CLUSTERING VON DRUCKLUFTVOLUMENSTRÖMEN ZUR ERKENNUNG VON ANOMALIEN MIT MASCHINELLEM LERNEN

Christian Dierolf[1], Alexander Sauer[2]

[1]Fraunhofer Institut für Produktionstechnik und Automatisierung, Nobelstraße 12, 70569 Stuttgart, Deutschland, +49 711 9703650, christian.dierolf@ipa.fraunhofer.de, www.ipa.fraunhofer.de

[2]Universität Stuttgart, Institut für Energieeffizienz in der Produktion , Nobelstraße 12, 70569 Stuttgart, Deutschland, +49 711 9703600, alexander.sauer@eep.uni-stuttgart.de, www.eep.uni-stuttgart.de

Kurzfassung: Ziel der vorliegenden Arbeit ist es, Volumenströme einer Druckluftanlage mit unüberwachtem maschinellem Lernen zu clustern. Die Untersuchungen zeigen, wie sich die Clusteranalyse eignet, um charakteristische Verläufe zu identifizieren und Anomalien zu erkennen. Zuerst werden datenbasierte Merkmale, wie Minima und Maxima, aus den jeweiligen Tagesmessungen auf 15 Minutenbasis erstellt. Anschließend werden diese auf zwei bzw. drei Hauptkomponenten reduziert. Zum Clustering der Daten werden neben kMeans weitere Algorithmen mit unterschiedlicher Clusteranzahl untersucht. Die Analysen zeigen eine nachvollziehbare Gruppierung der Volumenstromverläufe mit vier Clustern. Weichen Tagesprofile von einem vorgegebenen Toleranzband ab, sind diese im Signalverlauf als Anomalie erkennbar. Die Erhöhung der Anzahl von vier auf neun zu gruppierende Cluster führt bei der Untersuchung dazu, dass Anomalien durch verschiedene Cluster unterschieden werden.

Keywords: Druckluft, maschinelles Lernen, Clustering

1 Druckluftverwendung, -bedarf und -leckage

Druckluft wird in Unternehmen in verschiedenen Bereichen angewendet, dazu zählen [1]:

- Antreiben von Druckluftwerkzeugen und -motoren;
- Transportieren in pneumatischen Förderanlagen;
- Spannen, klemmen und pressen in pneumatischen Haltevorrichtungen und Pressen;
- Spritzen in Spritzgussmaschinen;
- Füllen von Nahrungsmittel;
- Verpacken in Verpackungsmaschinen oder Sperrluft bei Werkzeugmaschinen

Für die Erzeugung von Druckluft werden in Deutschland 3 % des nationalen Gesamtstrombedarfs eingesetzt. In Industriebetrieben ist Druckluft eine teure Energieform mit jährlich verursachenden Kosten von fast 1,5 Mrd. EUR [2] [3] [4]. Druckluftleckagen verursachen in Deutschland bis zu einem Drittel des gesamten Druckluftverbrauchs [5].

Ein erhöhter bzw. im Profil abweichender Druckluftverbrauch kann sich abhängig von dessen Signifikanz deutlich im Tagesprofil ausprägen. Lediglich auf Basis der dargestellten Zeitreihen der Verbrauchsdaten in einem Monitoringsystem fällt es schwer, eine Aussage über ein normales oder anormales Verhalten zu treffen. Es wird in einem „Proof of Concept“ untersucht, wie sich unüberwachtes maschinelles Lernen im Speziellen – Clustering – zur Verbesserung der Anomalieerkennung und damit zur Reduzierung des Druckluftverbrauchs eignet. Clustering beschreibt das Finden von Ähnlichkeitsstrukturen in Daten und deren anschließende Zuordnung zu Gruppen.

2 Warum maschinelles Lernen?

Mit zunehmender Steigerung des Digitalisierungsgrad werden Daten immer entscheidender zur Unterstützung bei Vorhersagen, predictive analytics („Was wird passieren?“) und bei der automatisierten Entscheidungsfindung prescriptive analytics („Wie soll reagiert werden?“). Maschinelles Lernen generiert Wissen aus Erfahrung. Im untersuchten Anwendungsfall aus historischen Messdaten. Methoden des maschinellen Lernens sind unter anderem geeignet, um automatisch Anomalien in großen Datenmengen zu erkennen. Hierbei ist das Lernen aus Beispielen und die Verallgemeinerung nach der Lernphase entscheidend, um darin Muster und Gesetzmäßigkeiten zu erkennen. Es existiert eine Vielzahl von Algorithmen, wobei die Auswahl nach Einsatzzweck und Modellleistung erfolgt. Für die Auswahl der Algorithmen stellen sich drei Kernfragen:

- Welche Messdaten sind wertvoll, um Wissen zu generieren?
- Welche Gesetzmäßigkeit kann aus den vorliegenden Messdaten abgeleitet werden?
- Wie kann schnell und nachvollziehbar eine Handlungsempfehlung ermittelt werden?

Im untersuchten Anwendungsfall liegt der Fokus auf der Erkennung von charakteristischen Verläufen bzw. den Abweichungen in Form von Anomalien. Sie sind Abweichungen von einem erwarteten Verhalten, Muster oder einer Struktur. In gut aufbereiteten Daten können Menschen Abweichungen leichter erkennen, die von der eigenen Erwartung oder dem Verständnis abweichen. Gut aufbereitete Daten sind für Menschen Grafiken mit bspw. Minima, Maxima oder die Änderung von Steigungen. Die menschliche Analysefähigkeit erreicht ihre Grenze, wenn es sich um Datenbanken bzw. im Tabellenformat vorliegenden Maschinen- oder Sensordaten handelt.

Modelle des maschinellen Lernens werden aus historischen Daten trainiert. Die Trainingsart kann in überwachtes und unüberwachtes Lernen unterschieden werden. Beim überwachten Lernen ist die Zuordnung von Merkmalen (X) und Zielgrößen (Y bzw. Label) bekannt, da sie explizit vorgegeben wird. Dafür ist ein Grundverständnis der zu analysierenden Daten durch den Anwender erforderlich. Die aufwendige Zuordnung der Daten kann manuell oder automatisiert bspw. unter Verwendung von Betriebsdaten durchgeführt werden. Das unüberwachte Lernen hingegen charakterisiert sich durch die Verwendung einer unbekannten nicht festgelegten Zuordnung der Daten. [6] Hieraus entsteht für den Anwender kein zusätzlicher Aufwand, da die Zuordnung nicht vorab festgelegt wird. Bei dieser Art des Lernens wird versucht, typische Daten bzw. Datengruppen wie bspw. „Maschine in Leerlauf“ im Signalrauschen zu erkennen.

Untersuchungen an elektrischen Tageslastprofilen haben gezeigt, dass das Clustering für Mittelständische Unternehmen bei hinsichtlich Lastmanagement unterstützen kann. [7] Die Zuordnung der Daten zu den Gruppen, wie bspw. Tage mit typischen Verbrauchsprofilen ist nicht explizit vorzugeben. Der Ansatz eignet sich daher für die Untersuchung, da keine zusätzlichen Informationen, wie die Produktionsauslastung oder die Betriebszustände der Kompressoren verwendet werden.

3 Untersuchte Messungen und angewendete Methoden

Im folgenden Kapitel wird die Datenbasis der vorliegenden Arbeit beschrieben, gefolgt von der Merkmalserstellung und –reduktion. Abschließend werden verschiedene Clustering-Algorithmen untersucht und die optimale Clusteranzahl bestimmt.

3.1 Beschreibung Messdaten

In der Arbeit wird der Gesamtvolumenstrom einer Druckluftanlage analysiert. Bestehend aus sechs Kompressoren mit insgesamt 130 kW Nennleistung und einem Druckluftspeicher. Im Zwei-Schicht-Betrieb des Unternehmens sind im Messzeitraum von fast sechs Monaten Messdaten im Messintervall von 15 Minuten aufgezeichnet worden. Gemessen wurde zentral nach der Druckluftaufbereitung.

3.2 Merkmalserstellung und -reduktion

Zunächst werden die Messdaten in 163 Tagesmessungen unterteil. Für jede Tagesmessung wird ein Merkmalssatz von 78 unterschiedlichen Merkmalen (d) erstellt. Die Merkmale im Zeitbereich (72) sind die Werte aus dem Stundenmittel, Stundenminimum und Stundenmaximum der Stundenwerte. Im Frequenzbereich sind die Merkmale (6) die Minima, Maxima und die Mittelwerte der Amplituden und Frequenzen aus dem Tagesspektrum.

Die Merkmale sind aufgrund des „Curse of Dimensionality“ zu reduzieren, um den aufgespannten Merkmalsraum ausreichend zu befüllen. Für jedes unterschiedliche Merkmal wird daher die Potenzierung der Daten empfohlen (10^d). [8]

Im beschriebenen Anwendungsfall sind die Merkmale nicht manuell, sondern über die Hauptkomponentenanalyse (PCA) reduziert [9]. Anders als beim überwachten maschinellen Lernen kann die Signifikanz jedes Merkmals nicht ausgewiesen werden, weil die Merkmale selbst auf die definierte Anzahl an Hauptkomponenten informationsverdichtet sind. Zudem ist die Reduktion auf zwei bzw. drei Dimensionen hilfreich, um die geclusterten Punktewolken zu visualisieren. Das Prinzip der Hauptkomponentenanalyse von Punktewolken aus zwei Merkmalen zur Reduktion auf eine Hauptkomponente veranschaulicht Abbildung 1. Die Richtungen mit der größten Streuung (Varianz) beinhalten die meisten Informationen. Am Beispiel der blauen Punktewolke enthält Merkmal 2 weniger Information als Merkmal 1.

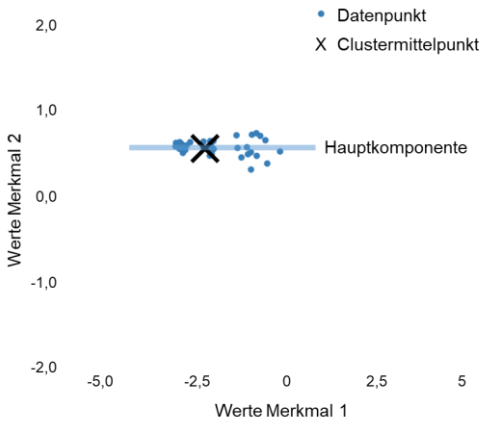

Abbildung 1: Prinzip der Hauptkomponentenanalyse am Beispiel der Reduktion der Merkmale eines Clusters auf eine Hauptkomponente.

3.3 Untersuchte Clustering-Algorithmen und Bestimmung der optimalen Clusteranzahl

Nach der Reduzierung der 78 Merkmale auf deren zwei bzw. drei Hauptkomponenten, gruppiert der Clustering-Algorithmus kMeans diese. Die Hauptkomponenten der Merkmalssätze werden im Algorithmus wie folgt geclustert [10]:

(i.) Zufällige Mittelpunkte mit der Anzahl der definierte Clustern werden gewählt.
(ii.) Jeder Datenpunkt wird dem am nächsten liegenden Cluster-Mittelpunkt zugeordnet.
(iii.) Das geometrische Zentrum der zugeordneten Punkte wird bestimmt und der Mittelpunkt dorthin verschoben.
(iv.) Nach anfänglich zufälliger Wahl der Mittelpunkte werden die Schritte ii.) bis iv.) so lange wiederholt, bis sich die Mittelpunkte unwesentlich verändern.

KMeans liefert im Vergleich zu den anderen untersuchten Algorithmen Gaussian mixture models [11] und Ward hierarchical clustering [12] die aussagekräftigsten Clusterergebnisse nach Kapitel 4. Das zeigt sich in ähnlichen Verläufen des Volumenstroms von Tagesmessungen desselben Clusters.

Die optimale zu findende Clusteranzahl wird durch die Analyse des Silhouettenkoeffizients ermittelt. Der Silhouettenkoeffizient ist ein Maß dafür, wie ähnlich ein Clusterpunkt zu seinem zugeordneten Cluster im Vergleich zu anderen Clustern ist. Ist der Wert nahe 1, impliziert dies die korrekte Zuordnung zum Cluster. Die falsche Zuordnung zum Cluster beschreibt ein Wert nahe -1. Die optimale Anzahl an Clustern ist erreicht, wenn der Mittelwert der Silhouttenkoeffizienten über alle geclusterten Trainingsbeispiele den Maximalwert erreicht. [13]

Eine weitere untersuchte Methode ist die Ellenbogen-Methode. Sie vergleicht die prozentuale Varianz gegenüber der Anzahl an Clustern. In der grafischen Analyse ist die optimale Anzahl diese, bei der die Kurve einen ellenbogenähnlichen Knick aufweist [14]. Sie lieferte in der Untersuchung vergleichbare Ergebnisse basierend auf den verwendeten Daten.

4 Ergebnisse

Die beschriebene Vorgehensweise zur Erkennung von Anomalien in Tageslastprofilen liefert für die Reduktion auf drei Hauptkomponenten mit der optimalen Clusteranzahl von vier die besten Ergebnisse. Abbildung 2 und Abbildung 3 zeigen das Clustering nach Kapitel 3.3 für drei bzw. zwei Hauptkomponenten. Trotz der optisch deutlicheren Trennung der Punktewolken für zwei Hauptkomponenten mit kMeans, lieferte das Clustering für drei Hauptkomponenten aussagekräftigere Ergebnisse gemäß der nachfolgenden Validierung.

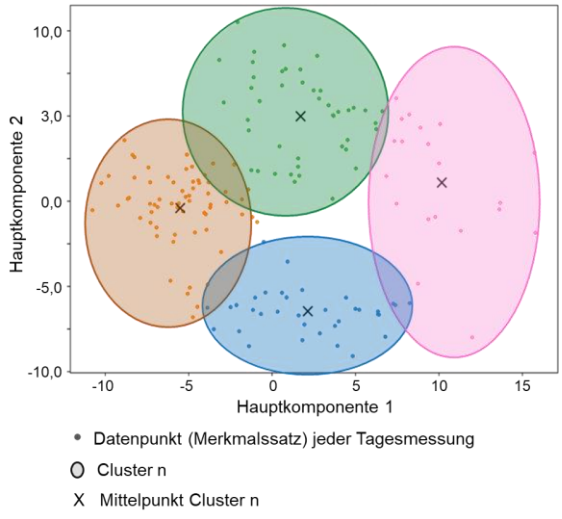

Abbildung 2: KMeans Clustering der zwei Hauptkomponenten der Merkmalsätze für vier Cluster

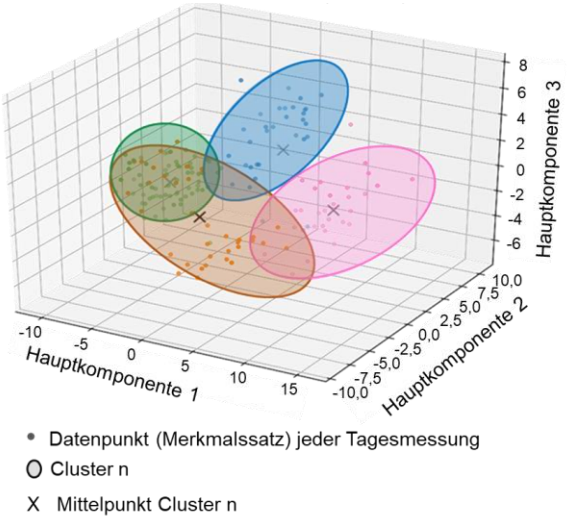

Abbildung 3: KMeans Clustering der drei Hauptkomponenten der Merkmalsätze für vier Cluster

Zur Validierung sind die Tagesmessungen des jeweiligen Clusters nachfolgend als Zeitreihen dargestellt. Die Signalverläufe sind auf den Maximalwert aller Messpunkte normalisiert. Der Median aus allen Signalverläufen eines zugeordneten Clusters ist in Rot dargestellt. Er ist als charakteristischer Volumenstromverlauf des jeweiligen Clusters zu interpretieren. Die Häufigkeit der auftretenden Wochentage ist darunter aufgetragen. Abbildung 4 bis Abbildung 11 zeigen die Ergebnisse der Cluster mit der Häufigkeit des Auftretens für die Tage bzw. Gruppen von Tagen Montag, Dienstag bis Freitag, Samstag und Sonntag.

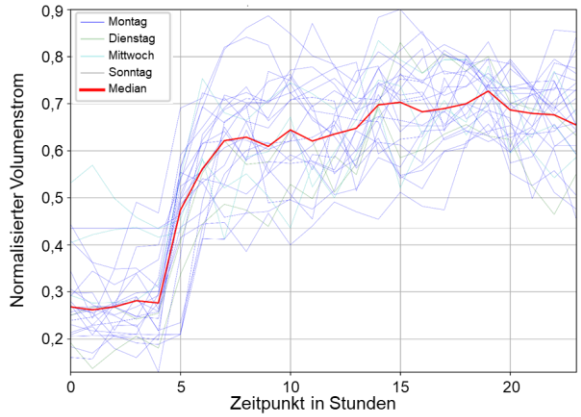

Abbildung 4: Charakteristisches Profil des Volumenstrom für Montag mit Median (rot) mit signifikantem Anstieg ab etwa 4 Uhr.

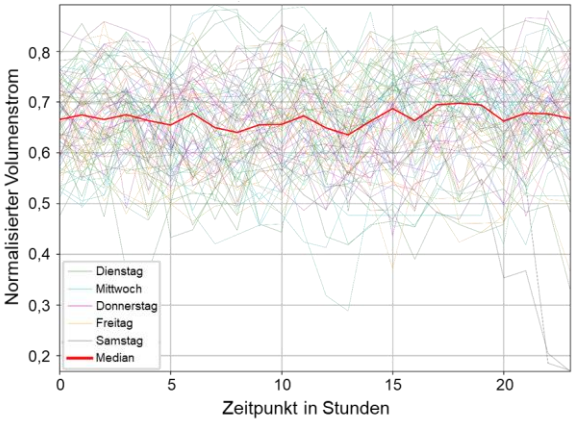

Abbildung 6: Charakteristisches Profil des Volumenstrom für Dienstag bis Freitag mit Median (rot).

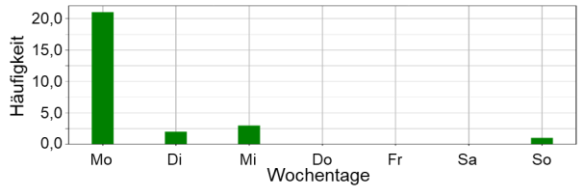

Abbildung 5: Häufigkeit der charakteristischen Profile des Volumenstroms für Montag

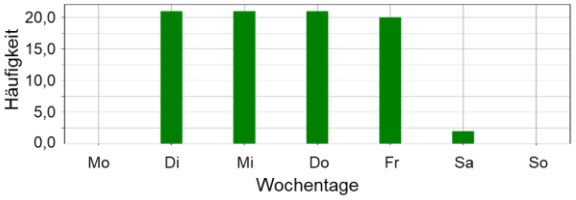

Abbildung 7: Häufigkeit der charakteristischen Profile des Volumenstroms für Dienstag bis Freitag

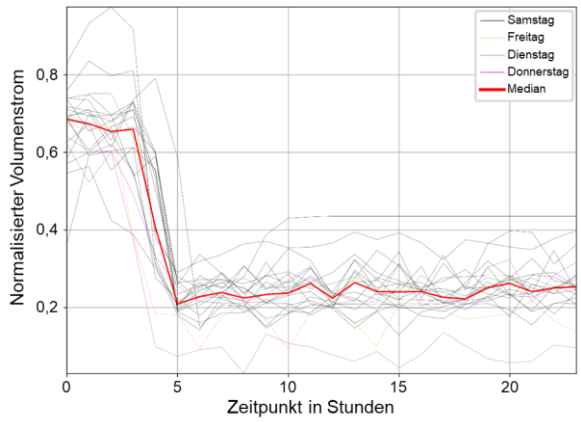

Abbildung 8: Charakteristisches Profil des Volumenstrom für Samstag mit Median (rot) mit signifikantem Abfall ab etwa 4 Uhr.

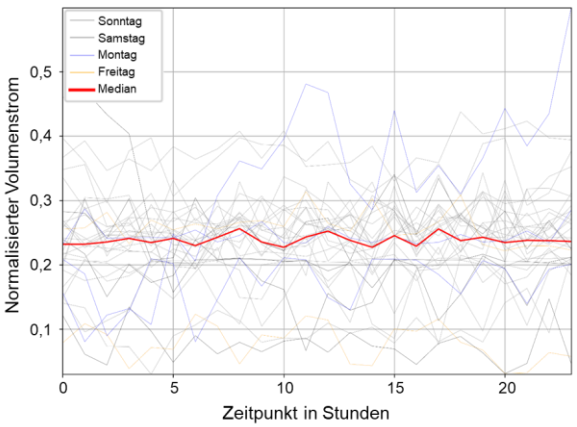

Abbildung 10: Charakteristisches Profil des Volumenstrom für Sonntag mit Median (rot)

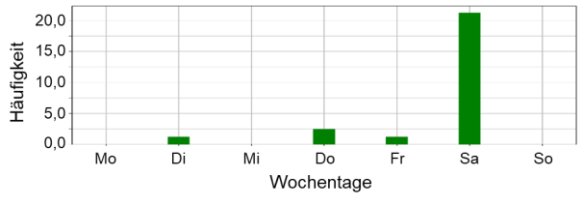

Abbildung 9: Häufigkeit der charakteristischen Profile des Volumenstroms für Samstag

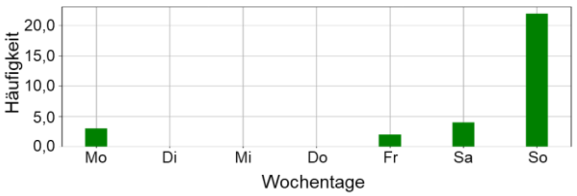

Abbildung 11: Häufigkeit der charakteristischen Profile des Volumenstroms für Sonntag

Mit den beschriebenen Methoden angewendet auf die Messdaten ist es möglich, Anomalien mit unüberwachtem maschinellem Lernen zu clustern und anschließend im Signalverlauf des Volumenstroms tageweise zu interpretieren. Weicht ein Tagesprofil einem vorgegebenen Toleranzband exemplarisch nach Abbildung 12 ab, ist dies im Signalverlauf als Anomalie erkennbar.

Die Erhöhung der Clusteranzahl führt bei der Untersuchung dazu, dass Anomalien in diesem Fall durch neun unterschiedliche Cluster erkannt werden können. Zwei der Cluster gruppieren bspw. Signalverläufe von Samstagen wie in Abbildung 13 so, dass häufig auftretende ähnlich normale und weniger häufig auftretender anormaler Signalverlauf jeweils darin enthalten sind. Wird das trainierte Clustermodell mit neun Clustern verwendet, können Anomalien ohne den Vergleich des Verlaufs der Zeitreihe mit definiertem Toleranzband, direkt mit entsprechenden Cluster für einen anomalen charakteristischen Tagesverlauf identifiziert werden.

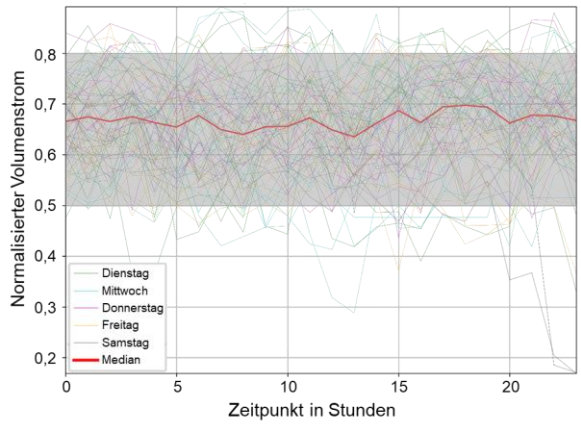

Abbildung 12: Charakteristisches Cluster für Dienstag bis Freitag mit exemplarischen Toleranzband von 0,5 bis 0,8 des normalisierten Volumenstroms (grau)

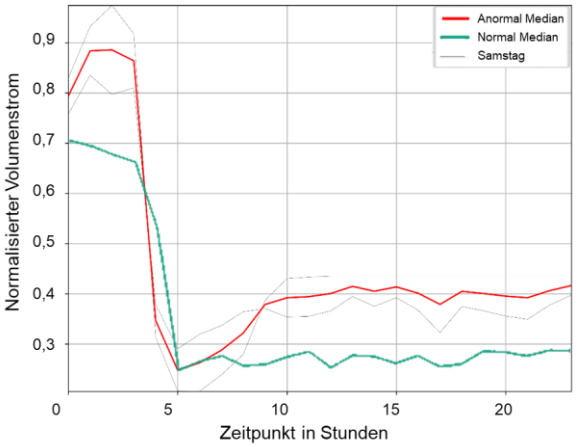

Abbildung 13: Signalverlauf Cluster für Samstag normal (grün) und Signalverlauf Cluster Samstag anormal (rot) als Clustermedian

5 Bewertung und Ausblick

Die Wahl des Toleranzbandes ist nach dem erfolgten Clustering entscheidend. Anomalien sollen hierdurch bei einer geringen Fehlerkennungsrate detektiert werden. Durch die dargestellten Medianverläufe eines jeweiligen Clusters kann das Toleranzband durch den Anwender entsprechend definiert werden. Die Abweichung von der optimalen Clusteranzahl ist nur dann sinnvoll, wenn die geclusterten Verläufe plausible Ergebnisse liefern, wie bspw. zwei unterschiedliche Cluster für Samstag. Die Übertragbarkeit der vorliegenden Arbeit auf Druckluftverbräuche weiterer Unternehmen ist in darauf aufbauenden Forschungsvorhaben zu validieren.

Charakteristische Tagesprofile können zur Kosteneinsparung verwendet werden, indem sie als Bedarfsprognosen für Folgetage, Monitoring von Betriebszeiten für Effizienzmaßnahme zur Grundlastsenkung oder Identifikation wertschöpfender und nicht wertschöpfender Zeiten wie Sonntage eingesetzt werden. Eine detailliertere Unterteilung in Betriebszustände ist möglich, indem Betriebsdaten aus der Kompressor-Steuerung verwendet werden.

Das Clustering kann durch zusätzliche Kontextinformation verbessert werden, indem Untermessungen wie Einzelmessung der Kompressoren installiert, charakteristische Stundenprofile pro Tag eingeführt oder die Messdaten mit der Produktionsplanung oder Produktionsauslastung verknüpft werden. Dies gilt es in nachfolgenden Forschungsvorhaben weiter zu untersuchen.

Danksagung

Diese Arbeit wird vom Ministerium für Umwelt, Klima und Energiewirtschaft Baden-Württemberg im Rahmen des Projekts „Die Ultraeffizienzfabrik – Symbiotisch-verlustfreie Produktion im urbanen Umfeld“ gefördert.
Bei Herrn Reisinger bedanken wir uns für die konstruktiven Anmerkungen zum Manuskript.

Referenzen

[1] R. Gloor, „Energieeinsparungen bei Druckluftanlagen in der Schweiz - Programm Elektrizität Forschungsprojekt," Gloor Engineering, Sufers, 2000.

[2] Bayerisches Landesamt für Umweltschutz, „Effiziente Druckluftsysteme," 2004. [Online]. Verfügbar unter: https://www.bestellen.bayern.de/application/eshop_app000003?SID=580600777&ACTIONxSESSxSHOWPIC(BILDxKEY:%27lfu_klima_00028%27,BILDxCLASS:%27Artikel%27,BILDxTYPE:%27PDF%27). [Zugriff am 28 Januar 2020].

[3] Umweltbundesamt, „Entwicklung des Stromverbrauchs nach Sektoren," 2018. [Online]. Verfügbar unter: https://www.umweltbundesamt.de/sites/default/files/medien/384/bilder/dateien/2_und_4_datentabelle_eev_2018-02-14.pdf. [Zugriff am 22 Januar 2020].

[4] Statista GmbH, „Nettostromverbrauch in Deutschland in den Jahren 1991 bis 2018 (in Terawattstunden). Statista," Februar 2019. [Online]. Verfügbar unter: https://de.statista.com/statistik/daten/studie/164149/umfrage/netto-stromverbrauch-in-deutschland-seit-1999/. [Zugriff am 11 Oktober 2019].

[5] Zentrale Koordinierungsstelle KEFF, „KEFFIZIENZ-LEITFADEN DRUCKLUFT," Zentrale Koordinierungsstelle KEFF, Stuttgart, 2017.

[6] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol: O'Reilly Media, Inc., 2019.

[7] M.-A. Richard, H. Fortin und M. Fournier, „Daily load profiles clustering: a powerful tool for medium-sized industries demand side management," in ACEEE summer study on Energy Efficiency in Industry, Denver, 2017.

[8] M. Verleysen und D. François, „The Curse of Dimensionality in Data Mining and Time," Computational Intelligence, pp. 758-770, 2005.

[9] L. Smith, „A tutorial on Principal Components Analysis," Februar 2002. [Online]. Verfügbar unter: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. [Zugriff am 31 Oktober 2019].

[10] D. Arthur und S. Vassilvitski, „k-means++: The advantages of careful seeding," in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics , 2007.

[11] J. Karter, Statistical Methods With Matlab, CreateSpace Independent Publishing Platform, 2016. ISBN: 978-1-5395-4238-4

[12] F. Murtagh und P. Legendre, „Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm," Journal of Classification, pp. 274-295, 18 Oktober 2014.

[13] L. Kaufman und P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.

[14] R. L. Thorndike, „Who belongs in the family?," in Psychometrika, Springer-Verlag, 1953, pp. 267-276.