Zustandsdiagnose von Maschinen im Kontext von Industrie 4.0 unter Einsatz von Data-Mining Methoden

Christof Küstner[1], Jürgen Mitsch[1], Matthias Hegwein[1],
Nico Meintker[2], Konrad Mönks[2], Moritz Fröhlich[2],
Sandro Wartzack[1]

[1] *Lehrstuhl für Konstruktionstechnik (KTmfk);*
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

[2] *GE Jenbacher GmbH & Co OG, Achenseestr.1-3, 6200 Jenbach, Österreich*

Abstract

Today, the machine industry offers a range of After-Sales-Services, including the installation and commissioning of the purchased products, the provision of spare parts, inspections or comprehensive maintenance contracts. Studies prove that After-Sales-Services often generate revenues that are several times higher than the original purchase price, which is why these services significantly contribute to a company's balance sheet. Worldwide operation of machines could lead to long machine downtimes before service technicians reach the operating site. Therefore, companies are interested in machine prognostics and predictive maintenance techniques. To accomplish this, an approach for condition-based maintenance in context of the internet of things and the use of data mining is presented in this contribution.

Keywords: condition-based maintenance, machine prognostics, internet of things, data mining

1 Einleitung und Herausforderung

Industrieunternehmen bieten heutzutage eine Vielzahl an Zusatzprodukten und Dienstleistungen an, die den Kunden über den Erwerb eines Produktes hinaus unterstützen und an das Unternehmen binden sollen. Diese werden unter dem Begriff After-Sales-Services zusammengefasst und beinhalten bspw. die Installation und Inbetriebnahme der erworbenen Produkte, die Bereitstellung von Ersatzteilen, Inspektionen oder umfassende Wartungsverträge. Doch die Bedeutung von After-Sales-Service geht heute weit über die Kundenbindung hinaus. Sie sind sowohl Unterscheidungsmerkmale im globalen Wettbewerb zur Neukundengewinnung, als auch zuverlässige und vor allem langfristige Einnahmequellen [3]. Studien zeigen, dass im Verlauf des Lebenszyklus eines Produktes durch die After-Sales-Services oftmals ein Vielfaches des ursprünglichen Kaufpreises erwirtschaftet und so ein erheblicher Beitrag zur Gesamtbilanz eines Unternehmens geleistet wird [5]. Besonders bei komplexen Maschinen und Anlagen wie z. B. Automobilen, Turbinen oder industriellen Großmotoren sind Wartungsverträge weit verbreitet, um einen reibungslosen Betrieb für den Kunden zu gewährleisten. Diese Maschinen werden i. d. R. weltweit betrieben (bspw. Gasmotoren für Biogasanlagen), weshalb im Fall von kleinen Wartungsintervallen oder Störungen der Betrieb für mehrere Tage ausfallen kann, bis die Wartungstechniker den Betriebsort erreichen.

Aus diesem Grund sind Unternehmen an einer prädiktiven Zustandsdiagnose von Maschinen interessiert, um u. a. dem Aspekt des *Designs for Maintenance* im Kontext von Industrie 4.0 gerecht zu werden. In diesem Beitrag wird in diesem Kontext eine Methode vorgestellt, womit die datengetriebene, prädiktive Zustandsdiagnose von Maschinen ermöglicht wird. Hierdurch können Störungen vorhergesagt, bzw. im Fall einer Störung die Dauer des Maschinenstillstandes verkürzt werden.

2 Wartungskonzepte von Maschinen im Kontext von Industrie 4.0

Allgemein betrachtet gibt es drei Konzepte, wie die Wartung und Instandhaltung einer Maschine erfolgen kann [7]: (vgl. Bild 1) Reaktive Wartung: Wartungs- und Reparaturmaßnahmen erfolgen erst nach Eintritt einer Störung; Zeitbasierte/vorbeugende Wartung: Inspektionen und Wartungen erfolgen in festen, periodischen Zeitintervallen; Zustandsabhängige Wartung: der Maschinenzustand wird kontinuierlich überwacht, die Wartungen erfolgen, sobald sich der Maschinenzustand verschlechtert und eine Störung droht.

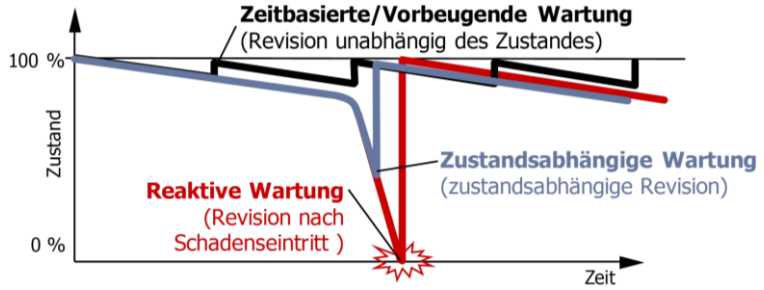

Bild 1: Wartungskonzepte von Maschinen nach [6]

In der Praxis am weitesten verbreitet ist die klassische, vorbeugende Wartung. Sie bietet für viele Unternehmen den besten Kompromiss aus Planbarkeit, Realisierbarkeit und Gewährleistung einer hohen Maschinenverfügbarkeit, denn sie kann durch die klassischen Dimensionierungsansätze ermittelt werden. Die vorbeugende Wartung erfolgt stets mit einer ausreichend großen Sicherheit (dementsprechend Vorlaufzeit), um eine mögliche Störung und somit Stillstand der Maschine zu vermeiden.

Im Kontext von Industrie 4.0 und der stetig zunehmenden Vernetzung von Maschinen werden sehr viele SCADA-Daten (Supervisory Control And Data Acquisition) der Maschinen erhoben. Diese SCADA-Daten stellen die sensorisch erfassten Zustandsdaten einer Maschine dar, welche die Basis für die Steuerung und Überwachung einer Maschine darstellen und deshalb kostenneutral zur Verfügung stehen. Ziel der zustandsabhängigen Wartung ist, auf Basis der SCADA-Daten eine Verschlechterung des Maschinenzustandes zu antizipieren und im Idealfall selbstständig eine Wartung einzuleiten. Dies beugt unnötig frühen Austausch von Komponenten und auf lange Sicht häufigere Wartungsarbeiten vor, wodurch im Vergleich zur vorbeugenden Wartung erhebliche Kosten eingespart werden können. [7]

3 Zeitabhängige Daten

SCADA-Daten sind zeitabhängige Daten, auch Zeitreihen genannt. Sie lassen sich allgemein als eine Menge von Werten eines Vektors $Y = \langle y_1, y_2, y_3, \dots, y_T \rangle$ beschreiben, welcher über einen definierten Zeitraum (Index T) nacheinander gemessen und nach der Zeit geordnet sind [1]. Prinzipiell wird zwischen stationären und instationären Zeitreihen unterschieden. Stationäre Zeitreihen schwanken stets um einen konstanten Mittelwert (Niveau), z. B. mechanische Spannungsschwingungen (vgl. Bild 2-a). Instationäre Zeitreihen hingegen weisen Trends oder zyklische Wiederholungen auf, sodass kein konstantes Niveau vorliegt, z. B. Aktienkurse (vgl. Bild 2-b).

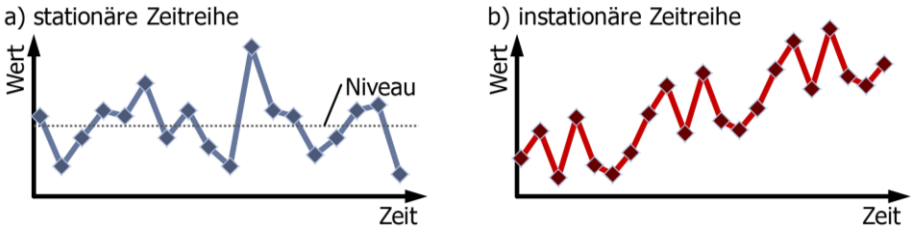

Bild 2: Qualitative Darstellung einer stationären und instationären Zeitreihe

Stationäre Zeitreihen lassen sich in erster Linie durch ihr Niveau beschreiben, instationäre Zeitreihen hingegen können zur Beschreibung in mehrere Komponenten aufgeteilt werden [10]: Trendkomponente, Saisonkomponente, Zyklische Komponente und Restkomponente. Die Trendkomponente beschreibt die langfristige Veränderung, d. h. den Anstieg oder den Abfall des Niveaus über den betrachteten Zeitraum. Die zyklische Komponente und die Saisonkomponente hingegen beschreiben wiederkehrende Schwankungen in unterschiedlich langen Perioden. So bezieht sich die Saisonkomponente meist auf Schwankungen innerhalb eines Jahres, während die zyklische Komponente Schwankungen über den Zeitraum mehrerer Jahre beschreibt. Schließlich werden in der Restkomponente sämtliche Einflüsse gesammelt, welche sich keiner der anderen Komponenten zuordnen lassen. Darunter fallen unvorhersehbare, kurzfristige Einflüsse wie z. B. „Naturkatastrophen" eines Aktienkurses. Dieses Komponentenmodell nach Schlittgen [10] kann für instationäre Zeitreihen für die Prognose des zukünftigen Verlaufs der Zeitreihe verwendet werden. Sensoren an technischen Systemen erzeugen i. d. R. stationäre Zeitreihen, weshalb auf deren Charakterisierung und Analyse in den folgenden Abschnitten näher eingegangen wird.

4 Einsatz von Data-Mining zur Mustererkennung in Zeitreihen

Aufgrund des stetig steigenden Daten- und Informationsaufkommens werden das gezielte Finden, Auslesen und Verarbeiten von Mustern aus Datenbanken oder -strömen zu immer größeren Herausforderungen. Ohne technische Hilfsmittel können diese in vielen Fällen nicht mehr identifiziert werden. Dies liegt an der Anzahl der Variablen eines Systems oder der Komplexität der Daten einer betrachteten Maschine, welche vom Ingenieur nicht mehr interpretiert werden können. Die computerunterstützte Analyse großer Datensätze unter Anwendung statistischer Methoden wird unter dem Begriff *Data-Mining* zusammengefasst. Scarpa [9] bspw. definiert Data-Mining als die Tätigkeit grafischer oder numerischer Analyse von großen, Datensätzen o-

der -strömen, um für den Dateneigentümer nützliches Wissen zu finden. Der Ausdruck nützliches Wissen ist hierbei bewusst allgemein gehalten, da oft vor Beginn der Analyse noch nicht feststeht, was der Gegenstand des Interesses ist. Bissantz und Hagedorn [2] beschreiben das durch Data-Mining zu extrahierende Wissen allgemein als ein Muster, das im Auge des Betrachters interessant ist und mit ausreichender Sicherheit tatsächlich existiert. Der Begriff Muster wiederum bezeichnet Beziehungen von Daten und Regelmäßigkeiten zwischen mehreren Datensätzen oder innerhalb eines Datensatzes. Im Folgenden werden die wichtigsten Aspekte vorgestellt, die bei der Mustererkennung in stationären Zeitreihen mittels Methoden aus dem Bereich des Data-Minings berücksichtigt werden müssen.

4.1 Auswahl der Data-Mining Methodengruppe zur Mustererkennung in stationären Zeitreihen

Die Data-Mining Methoden werden generell in überwachtes und unüberwachtes Lernen unterteilt (s. Bild 3). Welche der beiden Verfahren durchgeführt werden kann, hängt von den verfügbaren Daten ab: Ist die Zielgröße (zu analysierendes Attribut eines Datensatzes) nicht bekannt oder zu aufwändig zu bestimmen, kommen unüberwachte Lernverfahren zum Einsatz [11]. Dies ist der Fall, wenn zwar verschiedene Sensordaten zur Verfügung stehen, aber nicht bekannt ist, ob, wann und wo eine Störung oder ein Wartungsfall eingetreten ist. Hierbei werden anhand der Struktur der Sensordaten eindeutige Attribute (engl. features) identifiziert, welche die Daten bspw. in Störungsklassen (Art der Störung) oder zeitliche Priorität für eine Wartung unterteilen. Ist die Zielgröße der Daten hingegen bekannt, d. h. das zu analysierende Attribut ist bereits Teil des Datensatzes, wird von überwachtem Lernen gesprochen [11]. Dies ist der Fall, wenn bspw. für verschiedene Sensordaten vorab die Störungsklassen oder zeitlichen Prioritäten für eine Wartung bekannt sind. Beide Verfahrenstypen suchen Muster, welche die Beziehung zwischen den bekannten Zielgrößen in den Datensätzen und den Attributen möglichst genau beschreiben. Dieser Vorgang wird als Lernen oder Trainieren eines Modells bezeichnet. Das Resultat dieser Trainingsalgorithmen werden u. a. Metamodelle genannt.

Für jedes Data-Mining Verfahren stehen charakteristische Methodengruppen zur Verfügung (vgl. Bild 3). Beim überwachten Lernen kommen Algorithmen aus den Methodengruppen der Klassifikation oder Regression zum Einsatz mit welchen prognosefähige Metamodelle trainiert werden können [11]. Unter der Prognose kann eine Inter- und Extrapolation verstanden werden. Diese Metamodelle können bspw. direkt auf den Maschinensteuerungen betrieben werden und somit das zukünftige Verhalten der Maschine abschätzen,

womit Störungen vorhergesehen werden können. Beim überwachten Lernen im Kontext der Zeitreihenanalyse eignet sich besonders die Klassifikation zur Unterscheidung von Störungsklassen, wobei die Regression beim Extrapolieren von Sensorsignalen zum Einsatz kommt. Hier sind rekurrente neuronale Netze als bekannteste Algorithmengruppe zu nennen, welche eine Rückkopplung in ihrer internen Struktur aufweisen, womit eine Prognose des zukünftigen Sensorsignalverlaufs auf Basis der Signalhistorie möglich ist [8]. Beim unüberwachten Lernen gibt es vier Methodengruppen (vgl. Bild 3). Hier eignen sich die Visualisierung und vor allem das Clustern für die Zeitreihenanalyse. Beim Clustern können Maschinen identifiziert werden, welche sich nicht durchschnittlich verhalten.

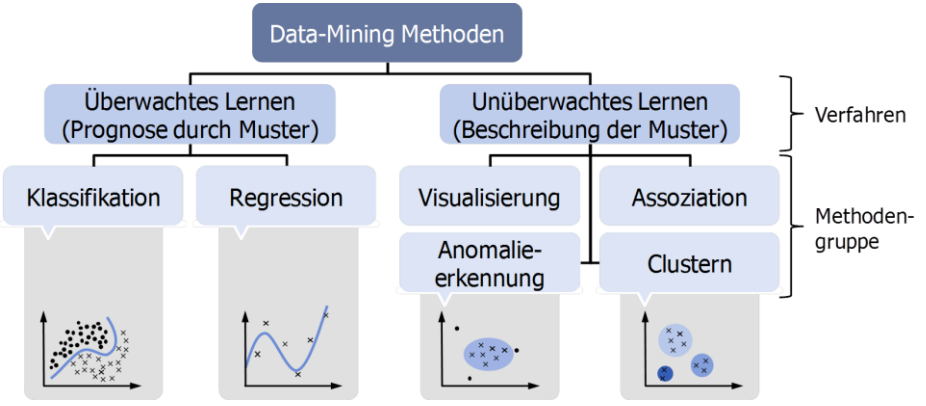

Bild 3: Strukturierung der Data-Mining Methoden nach [11]

4.2 Klassifikation stationärer Zeitreihen

Zeitabhängige Daten sind in ihrer „rohen Form" (s. Kapitel 3) aufgrund ihrer quantitativen Skalierung für Klassifikationsverfahren nicht geeignet, weshalb die Daten durch eine Transformation reduziert werden müssen. Diese Datenreduktion, welche ebenso als Identifikation von Attributen bezeichnet wird, wird allgemein Featureextraktion genannt. Möglichkeiten der Featureextraktion sind u. a. die Berechnung von Ähnlichkeits- oder Distanzmaßen zwischen Zeitreihen, Fourier- und Wavelettransformation und die Hauptkomponentenanalyse [4]. Neben der Featureextraktion stellt die hohe Dimensionalität der Zeitreihendaten bei häufig gleichzeitig kleiner Anzahl von Trainingsdaten im Kontext einer Klassifikationsaufgabe eine weitere Herausforderung dar. Dies zeigt sich besonders darin, dass Zeitreihen mit einer hohen Zahl an Da-

tenpunkten durch die Transformation auf eine einzige Instanz (Datenpunkt) reduziert werden können.

Eng verknüpft mit der Anzahl der verfügbaren Trainingsdaten ist auch die maximale Anzahl der für das Training nutzbaren Attribute. Mit steigender Anzahl von Attributen erhöht sich die Dimensionalität der Klassifikationsaufgabe. Wird die Dimension der Klassifikation überproportional groß, werden die Modelle entweder sehr instabil oder es kommt zu einem Overfitting. In der Literatur findet sich für dieses Problem häufig die Bezeichnung *curse of dimensionality* (deutsch: Fluch der Dimensionalität) [12]. Veranschaulichen lässt sich dies anhand eines Beispiels: Gegeben sei ein Datensatz von zehn Bildern, auf denen jeweils Äpfel oder Birnen abgebildet sind (s. Bild 4). Ziel ist die Entwicklung eines Klassifikationsmodells zur Unterscheidung der beiden Obstsorten. Hierfür wird ein erstes Attribut gewählt. Der offensichtlichste in diesem Fall ist bspw. der durchschnittliche Grünanteil im Farbwert der Bilder (s. Bild 4-a).

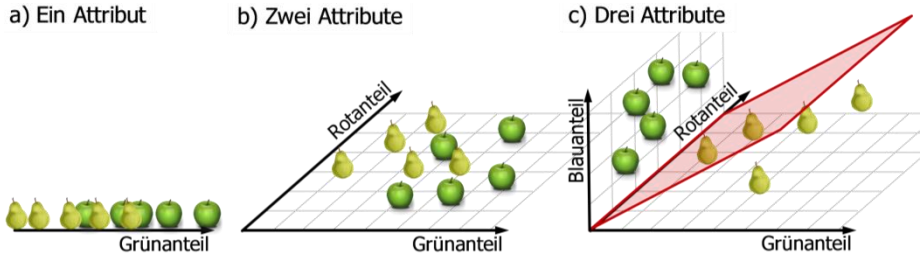

Bild 4: Klassifikation von Äpfeln und Birnen bei der Nutzung einer unterschiedlichen Anzahl von Attributen

In Bild 4-a ist zu sehen, dass sich die einzelnen Instanzen anhand nur eines Attributs nicht eindeutig und fehlerfrei voneinander trennen lassen, es kommt zu Überschneidungen. Wird der Betrachtung nun ein zweites Attribut „durchschnittlicher Rotanteil" hinzugefügt, so erweitert sich der Featureraum um eine Dimension (s. Bild 4-b). Nun sind die Instanzen zwar besser voneinander getrennt als vorher, eine vollständige Trennung von Äpfeln und Birnen ist jedoch nach wie vor nicht möglich. Ebenfalls wird deutlich, dass sich durch die Erhöhung der Dimension die Dichte der Trainingsdaten im Featureraum deutlich verringert hat. Durch die erneute Erweiterung um ein Attribut verteilen sich die Instanzen noch stärker im Featureraum und eine eindeutige Trennung wird nun möglich (Bild 4-c). Der Prozess zur Auswahl der notwendigen Attribute zur Lösung einer Klassifikationsaufgabe wird Featureselektion bezeichnet. Diese ist i. d. R. ein iterativer Vorgang beim Training eines Klassifikationsmodells.

Generell nimmt die Wahrscheinlichkeit für eine perfekte Trennung der Trainingsdaten mit steigender Anzahl an Attributen zu. Gleichzeitig sinkt jedoch die Dichte der Instanzen im Raum mit jeder weiteren Dimension. Das hat zur Folge, dass bei steigender Anzahl von Attributen die Wahrscheinlichkeit sinkt, dass das trainierte Klassifikationsmodell für die Gesamtheit der zu unterscheidenden Klassen gültig ist. Bei der Klassifikation von Zeitreihen entsteht sehr häufig das Problem der zu hohen Dimensionalität, da je nach genutzter Featureextraktionsmethode eine Vielzahl an Attributen generiert wird. Insbesondere im Fall von multivariater Zeitreihen können so mehr Attribute entstehen, als die Menge vorhandener Instanzen in den Zeitreihen. [11, 12]

4.3 Validierung von Klassifikationsmodellen

Die gängigsten Validierungsmethoden für die trainierten Klassifikationsmodelle basieren auf der Konfusionsmatrix (s. Bild 5-a). Sie bildet die Grundlage für die Berechnung von vielen Validierungskennzahlen, welche im Kontext von Data-Mining als Performanz bezeichnet werden. Die Konfusionsmatrix stellt die vorhergesagten Klassen der Testdatensätze den tatsächlichen Klassen gegenüber. Die Testdaten sind Datensätze zur Validierung des Klassifikationsmodells, in welchen die tatsächliche Klasse bekannt ist, wodurch eine Gegenüberstellung möglich ist. In Bild 5 ist eine Konfusionsmatrix für den binären Klassifikationsfall dargestellt.

a) Konfusionsmatrix

		Vorhergesagte Klasse	
		Klasse A (•)	**Klasse B (×)**
Tatsächliche Klasse	**Klasse A (•)**	True-Positives (TP)	False-Negatives (FN)
	Klasse B (×)	False-Positives (FP)	True-Negatives (TN)

b) Visualisierung

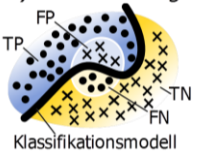

Bild 5: Konfusionsmatrix und entsprechende Visualisierung für den binären Klassifikationsfall nach [11]

Die True-Positives (TP) beschreiben die Anzahl an Datensätzen, welche der Klasse A angehören und auch korrekt vom Klassifikationsmodell als A klassifiziert wurden. Die False-Positives (FP) beschreiben die Anzahl an Datensätzen, welche der Klasse B angehören, jedoch fälschlicherweise als A klassifiziert wurden. Analog dazu beschreiben die True-Negatives (TN) korrekt als B klassifizierte B-Fälle und die False-Negatives (FN) fälschlicherweise als B klassifizierte A-Fälle. [11] Eine Visualisierung hierfür ist in Bild 5-b dargestellt. Im Kontext der Zeitreihenanalyse werden TP als „echte Störungen", FP als

„Fehlalarme", FN als „nicht erkannte Störungen" und TN als „Normalbetrieb" bezeichnet. Zur Quantifizierung der Performanz eines Klassifikationsmodells gibt es verschiedene Kennwerte. Die Performanzkennwerte, die am häufigsten eingesetzt werden, sind im Folgenden aufgeführt: [11]

$$\text{Genauigkeit (engl. accuracy): } \mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{FN}} \tag{4.1}$$

$$\text{Präzision (engl. precision): } \mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.2}$$

$$\text{Trefferquote (engl. recall): } \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.3}$$

$$\text{F1-Wert: } \mathrm{F1} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4.4}$$

Die Genauigkeit Acc errechnet sich aus der Konfusionsmatrix (vgl. Bild 5-a) als das Verhältnis aller richtigen Vorhersagen beider Klassen zu der Gesamtanzahl aller Instanzen. Sie ist in der Praxis zwar weit verbreitet und auf den ersten Blick ein stimmiger Performanzkennwert, hat jedoch Schwächen bei ungleichen Klassenhäufigkeiten in den Trainingsdaten oder bei Fällen, in welchen es von großer Bedeutung ist, dass eine Klasse in allen Fällen richtig vorhergesagt wird [11]. Die Präzision P, Trefferquote R und $\mathrm{F1}$-Wert sind Performanzkennwerte, die sich ebenfalls aus der Konfusionsmatrix errechnen. Im Gegensatz zur Genauigkeit Acc, konzentrieren sich die drei zuletzt genannten Kennwerte auf die positive Klasse. [11] Die Fokussierung auf eine interessante Klasse, für welche möglichst gute Vorhersagen erzielt werden sollen, ist auch im Kontext der Klassifikation des Maschinenzustandes von hoher Relevanz. Aus diesem Grund sind die zuletzt genannten Kennwerte der Genauigkeit Acc vorzuziehen. Dies liegt daran, dass vor allem die Wartungszustände als positive Klasse (TP) ohne teure Fehlalarme (FP) klassifiziert werden müssen und somit der Fokus auf der positiven Klasse liegt.

5 Prozess zur Zustandsdiagnose von Maschinen unter Einsatz von Data-Mining Methoden

In diesem Kapitel wird ein datengetriebener Ansatz zur Klassifikation von Maschinenstörungen vorgestellt. Das Ziel ist die Beurteilung einer Störung nach einer erfolgten Abstellung aufgrund einer kritischen Warnung der Maschinensteuerung. Die Grundlage sind Sensordaten sowohl zu Abstellungen mit Störungen, als auch Maschinenabstellungen, bei welchen im Nachhinein keine Schäden an der Maschine festgestellt wurden (Sicherheitsabstellung im Normalbetrieb). Das wichtigste Kriterium, welcher der entwickelte Prozess zur Charakterisierung der Störung zu erfüllen hat, ist eine geringe Fehlalarmrate.

Fälschlicherweise als Schaden klassifizierte Störungen führen zu Maschinenstillstand und unnötigen Inspektionen und somit zu hohem Kostenaufwand und sind daher unbedingt zu vermeiden. Die Priorität liegt demnach in der Minimierung von Fehlalarmen, wodurch die Wahrscheinlichkeit für einen nicht erkannten Wartungsfall steigt. Der Gesamtprozess in Bild 6 dargestellt.

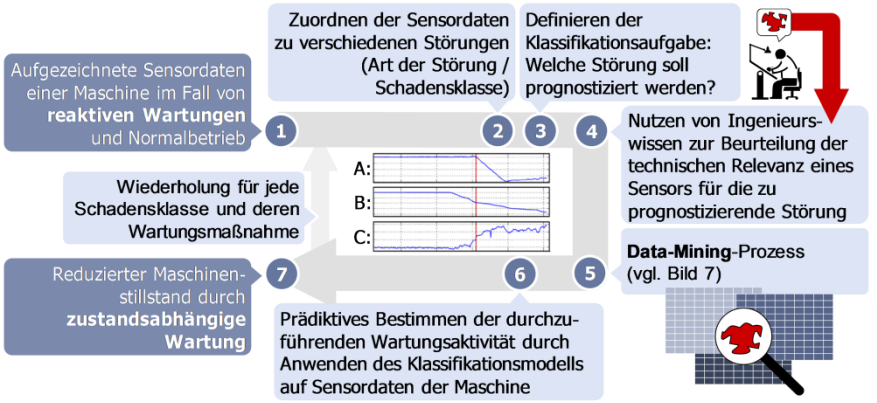

Bild 6: Prozess zur Nutzung von Sensordaten reaktiver Wartungsaktivitäten für die zustandsabhängige Wartung

Der Prozess in Bild 6 beginnt bei den aufgezeichneten Sensordaten einer Maschine im Fall von reaktiver Wartung und im Normalbetrieb. Die verfügbaren Aufzeichnungen der Sensoren müssen zunächst den verschiedenen Schadensklassen zugeordnet werden. Nicht alle verfügbaren Sensoren haben für eine bestimmte Störung eine technische Relevanz, z. B. der Lagerschaden mit den Sensordaten des Abgasstroms. Aus diesem Grund wird mithilfe von Entwicklungsingenieuren eine Vorselektion vorgenommen, um die Dimension des Datensatzes für den Data-Mining-Prozess im Vorfeld zu reduzieren. Danach wird für jede Schadensklasse und deren Wartungsmaßnahme ein Klassifikationsmodell trainiert (s. Bild 7). Dieses Klassifikationsmodell erkennt, unter der Randbedingung der Minimierung von Fehlalarmen, ob ein Schaden vorliegt. Durch das Konsultieren aller trainierten Klassifikationsmodelle findet eine sehr sichere Schadensklassifizierung nach einer Störung statt. Sofern alle Schlussfolgerungen der Klassifikationsmodelle negativ ausfallen, kann von einer Sicherheitsabstellung im Normalbetrieb ausgegangen werden, d. h. der Zustand der Maschine ist in Ordnung, und der Maschinenbetrieb kann ohne das Einleiten einer Wartungsaktivität fortgesetzt werden. Hierdurch kann der Maschinenzustand prädiktiv ermittelt werden, ohne einen Wartungstechniker an den Betriebsort zu schicken.

In Bild 7 ist der Prozess zum Trainieren eines Klassifikationsmodells abgebildet. Dieser beinhaltet die Ansätze, welche in Kapitel 4 vorgestellt wurden. Er startet bei den verfügbaren Sensordaten einer Schadensklasse, welche bereits von Entwicklungsingenieuren hinsichtlich der Schadensklasse auf Relevanz geprüft wurden. In der Datenvorbereitung werden aus den Sensorrohdaten nach einer Datenaufbereitung die Attribute extrahiert (Featureextraktion). Der „Holdout Split" spaltet die vorbereiteten Daten zufällig in einem festen Verhältnis in Trainings- und Validierungsdaten auf. Beim Training des Modells findet die Featureselektion statt. Durch die Validierung des Modells kann die Wahrscheinlichkeit für einen Fehleralarm berechnet werden.

Bild 7: Prozess zum Trainieren eines Klassifikationsmodells

6 Zusammenfassung und Ausblick

In diesem Beitrag wurde ein datengetriebener Ansatz zur Klassifikation von Maschinenfehlerzuständen unter Einsatz von Data-Mining Methoden vorgestellt. Einsatzzweck ist die Beurteilung des Maschinenzustandes nach einer erfolgten Abstellung aufgrund einer kritischen Warnung. Datengrundlage sind Sensordaten (SCADA-Daten) sowohl zu Abstellungen mit vorliegendem Maschinenschaden, als auch zu gewöhnlichen Maschinenabstellungen, bei welchen keine Beschädigungen festgestellt wurde. Anhand des in diesem Beitrag vorgestellten Prozesses zur prädiktiven Zustandsdiagnose können Maschinenstillstandszeiten gegenüber der klassischen, vorbeugenden Wartung reduziert und gleichzeitig die Ausfallzeiten durch Maschinenschäden vermieden werden. Die Maschinen werden hierdurch zuverlässiger, wodurch Unternehmen die Wartungsarbeiten im Zuge des After-Sales-Service im Kontext des *Designs for Maintenance* kostengünstiger gestalten können.

Danksagung

Die Autoren danken der DFG für die Förderung des Teilprojekte B1 „Entwicklung eines selbstlernenden Assistenzsystems" im Rahmen des SFB/TR73 für die Unterstützung bei der hier vorgestellten Arbeit.

Literatur

[1] Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R.: "Multivariate Analysemethoden. Eine anwendungsorientierte Einführung", 14. Auflage, Springer, Berlin, 2016.

[2] Bissantz, N.; Hagedorn, J.: "Data Mining", Business & Information Systems Engineering, Vol. 1 No. 1, 2009, S. 118–122.

[3] Brock, D.: "Aftersales management. Creating a successful aftersales strategy to reduce costs, improve customer service and increase sales", Kogan Page, London, 2009.

[4] Buza, K. A.: "Fusion Methods for Time-Series Classification", Dissertation, Universität Hildesheim, 2011.

[5] Cohen, M.; Agrawal, N.; Agrawal, V.: "Winning in the Aftermarket", Harvard Business Review, Bd. 84 (2006) Nr. 5.

[6] Endig, M. et al.: "Instandhaltung technischer Systeme – Methoden und Werkzeuge zur Gewährleistung eines sicheren und wirtschaftlichen Anlagenbetriebs", Springer Science + Business Media, 2010.

[7] Kolerus, J.; Wassermann, J.: "Zustandsüberwachung von Maschinen. Das Lehr- und Arbeitsbuch für den Praktiker", 6. Auflage, Expert, Renningen, 2014.

[8] Petersohn, H.: "Data Mining: Verfahren, Prozesse, Anwendungsarchitektur", Oldenbourg, 2005.

[9] Scarpa, B.: "Data-Mining", In: Lovric, M. (Hrsg.): "International Encyclopedia of Statistical Science", Springer, 2011, S. 336–339.

[10] Schlittgen, R.; Streitberg, B.: "Zeitreihenanalyse", Oldenbourg, 2001.

[11] Tan, P.-N.: "Introduction to Data Mining", Addison-Wesley, 2006.

[12] Verleysen, M.; François, D.: "The Curse of Dimensionality in Data Mining and Time Series Prediction", International Work-Conference on Artificial Neural Networks, 2005, S. 758–770.