

EXAMENSARBETE Anomaly Detection with Machine Learning in OpenStreetMap Changesets**STUDENT** Dan Svenonius**HANDLEDARE** Patrik Edén (LU), Hampus Londögård (AFRY AB)**EXAMINATOR** Jacek Malec (LTH)

Fungerar maskininlärning för att hitta fel i stora mängder geografisk data?

POPULÄRVETENSKAPLIG SAMMANFATTNING **Dan Svenonius**

OpenStreetMap är som Wikipedia för kartor, där vem som helst kan bidra. Eftersom databasen är så stor är det svårt att ha en bra kvalitetskontroll, vilket gör validering av bidrag med maskininlärning intressant, och det visar sig att det funkar ganska bra.

Tänk er att ni får i uppgift att skapa ett system för kvalitetskontroll av bidrag till hela Wikipedia. Svårt? Det är vad vi försökt göra i det här examensarbetet, fast med OpenStreetMap (OSM) och geografisk data. Ett bidrag är en samling individuella redigeringar, som att ändra ett adressnamn eller lägga till ett hus. När det är så många sorters fel som kan uppstå blir det orimligt att skapa ett antal regler, exempelvis att ett hus inte kan ligga i havet, för att göra denna kontroll. Då blir det naturligt att istället testa om maskininlärning kan fungera, då det hade kunnat flagga misstänksamma bidrag som får genomgå en mer ordentlig validering.

Här behöver vi dock veta om ett bidrag var bra eller dåligt, så vi är till synes är tillbaka på ruta ett. Istället hittar vi bidrag som av en annan användare har dragits tillbaka (utan att vi vet varför), och säger att dessa borde innehålla något konstigt. För de bra bidragen så drar vi slumpmässiga bidrag och antar att en stor majoritet av bidragen är bra.

Vi testar detta i fyra olika maskininlärningsmodeller (två baserade på neurala nätverk och två på beslutsträd) och får oväntat bra resultat - den bästa modellen har rätt på 9 av 10 bidrag i genomsnitt. Det visar sig även att modellerna baserade på beslutsträd med råge fungerar bäst.

Något väldigt intressant är att modellerna fungerar så bra *trots* att de inte har tillgång till information om specifika redigeringar, utan snarare om bidraget som helhet och dessutom användarens erfarenhet. Man hade kunnat tänka sig att den går mest på användarerfarenhet, men det är relativt balanserat mellan dessa två kategorier.

Detta visar på den otroliga potentialen i maskininlärning - att algoritmen hittar regler och mönster snarare än vi - både i denna uppgift, men egentligen överallt där stora datamängder finns.



Även LTH:s Sjön Sjon finns med i OpenStreetMap, inlagd här som "Lake Lake".