la diferència no és mot significativa i el SVM te un percentatge de no classificats més alt (un 32% del SVM contra un 23% del KNN).

També s'observa una major precisió a l'hora de classificar si una sessió és hiperglucèmia o no hiperglucèmia respecte les altres dos classes. És un efecte lògic degut al desbalanceig de les dades.

7. Conclusions

Queda clar que la puntuació del sistema no és suficientment bona com per determinar que es classifiquen correctament les seqüències de sessions dels pacients.

Aquest fet pot ser degut a dos raons (o a una combinació d'elles):

- Realment no existeix una correlació entre les corbes de glucosa que impliqui que la aparició d'una comportí l'aparició de la següent. És a dir, que una corba de glucosa i l'anterior son independents entre elles i no segueixen cap patró.
- Incompletesa de les dades: s'ha detectat clarament que les dades amb les que es treballava eren força incompletes (es podia donar el cas que en un dia hi haguessin la mitat de lectures que en el dia següent, o fins i tot dies sense lectures), desbalancejades, i fins i tot incoherents (pics de glucosa, dies amb 5 àpats). En altres projectes en que s'ha treballat amb les mateixes dades també s'han trobat amb aquest problema [6].

En resum, és possible que les corbes de glucosa siguin unes dades insuficients per ser utilitzades directament en algorismes de classificació Machine Learning. En tot cas, s'ha posat de manifest els límits de les dades actuals disponibles.

Referències

[1] Albert Pla, Natalia Mordvanyuk, Beatriz Lopez, Marco Raaben, Taco J. Blokhuis, Herman R. Holstlag. Bag-of-steps: Predicting Lower-limb Fracture Rehabilitation Length.

[2] Fermin Pitol. KNN, el algoritmo del vecino mas cercano (22/3/2014). http://ferminpitol.blogspot.com.es/2014/03/k-nn-k-nearest-neighbor-el-algoritmo.html (01/06/2017)

[3] Sunil Ray. Understanding Support Vector Machine algorithm (06/10/2015). https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/ (01/06/2017)

[4] Tan, Steinbach, Kumar. The k.means algorithm. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html (01/06/2017)

[5] Kenneth Jensen (IBM). K-fold Cross-validation in IBM SPSS Modeler (02/03/2013). https://developer.ibm.com/predictiveanalytics/2016/03/02/k-fold-cross-validation-ibm-spss-modelar (01/06/2017)

[6] Fabien Dubosson, Natalia Mordanyuk, Beatriz Lóopez, and Michael Schumacher. Prediction of postprandial hypoglycemias from insulin intakes and carbohydrates: analysis and comparison between real and simulated datasets. 2nd Workshop on Artificial Intelligence for Diabetes, MIE, Vienna, 2017