
Automatische Klassifikation von Datenpunkten in raumluftechnischen Anlagen

Noah Mertens¹, Andreas Wilde¹

¹Fraunhofer IIS/EAS

Zusammenfassung

Zur vereinfachten, möglichst weitgehend automatisierten Konfiguration von Monitoring-Systemen in Lüftungstechnischen Anlagen soll aus Betriebsdaten der Anlagen automatisiert das Anlagenschema mit den räumlichen und funktionalen Beziehungen der Komponenten und deren semantischen Zusammenhänge abgeleitet werden. Die vorliegende Arbeit befasst sich mit dem Teilproblem der automatisierten Klassifikation von Datenpunkten und schafft dadurch eine Grundlage für weiterführende Arbeiten zur Lösung der übergeordneten Forschungsfrage. Dazu kommen Verfahren aus dem Gebiet der künstlichen Intelligenz zum Einsatz. Der verwendete Datensatz beinhaltet Messwerte aus insgesamt 76 Gebäuden und umfasst die Monate Januar bis November des Jahres 2018. Anhand dieser Daten werden relevante Merkmale der Zeitreihen zur Unterscheidung von Datenpunktklassen definiert und die Merkmalsausprägungen für jede Zeitreihe berechnet. Auf Basis der Merkmale werden verschiedene Klassifikationsmodelle erstellt. Insgesamt liefern die verwendeten Klassifikationsmodelle bereits sehr gute Ergebnisse und ordnen im Schnitt etwa 92% aller Datenpunkte der korrekten Klasse zu. Die Datenpunkte der am schlechtesten erkannten Klasse werden zu rund 88% (Validierungsdaten) bzw. 90% (Testdaten) korrekt klassifiziert. Es stellt sich heraus, dass die am schlechtesten erkannte Klasse, die Stellung von Lüftungsclappen, sehr grob definiert ist und durch eine feinere Aufteilung in Subklassen die Gesamtergebnisse voraussichtlich deutlich verbessert werden könnten. Die erzielten Ergebnisse zeigen, dass sich in der Praxis mit diesen Verfahren sehr viel Arbeit bei der Konfiguration von Monitoringsystemen einsparen lässt.

Keywords: Datenpunkt, Klassifikation, Machine Learning

1 Einführung

Der Wirtschaftsnobelpreisträger William D. Nordhaus beschreibt in „Climate Change: The Ultimate Challenge for Economics“ den Klimawandel als größte ökonomische Herausforderung aller Zeiten [15]. Mit dem 1,5-Grad-Ziel des Pariser Abkommens als Richtlinie müssen ökonomisch sinnvolle Maßnahmen zur Energieeinsparungen erhöht werden. Der Gebäudesektor mit seinem Anteil von 40 % am Primärenergiebedarf steht dabei im besonderen Fokus. Da in Nichtwohngebäuden 90 % der Gebäudeflächen nicht effizient geregelt werden [19], wird der Automation von Heizung, Lüftung und Kühlung ein realistisches Einsparpotential von 13 % des Gebäudeenergiebedarfs bis 2035 zugerechnet [6].

Gebäudetechnische Anlagen liefern viele Daten über den Zustand des Gebäudes und der verbauten Geräte. Oft führen Fehler im Aufbau, der Programmierung oder dem Betrieb der Anlage zu einem erhöhten Energieverbrauch in Gebäuden. Um diese Prozessfehler möglichst schnell und präzise detektieren und beheben zu können, ist es notwendig, ein Monitoringsystem für die Anlagen einzurichten. Bei einer Untersuchung von 87 Softwareanwendungen aus dem Bereich der Gebäudeautomation wurde deutlich, dass für die erfolgreiche Implementierung der Software eine Vielzahl an semantischen Informationen (Metadaten) notwendig ist [2]. Während der Implementierung solcher Softwarelösungen werden die Anlagenkomponenten sowie deren Zusammenhänge in der Software verknüpft. Für kleine Anlagen stellt dies kaum ein Problem dar, doch bei Gebäuden mit komplexen Anlagenstrukturen und vielen Datenpunkten ist dieser Prozess sehr zeitintensiv und fehleranfällig [20, 14]. Der Grund hierfür liegt hauptsächlich an mangelnden Normen und Standards zur Vergabe der Datenpunktnamen. Die verschiedenen Hersteller der technischen Anlagen verfolgen häufig unterschiedliche Systeme zur Benennung des gleichen Bauteils, was eine vollautomatisierte Verknüpfung von Anlagenstruktur und Software bisher unmöglich macht.

Um den Einsatz von Monitoringssystemen und Automation im Gebäudesektor skalierbar auszubauen, stellt sich deshalb die Frage nach einer automatisierten Verknüpfung von Anlagenkomponenten und Software. Dabei soll das Anlagenschema mit den räumlichen und funktionalen Beziehungen der Komponenten und deren semantischen Zusammenhänge bestenfalls vollautomatisch, doch mindestens unter Minimierung menschlicher Interaktion erkannt werden. Diese Arbeit konzentriert sich dabei auf Datenpunkte aus Lüftungsanlagen. Ein wichtiger Zwischenschritt ist dabei die Erkennung von Datenpunktklassen, also eine Einteilung von Datenpunkten in beispielsweise Temperatur, Druck oder Klappenstellung. Diese Klassifizierung kann auf zwei unterschiedlichen Informationsquellen

basieren. Zum einen können die Zeitserien der Datenpunkte verwendet werden, zum anderen steht die Datenpunktbezeichnung zur Verfügung. Bei der Verwendung der Zeitreihen zur Klassifikation stößt man in der Praxis oft auf das Problem, dass historische Daten nur für begrenzte Zeiträume gespeichert zur Verfügung stehen, wobei die zeitliche Auflösung typisch zwischen 1 Minute und 1 Stunde pro Messwert variiert. Im ungünstigsten Fall stehen gar keine historischen Betriebsdaten zur Verfügung und müssen erst gemessen und gespeichert werden.

Das Problem der automatisierten Erkennung von Metadaten wird in den letzten Jahren immer öfter behandelt [20]. In [1] wird das ZODIAC-System vorgestellt, das mit Hilfe von Datenpunktbezeichnern und Merkmalen der zugehörigen Zeitreihen automatisch Metadaten extrahiert. Fütterer et al. [8] stellen ein Verfahren vor, das jedoch nur an einem einzelnen Gebäude getestet wird. Die Samplingrate der dabei verwendeten Datenpunkte ist mit $1/(60\text{ s})$ recht hoch. Für viele Gebäuden stehen Daten mit so hoher zeitlicher Auflösung nicht zur Verfügung. Weiter werden lediglich 7 statistische Standardkennwerte als Merkmale der Zeitreihen verwendet. Bode et al. [3] stellen einen Ansatz zur Klassifikation mit unüberwachten Lernverfahren vor. Die erzielten Ergebnisse bleiben aber deutlich hinter den Verfahren auf Basis überwachten Lernens zurück. Gao et al. [9] und Shi et al. [17] werten mehrere Ansätze zur Klassifikation von Datenpunkten anhand der Zeitreihen aus und vergleichen die jeweiligen Ergebnisse für viele Gebäude. Aufgrund der unterschiedlichen Herangehensweisen ist es jedoch schwierig, für die praktische Anwendung Schlüsse zu ziehen. In einer neuen Studie nutzen Chen et al. [4] nur Merkmale von Zeitreihen, um sowohl Klasse als auch Anlagenzugehörigkeit zu erkennen. Mit den vorgestellten Ansätzen werden Klassifikationsgenauigkeiten von 90% bzw. 85% erreicht, jedoch wird auf die praktischen Aspekte der Anwendung mit typischerweise in der Praxis verfügbaren Daten kaum eingegangen.

Für die Untersuchungen wird von folgendem Anwendungsfall ausgegangen: Von einem unbekanntem Gebäude liegen historische Betriebsdaten vor, wobei der Zeitraum der Datenaufnahme von Fall zu Fall sehr unterschiedlich sein kann. In der Praxis können das im Idealfall ein Zeitraum von einem kompletten Jahr und mehr sein, im negativen Extremfall sind nur einige Tage oder Wochen durch Betriebsdaten abgedeckt. Ein Klassifikator ordnet jedem Datenpunkt eine Datenpunktklasse sowie eine Klassifikations-Fehlerwahrscheinlichkeit zu. Dafür werden Konzept und Implementierung eines geeigneten Klassifikators unter Verwendung von Verfahren des maschinellen Lernens erarbeitet. Für den Klassifikator werden anschließend Einflüsse bestimmter Faktoren wie zeitliche Auflösung und Länge der Zeitreihen, sowie saisonale Effekte eruiert. Abschließend wird der Klassifikator mit den Daten zweier unbekannter Gebäude getestet und das Ergebnis mit Blick auf die praktische Anwendbarkeit, insbesondere des notwendigen Umfangs an Betriebsdaten bewertet.

2 Datensatz und Klassifikationsverfahren

Der hier vorgestellte Ansatz zur Klassifikation von Datenpunkten beruht nur auf den Eigenschaften der Zeitreihen und hat folgenden Ablauf:

1. Zunächst werden die Daten eines jeden Datenpunktes in jeweils 24 Stunden lange Abschnitte (Tages-Zeitreihen) aufgeteilt.
2. Aus den so gebildeten Tages-Zeitreihen werden charakterisierende Merkmale für die Klassifikation extrahiert.
3. Auf Basis der extrahierten Merkmale werden verschiedene Klassifikationsmodelle trainiert.
4. Die beiden vorigen Schritte werden mit verschiedenen Merkmalen und Klassifikationsverfahren wiederholt, bis die beste Kombination gefunden ist.
5. Die Klassifikation eines Datenpunktes erfolgt anschließend in zwei Schritten: Erst wird jede Tages-Zeitreihe des Datenpunktes einzeln klassifiziert. Dann wird die Klasse des Datenpunktes nach dem Mehrheitsprinzip bestimmt.

Das zweistufige Verfahren für die Klassifikation des Datenpunktes unterscheidet die vorliegende Studie von den meisten anderen Ansätzen in diesem Bereich. Die zugrunde liegende Idee ist, dass es für jede Datenpunktklasse typische Tagesverläufe gibt, sich aber an einzelnen Tagen untypische Verläufe ergeben. Durch die tageweise Klassifikation soll erreicht werden, dass untypische Verläufe die Berechnung von Merkmalen nur lokal beeinflussen.

Im folgenden Abschnitt 2.1 wird der Datensatz genauer beschrieben. Anschließend erfolgt in Abschnitt 2.2 eine Erläuterung des Verfahrens zur Klassifikation der Datenpunkte.

2.1 Beschreibung des Datensatzes

Der verwendete Datensatz stammt aus der Mortardata-Datenbank (Modular Open Reproducible Testbed for Analysis & Research)[7], welche im Jahr 2019 veröffentlicht wurde. Dort werden verschiedenste Messwerte und Metadaten von über 100 Gebäuden gesammelt und vereint. Zum Zeitpunkt der Untersuchung beinhaltet die Datenbank lediglich Gebäude aus dem Gebiet um San Francisco, Kalifornien (USA). Die sich daraus ergebenden möglichen Einflüsse der geografischen Lage der Gebäude auf die Ergebnisse der Arbeit werden in Kapitel 3 betrachtet.

Zur Beschreibung der Metadaten wird das Brick-Schema¹ verwendet. Dieses ermöglicht die Erklärung funktionaler Zusammenhänge der Gebäudeanlagen, sowie eine Einteilung in Klassen und Subklassen (hier als Bricktypes bezeichnet). Die in der Datenbank

¹<https://brickschema.org>

eingetragenen Informationen liegen überwiegend mit einer zeitlichen Auflösung von 15 Minuten vor (G. Fierro (UC Berkeley), persönliche Kommunikation, 11.3.2020).

Um bei der Klassifizierung von Datenpunkten saisonale Effekte berücksichtigen zu können, werden Daten eines gesamten Jahres, in diesem Fall des Jahres 2018 verwendet. Die Daten werden weiter in mehrere, repräsentative Abschnitte aufgeteilt. Im Bereich der Gebäudetechnik sind häufig typische Tagesverläufe zu beobachten, daher wurde die Länge der einzelnen Zeitreihen auf 24 Stunden (Tages-Zeitreihen) festgelegt. Ein Datenpunkt liefert also in einem Jahr 365 einzelne Tages-Zeitreihen von Messwerten.

Im gleichen Schritt wie die Aufteilung in Tages-Zeitreihen erfolgt eine Filtrierung der Daten. Die Tages-Zeitreihen, die über 24 Stunden weniger als 96 Messwerte liefern, werden als informationsarm verworfen. Das entspricht einem Messwert pro 15 Minuten und stellt einen weit verbreiteten Standard in der Gebäudetechnik dar.

In Abbildung 4 sind die Anzahl an Datenpunkten und Tages-Zeitreihen pro Klasse, sowie die untersuchten Klassen als Balkendiagramm dargestellt. Sie stammen aus insgesamt 76 Gebäuden. Es ist zu erkennen, dass für die einzelnen Klassen unterschiedlich viele Datenpunkte und Tages-Zeitreihen vorliegen. Bei der Betrachtung der Anzahl an Tages-Zeitreihen ist auffällig, dass sich die Klassengrößen stark voneinander unterscheiden. Diese Tatsache wirft ein praktisches Problem auf: Um das Training und die anschließende Evaluierung des Verfahrens durchzuführen, ist es notwendig, dass entsprechende Datenpunktklassen in allen Gebäuden im Datensatz vorhanden sind. Das trifft nur auf eine relativ kleine Auswahl von Datenpunktklassen zu und führte zu der Auswahl der 9 Klassen von Datenpunkten in Abbildung 4.

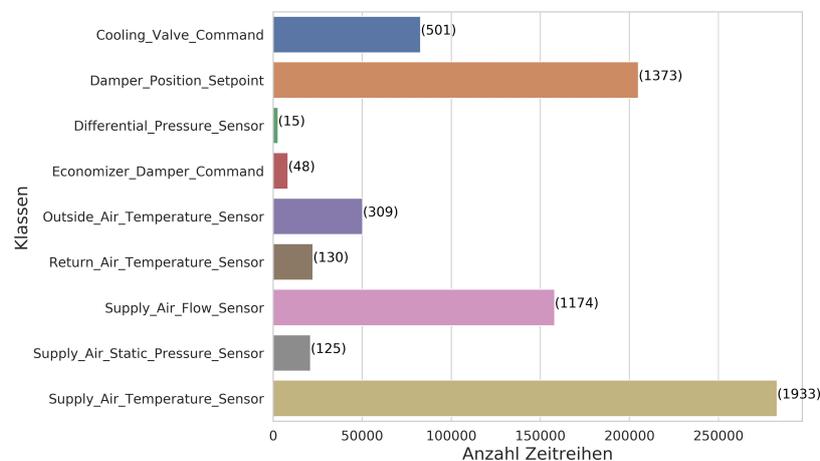


Abbildung 1: Anzahl an Tages-Zeitreihen pro Klasse (Anzahl an Datenpunkten in Klammern)

Weiter wird der Datensatz in Trainings-, Validierungs- und Testdaten aufgeteilt. Mit den Trainingsdaten werden Modelle trainiert und die Validierungsdaten sind notwendig zur Bewertung der Modelle hinsichtlich einer Optimierung von Merkmalen und Parametern, jedoch nicht zur abschließenden Bewertung der Modellgüte. Aufgrund der unterschiedlichen Anzahl von Tages-Zeitreihen pro Datenpunkt wird das Prinzip der Hold-Out-Validierung verfolgt. Dafür wird zu Beginn der Untersuchung ein Teil des Datensatzes für die Validierung ausgewählt und im weiteren Verlauf nicht verändert. Die Testdaten werden ebenfalls zu Beginn der Untersuchungen aus dem Datensatz extrahiert und kommen erst zur finalen Beurteilung eines ausgewählten Modelles zum Einsatz. Eine Aufteilung basierend auf den Zeitreihen hätte zur Folge, dass bereits beim Training der Modelle Informationen aller Gebäude bereitgestellt werden. Die Aufteilung des Datensatzes erfolgt stattdessen anhand von Gebäuden, da die Klassifizierung von Datenpunkten in fremden Gebäuden möglichst realitätsnah untersucht werden soll. Die Validierungsdaten stammen aus 12 Gebäuden, die insgesamt alle Klassen beinhalten. Die Testdaten bestehen aus 2 Gebäuden, die jeweils alle Klassen beinhalten.

Der Datensatz ist bezüglich der Anzahl der Tages-Zeitreihen je Klasse stark unausgeglich. Beim Trainieren von Modellen mit solchen Datensätzen kann es passieren, dass die Eigenschaften der großen Klassen größere Beachtung finden, als die der kleineren. Aus diesem Grund werden bei den Trainingsdaten durch Undersampling alle Klassen auf die Größe der zweitkleinsten Klasse reduziert. Dies ist ein Kompromiss, um nicht zu viele Informationen zu verlieren und gleichzeitig die Klassengrößen auszugleichen. Um einen Verlust von Informationen der saisonalen Effekte zu vermeiden, geschieht dies quartalsweise. Die Klassengrößen der Validierungs- als auch Testdaten bleiben unverändert.

2.2 Beschreibung des Klassifikationsverfahrens

Für die Klassifikation der Tages-Zeitreihen müssen Merkmale aus den Tages-Zeitreihen berechnet werden. Neben naheliegenden Merkmalen der beschreibenden Statistik wie Mittelwert, Extremwerte, Quantile oder Varianz bieten sich auch Merkmale an, welche auf Mustern und Formen der Tages-Zeitreihen basieren. Dazu zählen unter anderem Koeffizienten der Fouriertransformation (FFT-Koeffizienten). Es ist jedoch nicht apriori bekannt, welche Merkmale bezüglich der Klassifikationsaufgabe relevant sind. Weiter existieren eine Reihe von Algorithmen zur Klassifikation, aber auch hier ist nicht apriori bekannt, welches Verfahren die besten

Ergebnisse liefert. Um die beste Kombination aus Merkmalen und Klassifikationsverfahren zu finden, wurde wie folgt vorgegangen (siehe Abbildung 2): Zunächst wird anhand der Literatur eine erste Auswahl von für diese Aufgabe üblichen Merkmalen getroffen (L0).

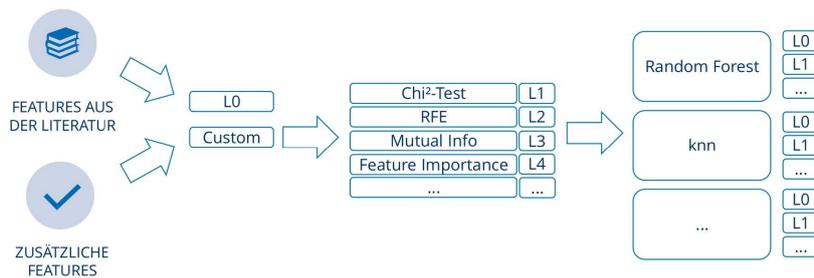


Abbildung 2: Schema des Prozesses zur Merkmal- und Modellauswahl

Anschließend wird die Liste L0 um weitere Merkmale aus der Bibliothek *TSFresh* erweitert, woraus die Liste Custom mit insgesamt 122 Merkmalen entsteht. Die Berechnung der Merkmale aus den Tages-Zeitreihen erfolgt dabei mit der Python-Bibliothek *TSFresh* (0.16.0) [5] und den dort hinterlegten Default-Einstellungen.

Mittels verschiedener Verfahren zur Bestimmung der Merkmalsrelevanz (Tabelle 1) werden jeweils Auswahl-Listen mit potentiell relevanten Merkmalen erstellt. Mit jeder dieser Auswahl-Listen werden anschließend wiederum alle Modelle aus Tabelle 2 trainiert und getestet. Dabei werden die Standardwerte der Hyperparameter aus den Python-Bibliotheken *Scikit-Learn*, bzw. *XGBoost* und *LightGBM* verwendet. Als Maß für die Performance der Kombination von Merkmalsauswahl und Klassifikationsalgorithmus wird der gewichtete F1-Score verwendet. Der F1-Score entspricht dem harmonischen Mittelwert von Precision und Recall [18]. Das Ergebnis ist eine Matrix von F1-Scores.

Tabelle 1: Verfahren zur Ermittlung relevanter Merkmale

Verfahren	Anzahl an Merkmalen	Kodierung
Merkmale in der Literatur	60	L0
Erweiterte Liste	122	Custom
Chi2 Unabhängigkeitstest	60	L1
Feature Importance RFE Default Random Forest	69	L2
Mutual Information	60	L3
Feature Importance Default Random Forest	58	L4
TSFresh Feature Selection	121	L5
Schnittmenge von Merkmale der Verfahren	23	L6
Merkmale von mindestens 4 Verfahren ausgewählt	45	L7

Tabelle 2: Zur Untersuchung der Merkmale verwendete Modelle (Überwachtes Lernen)

Modell	Verwendung in der Literatur
Linear Discriminant Analysis (LDA)	[9, 10]
k-Nearest Neighbours (kNN)	[9, 10, 8]
Random Forest	[9, 10, 8, 11, 12, 13, 16]
Gradient Boost	
ExtraTrees	
LightGBM	
XGBoost	

Für die weitere Optimierung werden aufgrund der erreichten F1-Werte die Modelle *ExtraTrees* (F1=0.87), *XGBoost* (F1=0.89) und *LightGBM* (F1=0.89) mit den Merkmalen der Liste L7 verwendet. Die Liste enthält insgesamt 45 Merkmale, die jeweils durch mindestens 4 verschiedene Feature Selection Verfahren (z.B. Mutual Information, Recursive Feature Elimination, Random Forest Feature Importance, etc.) als relevant erkannt wurden. Für die ausgewählten Merkmal/Klassifikator-Kombinationen wurde anschließend eine Hyperparameter-Optimierung durchgeführt. Alle weiteren Auswertungen wurden mit den so optimierten Modellen durchgeführt.

3 Ergebnisse

3.1 Einfluss saisonaler Effekte

Im Idealfall wird ein Klassifikator mit Daten eines gesamten Jahres trainiert und ermöglicht dadurch die Klassifikation von Datenpunkten unbekannter Gebäude unabhängig vom Aufnahme datum der Tages-Zeitreihen. Bei der Betrachtung des Einflusses saisonaler Effekte wird untersucht, ob die Performance eines Modells abhängig vom Datum der Tages-Zeitreihe ist. Dazu wird ein Modell mit den Trainingsdaten des gesamten Jahres trainiert und für jeden Monat mit Validierungsdaten evaluiert. Für diese Untersuchung wurde ein Extra Trees Modell verwendet, da es deutlich kürzere Trainingszeiten aufweist und annähernd gleiche Ergebnisse wie das XGBoost Modell erreicht. Die gewichteten F1-Scores der einzelnen Monate sind in Abbildung 3 als Balkendiagramm dargestellt. Die rote horizontale Linie stellt den Mittelwert aller berechneten F1-Scores dar. Da für den Monat Dezember weder Trainings noch Evaluierungsdaten vorliegen, kann dafür kein Ergebnis dargestellt werden. Im unteren Teil der Abbildung ist die Verteilung der Anzahl an Tages-Zeitreihen über die Monate aufgetragen.

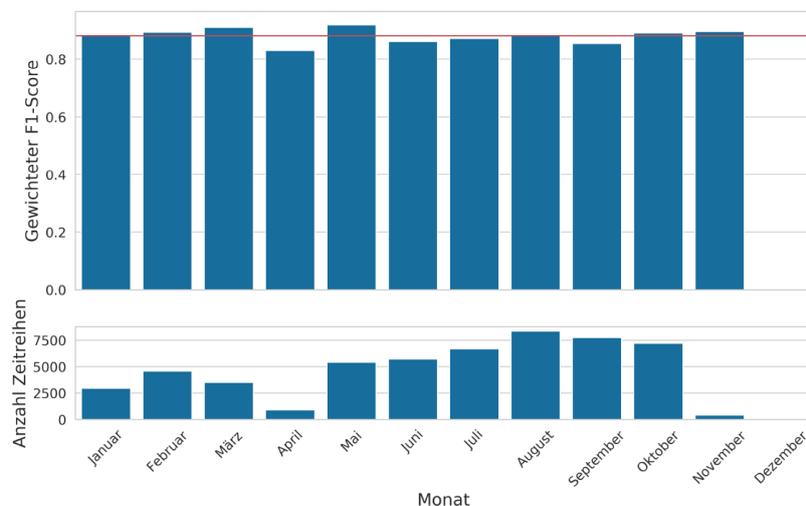


Abbildung 3: Klassifikator trainiert mit Daten des ganzen Jahres, Klassifikation mit Daten von jeweils einem Monat (Extra Trees Modell mit L7 Features)

Aus der Abbildung geht hervor, dass die Performance bei der Klassifikation weitestgehend unabhängig vom Datum der Tages-Zeitreihen ist. Lediglich im April ist eine deutliche Abweichung vom Mittelwert erkennbar. Ein möglicher Grund hierfür ist, dass für den Monat April deutlich weniger Trainingsdaten zur Verfügung stehen. Auffällig ist, dass im November ebenfalls sehr wenige Tages-Zeitreihen vorliegen, doch die Performance in diesem Monat nur unwesentlich vom Mittelwert abweicht. Die Anzahl an Trainingsdaten scheint demnach nicht der einzige Einflussfaktor zu sein und es ist möglich, dass saisonale Effekte eine wichtige Rolle spielen.

Das Klima in San Francisco, Kalifornien unterliegt besonders in den Übergangsmonaten großen Schwankungen. Sind in diesen Monaten nur wenige Tages-Zeitreihen vorhanden, wurden Informationen über Schwankung der Prozessgrößen nur unzureichend aufgezeichnet. Als Resultat kann der Klassifikator weniger Informationen aus diesem Zeitraum lernen. In Bezug auf die Flexibilität der Anwendung des Klassifikators für unbekannte Gebäude spielt der Jahresverlauf des Klimas eine wichtige Rolle. Für den vorliegenden Datensatz von Gebäuden aus dem Bereich um San Francisco nimmt der saisonale Temperaturverlauf eine untergeordnete Rolle ein. Im Vergleich dazu können in Deutschland deutlich stärker ausgeprägte Jahreszeiten beobachtet werden. Für eine Anwendung des Klassifikators in Deutschland und anderen mitteleuropäischen Ländern muss demnach zunächst der Einfluss der Jahreszeiten auf die Performance mit einem entsprechenden Datensatz überprüft werden.

Zusammenfassend kann festgestellt werden, dass beim Training mit Daten eines gesamten Jahres für den vorliegenden Datensatz kaum Unterschiede bei der Performance des Klassifikators in Abhängigkeit des Aufnahme datums von unbekanntem Tages-Zeitreihen vorhanden sind. Die Anzahl an Trainingsdaten scheint in Kombination mit saisonalen Effekten einen Einfluss zu haben, wenngleich dieser für den beobachteten Fall gering ausfällt. Für den Einsatz des Modells in Klimazonen mit stärker ausgeprägten Jahreszeiten muss der Einfluss des Klimas mittels eines anderen Datensatzes umfangreicher evaluiert werden.

3.2 Tages-Zeitreihe zu Datenpunkt

In diesem Abschnitt werden die Ergebnisse des Klassifikators mit Testdaten erläutert. Diese Daten waren dem Modell bisher gänzlich unbekannt und entsprechen damit denen eines realen Anwendungsfalles. Dazu wird das Modell mit einer Kombination von Trainings-

und Validierungsdaten trainiert. Wie in Abschnitt 2.2 beschrieben werden zunächst die Tages-Zeitreihen aller Datenpunkte klassifiziert und aus den Ergebnissen anschließend die Datenpunktklasse ermittelt. In Abbildung 4 (links) sind vergleichend die gewichteten F1-Scores für die Klassifikation von Tages-Zeitreihen und Datenpunkten als Balkendiagramm dargestellt. Die Klassifikation der Tageszeitreihen erreicht bereits einen F1-Score von über 0,925. Durch den zweiten Schritt der Datenpunktklassifikation, die Auswertung des Merheitsprinzips, wird das Ergebnis noch einmal deutlich verbessert und erreicht einen gewichteten F1-Score von etwa 0,975. Es stellt sich die Frage, wie viele Tages-Zeitreihen mindestens pro Datenpunkt vorliegen müssen, um eine verlässliche Aussage über die Datenpunktklasse machen zu können. Diese Frage wird unter Anwendung der Binomialverteilung beantwortet. Damit wird die Wahrscheinlichkeit $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ berechnet, dass von n Tages-Zeitreihen genau k korrekt klassifiziert werden, wobei p der Wahrscheinlichkeit entspricht, dass eine einzelne Tages-Zeitreihe korrekt zugeordnet wird. p ist für verschiedene Datenpunktklassen unterschiedlich. Für eine konservative Abschätzung wird der Wert von p genommen, der sich bei der auf Test- und Validierungsdaten am schlechtesten zu klassifizierenden Datenpunktklasse ergibt. Für diesen Ansatz wird angenommen, dass Klassifikationsfehler zufällig auftreten und die Einzelversuche stochastisch unabhängig voneinander sind. Das Diagramm in Abbildung 4 (rechts) zeigt, dass mit 13 Tages-Zeitreihen pro Datenpunkt die Gesamtwahrscheinlichkeit die korrekten Datenpunktklasse zu bestimmen bereits größer als 95% ist.

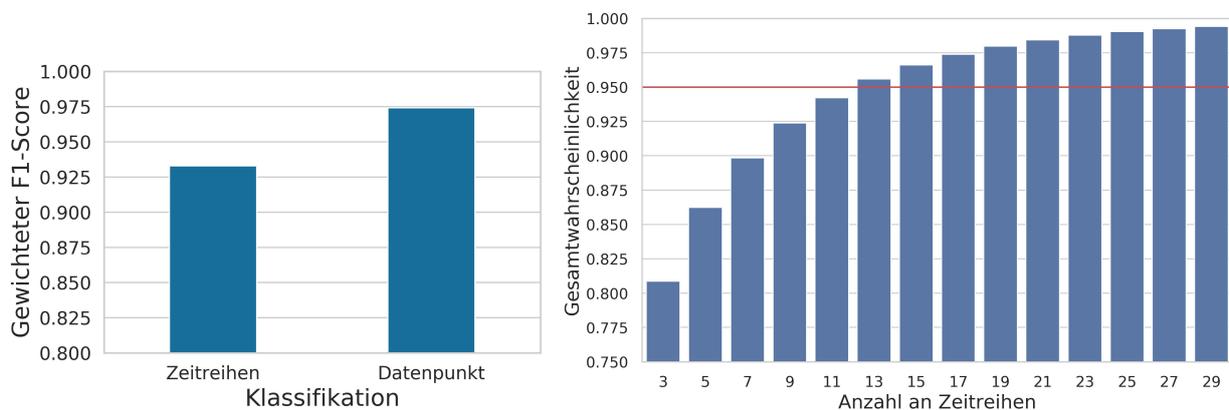


Abbildung 4: Links: Vergleich Klassifikation von Tages-Zeitreihen und Datenpunkten (XGBoost mit L7 Merkmalen, Testdaten), rechts: Gesamtwahrscheinlichkeit über der Anzahl an Tages-Zeitreihen

Eine genauere Betrachtung der Ergebnisse ermöglicht die detailliertere Bewertung des Klassifikators. Für die Verwendung des Modells mit den Daten unbekannter Gebäude ist es nicht nur wichtig, die vorhergesagte Klasse zu kennen, sondern auch die Wahrscheinlichkeit, mit der die Vorhersage korrekt ist. Dadurch kann die Klassenzuordnung für einige Klassen vollautomatisch stattfinden und der Einsatz von Expertenwissen auf einzelne Klassen konzentriert werden. Die Vorhersagewahrscheinlichkeit lässt sich direkt aus der Hauptdiagonalen der Klassifikationsmatrix in Abbildung 5 ablesen, da die Ergebnisse darin normalisiert dargestellt sind und dadurch unabhängig von der Anzahl an Datenpunkten interpretierbar sind. Sie drücken das Verhältnis korrekt zugeordneter Datenpunkte einer Klasse zu allen Datenpunkten dieser Klasse aus. Dabei wird die Annahme getroffen, dass die Anzahl klassifizierter Datenpunkte ausreichend groß ist. Über die Hälfte der Klassen werden vollständig korrekt erkannt. Lediglich die drei Klassen Dampfer_Position_Setpoint, Return_Air_Temperature und Supply_Air_Flow Sensor weisen Fehler auf.

4 Diskussion und Schlussfolgerungen

Für die betrachteten Datenpunkte konnte insgesamt eine sehr gute Erkennung erreicht werden, wenn jeweils genügend Tages-Zeitreihen vorliegen. Die Erkennungssicherheit nimmt mit der Anzahl der vorliegenden Tageszeihen zu, wobei im Mittel eine Klassifikationsgenauigkeit von 95% überschritten wird, wenn mehr als 12 Tages-Zeitreihen vorliegen. Damit sollte in der Praxis in sehr vielen Fällen bereits ein drastische Reduktion des Experten-Aufwandes zu erreichen sein. Dieses Erkennungsergebnis ist zumindest für die untersuchte Klimazone (San Francisco / USA) bezüglich der Jahreszeit, aus der die Tages-Zeitreihen stammen, stabil. Die Übertragbarkeit dieses Befundes auf andere Klimazonen muss noch geprüft werden. Es ist zu erwarten, dass in Klimazonen mit ausgeprägteren Jahrgängen der Temperatur erhebliche Probleme auftreten: Heizungsanlagen sind z.B. in Deutschland im Sommer wochenlang nicht in Betrieb, ebenso die Kühlung im Winter. In den Fällen, in denen die Systeme nicht in Betrieb sind, können nur sehr wenige Informationen entnommen werden. Insofern wird man für ein vollständiges Bild nicht auf die Nutzung von Tages-Zeitreihen aus verschiedenen Jahreszeiten verzichten können. Als grober Richtwert bleibt jedoch die Aussage, dass etwa 2 Wochen an Betriebsdaten die Datenpunkte der jeweils aktiven Systeme gut automatisch identifizierbar machen (vgl. Abbildung 4).

Im Laufe der Studie tauchten jedoch verschiedene methodische Probleme auf, die in Zukunft durch entsprechende Anpassung des Verfahrens noch behoben werden können. Es stellte sich heraus, dass die Klassifikationsergebnisse für einzelne Klassen wie z.B. „Dampfer_Position_Setpoint“ (Klappenstellung) vergleichsweise schlecht war. Die Ursache dafür ist, dass diese Datenpunktklasse

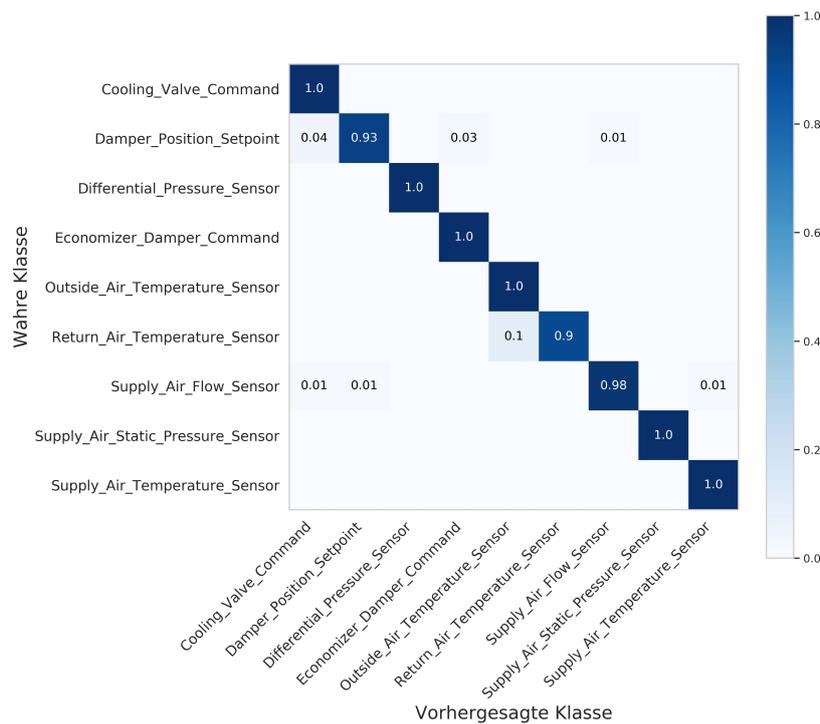


Abbildung 5: Klassifikationsmatrix (XGBoost mit L7 Merkmalen, Testdaten). Die Anzahlen wurden mit der Gesamtzahl der Datenpunkte in der jeweiligen Klasse normalisiert.

ungenau definiert ist. Klappen können an sehr unterschiedlichen Stellen in einer Lüftungsanlage eingesetzt werden und je nach entsprechender Funktion sehr unterschiedliche Betriebscharakteristiken aufweisen. Die vergleichsweise schlechte Erkennungsleistung weist hier also eher auf eine unscharfe Taxonomie der Datenpunkte hin als auf z.B. Probleme mit der Datenverfügbarkeit, der Merkmalsauswahl oder der Klassifikationsverfahren.

Auch einige Datenpunkte der Klasse Return_Air_Temperature werden nicht korrekt erkannt. Genauere Untersuchungen zeigten, dass die Tages-Zeitreihen der inkorrekt klassifizierte Datenpunkte einen deutlich größeren Wertebereich aufweisen als die Trainingsdaten. Eine mögliche Ursache dafür ist das unterschiedliche Nutzungsprofil eines einzelnen Raumes im vorhandenen Datensatz. Es ist zu erwarten, dass weitere Gebäude mit ähnlichen Nutzungsprofilen in den Trainingsdaten die Klassifikationsergebnisse verbessern können.

Die hier erzielten Ergebnisse zeigen, dass die Klassifikation von Datenpunkten anhand von Merkmalen der Zeitreihen einen praxistauglichen Weg darstellt, um den Experten-Aufwand bei der Konfiguration von Monitoring-Systemen in Lüftungstechnischen Anlagen drastisch zu reduzieren. In dieser Arbeit werden jedoch noch eine Reihe weiterer Aspekte wie z.B. die automatische Erkennung funktionaler Zusammenhänge und der Zugehörigkeit zu verschiedenen Anlagenteilen nicht betrachtet. Für diese Aufgaben werden weitere Informationsquellen wie Datenpunktbezeichner oder grafische Darstellungen der Anlagenstrukturen eine große Rolle spielen. Da diese Informationsquellen jedoch oft unvollständig und fehlerbehaftet sind, bietet die Betrachtung der Zeitreihen wertvolle Anhaltspunkte zur Plausibilisierung von Informationen aus anderen Quellen.

Danksagung

Gefördert durch:



aufgrund eines Beschlusses des Deutschen Bundestages unter dem Förderkennzeichen 03ET1567A.

Literatur

- [1] Bharathan Balaji u. a. „Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for Buildings“. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15. the 2nd ACM International Conference. Seoul, South Korea: ACM Press, 2015, S. 13–22.
 - [2] Arka Bhattacharya, Joern Ploennigs und David Culler. „Short Paper: Analyzing Metadata Schemas for Buildings: The Good, the Bad, and the Ugly“. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15. the 2nd ACM International Conference. Seoul, South Korea: ACM Press, 2015, S. 33–34.
 - [3] Gerrit Bode u. a. „A time series clustering approach for Building Automation and Control Systems“. In: Applied Energy 238 (März 2019), S. 1337–1345.
 - [4] Long Chen u. a. „A Metadata Inference Method for Building Automation Systems With Limited Semantic Information“. In: 17.4 (Okt. 2020), S. 2107–2119.
 - [5] Maximilian Christ u. a. „Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)“. In: Neurocomputing 307 (Sep. 2018), S. 72–77.
 - [6] Diedert Debusscher und Paul Waide. „A timely opportunity to grasp the vast potential of energy savings of Building Automation and Control Technologies“. In: (), S. 21.
 - [7] Gabe Fierro u. a. „Mortar: An Open Testbed for Portable Building Analytics“. In: ACM Transactions on Sensor Networks 7 (2019), S. 10.
 - [8] Johannes Fütterer, Maksymilian Kochanski und Dirk Müller. „Application of selected supervised learning methods for time series classification in Building Automation and Control Systems“. In: Energy Procedia 122 (Sep. 2017), S. 943–948.
 - [9] Jingkun Gao und Mario Bergés. „A large-scale evaluation of automated metadata inference approaches on sensors from air handling units“. In: Advanced Engineering Informatics 37 (Aug. 2018), S. 14–30.
 - [10] Jingkun Gao, Joern Ploennigs und Mario Berges. „A Data-driven Meta-data Inference Framework for Building Automation Systems“. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15. the 2nd ACM International Conference. Seoul, South Korea: ACM Press, 2015, S. 23–32.
 - [11] Dezhi Hong u. a. „Sensor-Type Classification in Buildings“. In: arXiv:1509.00498 (1. Sep. 2015).
 - [12] Dezhi Hong u. a. „The Building Adapter: Towards Quickly Applying Building Analytics at Scale“. In: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments - BuildSys '15. the 2nd ACM International Conference. Seoul, South Korea: ACM Press, 2015, S. 123–132.
 - [13] Jason Koh u. a. „Quiver: Using Control Perturbations to Increase the Observability of Sensor Data in Smart Buildings“. In: arXiv:1601.07260 (26. Jan. 2016).
 - [14] Jason Koh u. a. „Scrabble: transferrable semi-automated semantic metadata normalization using intermediate representation“. In: Proceedings of the 5th Conference on Systems for Built Environments - BuildSys '18. the 5th Conference. Shenzhen, China: ACM Press, 2018, S. 11–20.
 - [15] William Nordhaus. „Climate Change: The Ultimate Challenge for Economics“. In: American Economic Review 109.6 (1. Juni 2019), S. 1991–2014.
 - [16] June Young Park, Bertrand Lasternas und Azizan Aziz. „Data-Driven Framework to Find the Physical Association between AHU and VAV Terminal Unit – Pilot Study“. In: (), S. 9.
 - [17] Zixiao Shi u. a. „Evaluation of Clustering and Time Series Features for Point Type Inference in Smart Building Retrofit“. In: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. BuildSys '19: The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. New York NY USA: ACM, 13. Nov. 2019, S. 111–120. (Besucht am 03.04.2020).
 - [18] Wissam Sibli u. a. „Master Your Metrics with Calibration“. In: Advances in Intelligent Data Analysis XVIII. Hrsg. von Michael R. Berthold, Ad Feelders und Georg Kreml. Cham: Springer International Publishing, 2020, S. 457–469. ISBN: 978-3-030-44584-3.
 - [19] Paul Waide u. a. „The scope for energy and CO2 savings in the EU through the use of building automation technology“. In: Final Report for the European Copper Institute (2013).
 - [20] Weimin Wang u. a. „Automated point mapping for building control systems: Recent advances and future research needs“. In: Automation in Construction 85 (Jan. 2018), S. 107–123.
-