

Datavalidering

Datavalidering er afgørende for at sikre, at sensor-data er korrekte. Der er flere metoder hertil og man benytter nogle gange dem alle for at opnå den størst mulige sandsynlighed for at sensor-data er korrekte. Den første metode benyttes altid før en eller flere af de øvrige.

1. Grundlæggende validering

Dette drejer sig om simple undersøgelser af det enkelte data fra en sensor.

Datatype og format undersøges af software, fx er det et heltal som forventet eller et kommatal som forventet. Det kan også være der forventes et JSON format og software kan også undersøge om strukturen er JSON. Hvis type eller format er noget andet end forventet, kan sensoren have fejl eller softwareudvikler har misforstået hvilken datatype eller format sensoren skal levere.

Der kan være en forventning til hvilket interval data ligger indenfor, fx kan det forventes at et heltal som repræsenterer en temperatur i celsiusgrader, ligger indenfor -50 til +50. Hvis heltallet ligger udenfor intervallet, så er der noget som ikke virker og tallet kan ikke anvendes til beregninger eller beslutninger.

Nulværdier er sådanne værdier som en sensor leverer, hvis der slet ikke blev målt nogen værdi. For heltal er det ofte 0 (nul) eller kunne være et bestemt meget stort tal, der også samtidig ligger udenfor det forventede interval. Nulværdier indikerer også at noget ikke virker og nulværdier må ikke benyttes til beregninger eller beslutninger.

2. Tidsbaseret validering

Her undersøges en serie af data fra en sensor og hver værdi er mærket med klokkeslæt for hvornår den er målt/genereret.

Meget ofte er der nemlig en sammenhæng mellem efterfølgende data i en serie. Det kan fx være med måling af udendørstemperatur, hvor der om formiddagen typisk fås gradvis stigende værdier og fra sidst på eftermiddagen gradvis faldende værdier. I en sådan serie, vil fx et spring på 10 på en time være en klar fejl og værdien med det store udsving må ikke anvendes i beregninger eller beslutninger og skal fjernes fra serien.

Ved manglende data på tidspunktet hvor ny data forventes, bør der typisk ske en reaktion og systemet bør advare om at der kan være et teknisk problem, fx ustabil datakommunikation eller at sensoren kan være uden strøm. Fx hvis temperaturværdier normalt kommer med 30 minutters mellemrum, så forventes denne rytme at fortsætte.

3. Multisensor validering

I mere kritiske situationer, hvor det er af meget stor vigtighed at sensor-data er korrekte, benyttes flere sensorer til at måle det samme og så sammenlignes data fra flere sensorer.

Hvis man fx vil måle en vejtemperatur for at afgøre hvornår der skal saltes om vinteren, så placeres fx sensorer i vejen med 200 meter distance mellem dem. Hvis flere sensorer fortæller at vejens temperatur er 1 celsius grad, så er der stor sandsynlighed for at dette er korrekt. Hvis tre sensorer fortæller at temperaturen er 1 celsius grad, men én fortæller at den er 5 celsius grader, så anses den sidste værdi for ukorrekt og kan ikke anvendes i en beslutning om at starte med at sprede salt.

4. Statistisk filtrering

Dette er en systematisk måde til at mindske betydningen af afvigende værdier fra en serie.

En populær metode er at benytte glidende gennemsnit. Det kan fx være gennemsnit over 5 på hinanden følgende værdier i serien. I stedet for at benytte de enkelte sensorværdier i efterfølgende beregninger/beslutninger, så er det glidende gennemsnit, som benyttes. En enkelt afvigende sensorværdi får en langt mindre betydning, når den ikke benyttes direkte, men indgår i gennemsnitsberegninger.

Der er andre statistiske metoder, der reducerer betydningen af afvigende værdier i en serie eller fjerner værdierne helt fra serien. Der kan fx benyttes standardafvigelse og en værdi som afviger mere end den acceptable standardafvigelse, fjernes ganske enkelt fra serien.

5. Fejlhåndtering og alarmer

Automatisk fejlkorrektion kan benyttes ved manglende data, hvad enten data aldrig blev modtaget fra sensoren eller blev fjernet pba. en af valideringsmetoderne ovenfor.

En simpel fejlkorrektion kan være lineær interpolation, hvor en manglende værdi bliver sat til gennemsnittet af værdien lige før og lige efter i serien. Dette vil typisk fungere ganske godt for mange forskellige typer af data. Dette gælder dog ikke, hvis der fx måles på en proces, der vides at resultere i værdier der ikke følger en lineær udvikling. Hvis udviklingen vides fx at være eksponentiel, så skal benyttes en eksponentiel funktion for interpolationen.

Der kan alternativt benytte machine learning til at forudsige næste værdi i en serie og hvis værdien mangler (eller er helt skæv i forhold til forudsigelsen), så indsættes i stedet den forudsagte værdi i serien. Metoden bør dog ikke anvendes gennem længere tid, da nye forudsigelser i stigende grad baseres sig på tidligere forudsigelser af værdier i serien. Dermed kan forudsigelser gradvis blive mere og mere forkerte.

Hvis værdier afviger mere end normalt, men stadig ligger indenfor forventet interval eller hyppigheden af manglende værdier er stigende, bør det altid udløse en alarm fra systemet, da der sandsynligvis er tale om en teknisk fejl. Man skal dog være opmærksom på at baggrunden også kan være en form for angreb mod systemet. Fx kan en sensor som er placeret fast udendørs bliver stjålet og herefter måle temperatur i et helt andet miljø, der fx kunne være indendørs.

6. Validering fra edge devices

En edge device gennemfører lokale beregninger og benytter sig evt. også af simple machine learning modeller. Den vil også have indbygget flere sensorer men leverer værdier den har beregnet ud fra flere målte værdier og evt. som resultater af benyttelse af machine learning.

Sådan enhed kan også blive fejlbehæftet eller evt. udsat for angreb. Den kan så begynde at levere værdier, som er forkerte.

Principielt kan benyttes metoderne 1-5 ovenfor, men det kan også være at der ikke er noget oplagt sammenhæng mellem værdier i den serie enheden sender. Fx beregner enheden hastigheden af passerende cyklister og der er ikke rigtig noget sammenhæng mellem hastigheden af to cyklister som passerer med 10 minutters mellemrum. Så hvis én hastighed beregnes til 15 km/t og en anden til 30 km/t, så kan vi ikke sige at nogen af dem er forkerte.

Man kan dog stadig med statistiske metoder afgøre hvis der sker en udvikling over dage/uger/måneder og fx genere alarm, hvis daglig gennemsnitshastighed af cyklister stiger fra 20 km/t til 30 km/t over en måned.

Man kan også benytte machine learning centralt til at vurdere om der sker en udvikling væk fra en normalsituation. Algoritmen indlærer den normale situation med de normale/naturlige udsving og kan reagere på at data bevæger sig væk fra den normale situation.

Udarbejdet af Birger Andersen, DTU Engineering Technology, som led i InnoTech-TaskForce projektet 2022 - 2025