

Handbuch: 2.6. Daten-Voraussetzungen

Jede Datenanalyse hängt von der Qualität der zur Verfügung gestellten Daten ab. Wir gehen in zunächst davon aus, dass alle wichtigen Aspekte des Prozesses oder der Maschinerie im Datensatz aufgenommen werden. Fehlen wichtige Daten, wird die Modellierung nicht so gut funktionieren wie wenn diese Daten bereitstehen. Auf der anderen Seite ist es auch nicht gut, das Modell mit einer Vielzahl von unwichtigen Messungen zu überfrachten. Die wichtigste Tätigkeit zur Modellierung ist eine vernünftige Auswahl der Tags, die für die Modellierung bereit stehen sollen. Wir können sagen, dass bei einer Industrieanlage mit einem komplexen Leitsystem weniger als 10% aller verfügbaren Messungen für die modellhafte Darstellung des Prozesses nötig und wichtig sind.

Falls Sie im Zweifel darüber sind, ob eine Messung berücksichtigt werden soll oder nicht, sollten Sie sich eher für (und nicht gegen) diese Messung entscheiden. Der Grund dafür ist derselbe, der auch für menschliches Lernen entscheidend ist: Wenn Sie irrelevante Informationen bereitstellen, mag das zeitaufwändig und irritierend sein, aber diese Informationen werden Sie nicht daran hindern, Zusammenhänge zu begreifen. Fehlen aber wichtige Informationen, so kann es sein, dass Sie kein ausreichendes Verständnis gewinnen. Im Zweifelsfall ist es also besser, Messdaten mit einzubeziehen als sie wegzulassen.

Die meisten Archivsysteme verfolgen die Regel, nach der sie den neuen Wert eines bestimmten Tags nur dann aufzeichnen, wenn es von dem letzten aufgezeichneten Wert um einen Mindestbetrag abweicht. Diesen Mindestbetrag bezeichnet man als den Kompressionsfaktor. Das bedeutet, dass manche Messungen häufig aufgezeichnet werden und andere nur selten. Für den Zweck des Aufzeichnens ist dies ein enorm platzsparender Mechanismus.

Für die Analyse und das maschinelle Lernen müssen wir die Daten allerdings zeitlich abgleichen. Für jeden Zeitstempel müssen wir den Wert jedes Tags zu dieser Zeit kennen. Die meisten Methoden des maschinellen Lernens erfordern es, dass der Zeitabstand zwischen den fortlaufenden Messungen immer derselbe ist.

Um von der üblichen historischen Datenaufzeichnung zu dieser Tabelle der abgeglichenen Daten zu kommen, verwenden wir die allgemeine Regel, dass der Wert eines Tags solange als derselbe angenommen wird, bis wir einen neuen Wert erhalten. Eine Zeitreihe kann dann wie eine Treppe aussehen. In den meisten Fällen führt diese Datenauflistung zu einem Wachstum des gesamten Datenvolumens, weil viele Werte oftmals erscheinen. Das wird man aber nicht verhindern können.

Aus diesem Grund wählen wir ein vernünftiges Datenintervall. Das heißt, dass die Zeitabstände zwischen den sukzessiven Ablesungen in dieser Tabelle nicht zu klein sein dürften, um nicht zu viele Daten anzusammeln, aber auch nicht zu groß, weil sonst der dynamische Prozess nicht sichtbar wird. Das Zeitintervall sollte deshalb mit Sorgfalt ausgewählt werden, möglichst basierend auf der Erfahrung der inhärenten Zeitskala des Prozesses, den man modellieren möchte.

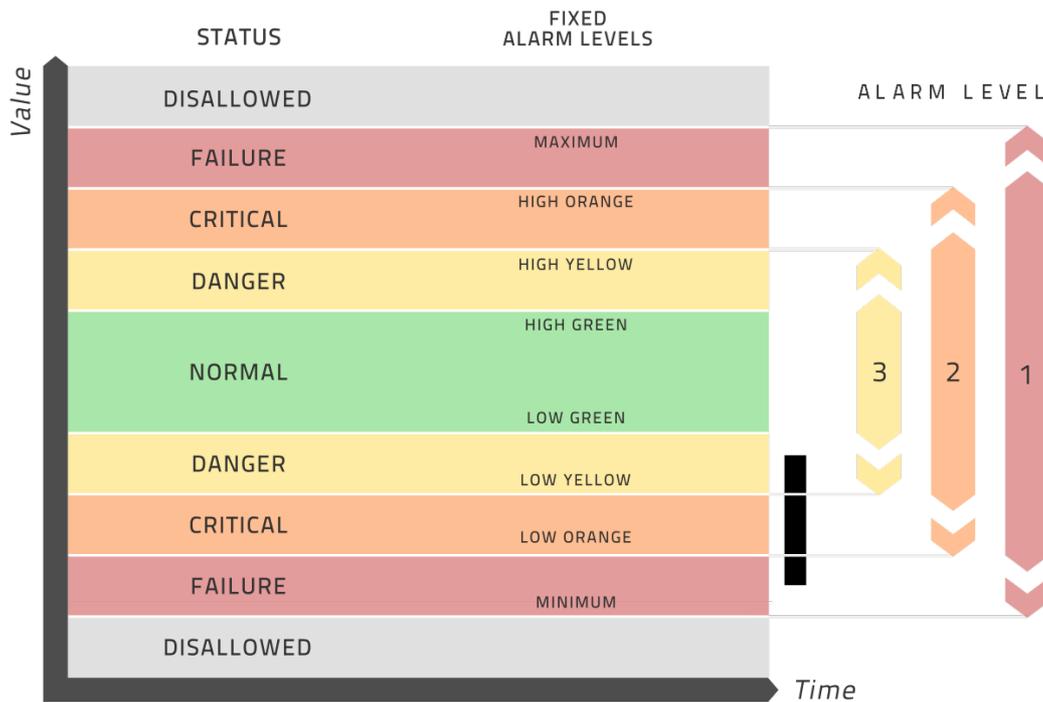
Ein Datensatz muss einen Anfang und ein Ende haben. Um ein IHM-Modell zu trainieren, wählen wir sorgfältig eine Trainingsperiode aus, so dass wir für eine bestimmte Zeitspanne das reibungslose Verhalten der zu überprüfenden Maschine

aufzeigen können. Sind wir davon überzeugt, dass externe Bedingungen, etwa die Jahreszeiten, für eine genaue Darstellung des Prozesses von entscheidender Bedeutung sind, so benötigen wir Daten für diese verschiedenen Bedingungen. Um den reibungslosen Zustand einer Maschine genau darzustellen, reicht in der Regel eine Zeitspanne von drei bis sechs Wochen aus. Diese Wochen müssen nicht unbedingt fortlaufend sein. Fallen saisonale Veränderungen ins Gewicht, wählen wir je eine Woche aus dem Frühling, dem Sommer, dem Herbst und dem Winter. Für ein Optimierungs-Modell empfehlen wir allerdings ein ganzes Jahr aufzuzeichnen, um die saisonalen Varianten ausreichend einzubeziehen.

Für die Zeitspannen, die wir für das Training des Modells auswählen, kann es aber vorübergehende Zustände geben, die wir nicht zugrunde legen sollten. Wenn eine Maschine beispielsweise jeden Abend abgeschaltet wird, wollen wir das Modell natürlich nicht auf diese Zeit ausrichten, während der die Maschine nicht funktionsbereit ist. Aus diesem Grund bietet die Software Ausschlussmöglichkeiten für bestimmte Messwerte an. Man könnte etwa bestimmen, dass alle Datenpunkte, bei denen die Rotationsrate einer Turbine weniger als 500 U/min beträgt, vom Modell ausgeklammert werden sollen. Die für solche Entscheidungen notwendigen Tags müssen dann in den Datensatz aufgenommen werden.

Nachdem diese Entscheidungen getroffen wurden, müssen die ausgewählten Tags dem Modell bekannt gemacht werden, indem gewisse Informationen bereitgestellt werden. Neben einigen verwaltungstechnischen Informationen wie Name und Werteinheit benötigen wir noch einige zusätzliche Fakten. Weil es zuweilen Messfehler gibt, müssen wir die erlaubte Bandbreite der Messungen kennen, damit sehr niedrige oder sehr hohe Messwerte als Ausreißer identifiziert und ausgeklammert werden können. Wir benötigen auch die Messunsicherheit, damit wir diese Unsicherheit in den Ergebnissen der Analyse (als Ergebnisunsicherheit) berücksichtigen können. Wenn Sie mehr darüber wissen möchten, schauen Sie im Abschnitt über den mathematischen Hintergrund nach. Für IHM müssen wir auch wissen, welche Messungen dynamische Begrenzungen erhalten sollen, die einen Alarm auslösen. In der Praxis werden die allermeisten Messungen des Modells keine Alarmsignale hervorrufen, sie liefern aber Informationen und den Kontext für diejenigen Messungen, die Alarme hervorrufen sollen. Für APO etwa müssen wir wissen, welche Tags direkt vom Anlagenfahrer kontrolliert werden können und welche nicht kontrolliert werden können.

Als Alternative könnten Sie für jede Messung bis zu drei statische Alarm-Spielräume spezifizieren. Das ist notwendig, damit IHM die traditionelle Zustandsüberwachung unterstützen kann – zusätzlich zum Ansatz der dynamischen Grenzwerte. Dieses System von Analyse und Alarm unterscheidet sich von der Modellanalyse und ist als völlig optional zu betrachten. Diese Bandbreiten werden am besten in der Grafik unten erklärt.



Summa Summarum: dies sind die Entscheidungen, die Sie zu Beginn im Hinblick auf Ihren Datensatz treffen müssen:

1. Welche Tags sollen aufgenommen werden?
2. Welcher Zeitabstand soll benutzt werden?
3. Über welchen Zeitraum soll der Datensatz laufen?
4. Über welchen Zeitraum soll das Training laufen?
5. Welche allgemeinen Zustände sollen ausgeklammert werden?
6. Welche Zustände sollen für den jeweiligen Tag ausgeklammert werden?
7. Für jeden Tag gilt es zu fragen:
 1. Welches soll die erlaubte Bandbreite der Messwerte sein?
 2. Wie unscharf ist die Messung?
 3. Soll dieser Tag modelliert werden und Alarmer auslösen?
 4. Kann dieser Tag direkt vom Anlagenfahrer kontrolliert werden?
 5. Als Option: Welches sind die Bandbreiten für die statischen Alarmmeldungen bei gelben, orange und roten Alarmsignalen?