Universitat de Girona
Escola Politècnica Superior

Treball Final de Màster

Estudi: Màster en Ciència de Dades

Títol: Detecció de dades anòmales en dades de cabal

Document: Resum

Alumne: Joan Saló Grau

Tutor: Marc Comas Cufí
Departament: Informàtica, Matemàtica Aplicada i Estadística
Àrea: Estadística i investigació operativa

Convocatòria (mes/any): Setembre 2023

Resum

1 Objectius

En aquest treball ens centrarem a detectar les anomalies de les dades majoritàriament causades per un mal funcionament de les estacions d'aforament, que són les estacions que mesuren el cabal d'aigua que circula pels diferents cursos fluvials de les CIC (Conques Internes de Catalunya), a diferència d'altres treballs, on es detecten anomalies relatives a patrons poc freqüents, però no incorrectes, com poden ser episodis de pluja extrema (que causen pics de cabal) o d'extrema sequera (com podria ser el cas actual). D'aquesta manera, el nou conjunt de dades filtrat pot ser utilitzat per entrenar models robustos que ajudin a comprendre els processos hidrològics, i no només això, sinó que també és interessant l'aplicació d'aquest treball pels gestors de les estacions d'aforament, ja que la presència de noves dades anòmales pot indicar un mal funcionament d'aquestes.

2 Desenvolupament i implementació

Per tal de detectar les anomalies, hem dissenyat dos mètodes: el primer (mètode 1), es basa a modelitzar les dades de cabal utilitzant l'algorisme de modelatge Prophet (dissenyat per Facebook) utilitzant GAM (Model Additiu Generalitzat), i considerar el que estigui fora d'un cert interval de confiança detectat pel mateix algorisme com a anomalia.

El segon mètode (mètode 2) parteix d'un etiquetatge previ de les mostres segons si són anòmales o no. Amb les dades etiquetades, podem modelitzar el cabal de les dades correctes de totes les estacions utilitzant una sola xarxa Long Short-Term Memory (LSTM) però no a partir de les mateixes dades de cabal, sinó a través de variables externes com la precipitació, temperatura, etc., d'aquesta manera, podem a arribar a esmenar errors que encara puguem tenir dins les dades. Un cop això, podem entrenar un LSTM Autoencoder, a partir de seqüències de n observacions i prediccions, on l'última observació és correcte. D'aquesta manera, i basant-se en el supòsit que les prediccions de les mostres correctes seran més exactes que la de les mostres incorrectes, podem dissenyar un classificador que a partir d'una sèrie de prediccions i observacions, ens ajudi a determinar si aquestes observacions són anòmales o no.

3 Conclusions

Els resultats dels algorismes proposats no han sigut satisfactoris, ja que aquests no han sigut capaços de detectar les anomalies, en gran part a causa de la complexitat d'aquestes, les quals no es poden detectar mitjançant mètodes estadístics clàssics, sinó que depenen en gran part del comportament de la sèrie i un conjunt de patrons concrets.

El mètode 1 és el que ha donat pitjors resultats, en gran a part a causa de la simplicitat de l'aproximació, i no és capaç de detectar gairebé cap anomalia de forma correcta (menys d'un 1% de les anomalies classificades són correctes). Pel que fa al mètode 2, sí que els resultats han sigut lleugerament positius, i tot i que les prediccions del cabal amb la xarxa LSTM són millorables, i l'error obtingut és elevat (tenim un NSE de 0.2 de mitjana per cada estació), el model entrenat és capaç d'entendre la dinàmica general, tot i que té problemes per predir els pics de cabal. Tot i això, el classificador LSTM Autoencoder no detecta bé les anomalies, però realitzant una inspecció visual de les prediccions obtingudes, i amb l'ajuda d'un expert, aconseguim detectar algunes anomalies. De totes maneres, caldria continuar treballant en les prediccions provant altres mètodes o integrant més dades per tal de millorar els resultats.