Das Data-Leakage-Problem für die prädiktive Instandhaltung

Martin Patrick Pauli, Martin Golz

Hochschule Schmalkalden, Fakultät Informatik, Blechhammer 4, 98573 Schmalkalden

Abstract

Für eine vibroakustische Zustandsüberwachung von kritischen Filterstufen einer Anlage zur Aufbereitung von Reinstwasser wurden im Rahmen einer Pilotstudie aus den vibroakustischen Sensordaten Merkmalsvektoren extrahiert und anschließend mit einer Gradient-Boosting-Methode analysiert, wie gut sich kritische Filterzustände von intakten diskriminieren lassen. Mit der üblichen zufälligen Kreuzvalidierung gelingt dies mit mittleren Genauigkeiten von 99,95 ± 0,01 % und mit der Kreuzvalidierung mit Partitionierung anhand einer kategorialen Variablen (hier: laufende Nummer eines Filterwechsels) brechen die erreichten Genauigkeiten auf 85,3 ± 8,7 % ein. Wir schlussfolgern, dass die Aufzeichnungen jedes der untersuchten Filtereinsätze individuelle Merkmalsausprägungen aufweisen, die zu einem Data-Leakage-Problem und zu beschränkt generalisierungsfähigen Problemlösungen führen.

1. Einleitung

In einer Reinstwasser-Aufbereitungsanlage, die in der Mikroelektronik-Industrie eingesetzt wird, befinden sich mehrere Filterstufen, die dem Hauptaggregat, einem Umkehrosmose-Filter, vorgeschaltet sind. Diese Vorfilter können über mehrere Wochen hinweg eingesetzt werden, bis ein Austausch des Filtereinsatzes erforderlich wird. Leider kommt es aus schwer nachvollziehbaren Gründen immer wieder dazu, dass seit längerem eingesetzte Filter sich binnen kurzer Zeit rapide zusetzen, sodass die erforderliche Flussmenge unterschritten wird und aufgrund der hohen Druckdifferenzen, die Förderpumpe abschaltet. Der Vorratstank wird folglich nicht mehr ausreichend befüllt und im schlechtesten Fall kann es zu einer Produktionsunterbrechung kommen mit erheblichen

Abbildung 1: Ausschnitt der Aufbereitungsanlage mit Vorfilterstufe und applizierten Sensoren.

finanziellen Verlusten. Mit einer vorausschauenden Wartung basierend auf automatischer Überwachung und Zustandsprognose soll dieses Problem gelöst werden. Hier wird ein Konzept mit vibroakustischen Sensoren vorgestellt [1], die außen an den Filtergehäusen befestigt wurden. Diese Sensoren haben den Vorteil, sensitiv zu sein und einen hohen dynamischen Bereich abzudecken. Sie messen im Wesentlichen die Strömungsgeräusche des Wassers im Inneren des Rohrleitungssystems. Aber auch anderenorts eingekoppelter Körperschall wird gemessen, da sich der Körperschall mit relativ geringer Dämpfung auf den Edelstahl-Bauteilen ausbreitet. Noch nicht untersucht wurde, wie gut sich auch frei Schallwellen aus der Umgebung einkoppeln und zu vibroakustischen Störsignalen werden [2].

2. Material und Methoden

Mit zwei vibroakustischen Sensoren wurde der Körperschall vor und nach den Vorfilterstufen mit einer Abtastrate von 16kHz aufgezeichnet. Zur Auswertung wurde ein Zeitraum von 48h vor bis 48h nach einem Filterwechsel ausgewählt. Die Musterbeispiele für die Klasse -1 (degenerierter Filterzustand) entstammen dem ersten Zeitraum, der mit einem Filterwechsel endet. Der zweite Zeitraum ist durch einen neuen Filtereinsatz gekennzeichnet und enthält somit Musterbeispiele für die Klasse +1 (intakter Filterzustand). Insgesamt lagen Aufzeichnungen zu fünf Filterwechseln vor. Durch Segmentierung der beiden 48h-Intervalle in 60s-Segmente konnten für jede Klasse jeweils 2.880 Signalsegmente analysiert werden. Aus jedem Segment wurden folgende Merkmale extrahiert:

- Perzentile der Amplitudenverteilung
- Momentanfrequenz
- Entropie-Maße und fraktale Dimensionen
- Entropie-Maße über Koeffizientenfolgen der diskreten Wavelet-Transformation
- Logarithmierte spektrale Leistungsdichten (LogPSD)

Diese Merkmale wurden unskaliert durch eine Gradient-Boosting-Methode zur Klassifikationsanalyse weiterverarbeitet. Es wurden parallel dazu folgende drei Skalierungsvarianten empirisch geprüft, ob sie zu höheren Klassifikationsgenauigkeiten führen:

- Minimum-Maximum-Skalierung
- Z-Skalierung
- Quantil-Transformation

Für das Gradient Boosting wurde LightGBM, eine rechenzeiteffiziente Implementierung von Microsoft Research, eingesetzt [3]. Die im nachfolgend präsentierten Ergebnisse wurden auf Basis einer Kreuzvalidierung vom Typ der wiederholten zufälligen Partitionierung in Trainings- und Validierungsmengen ermittelt (Partitionierungsverhältnis 4:1, 5 Wiederholungen). Die Leave-One-Subject-Out-Kreuzvalidierung (LOSO-CV) [4] wurde alternativ verwendet, bei der die Datenmenge ein und derselben Kategorie (subject) aus dem Training herausgehalten und nur zum Validieren genutzt wird; die anderen Datenmengen werden zum Training verwendet. Hier ist die laufende Nummer eines Filterwechsels die kategoriale Variable, sodass die aufgezeichneten Sensordaten von -48 h bis 0 h vor dem Filterwechsel der Klasse -1 (Filter degeneriert) und die Aufzeichnungen von 0 h bis +48 h nach einem Filterwechsel der Klasse +1 (Filter intakt) zugeordnet werden. Die Daten vor und nach einem Filterwechsel bilden somit die Validierungsmenge

und die Daten vor und nach allen anderen vier Filterwechseln bilden die Trainingsmenge. Dies wird wiederholt, sodass jede Kategorie einmal die Validierungsmenge festlegt.

Zur Merkmalsreduktion wurden LogPSD-Variablen, die in äquidistanten Frequenzintervallen liegen, gemittelt (sogenannte Bandmittelung). Die Parameter dieser Methode sind die untere Grenzfrequenz (Start-Frequenz), die obere Grenzfrequenz (Stop-Frequenz) und die Intervallbreite (Schrittweite). Diese Parameter wurden empirisch optimiert.

3. Ergebnisse

In den ersten Analysen zeigte sich, dass die LogPSD-Merkmale zu den höchsten mittleren Genauigkeiten an Validierungsmengen führen, sodass die anderen Merkmalstypen nicht weiter untersucht wurden. Durch empirische Optimierung der drei oben genannten Bandmittelungs-Parameter wurde die Genauigkeit von 79,84% auf 85,30% gesteigert (Abb. 2). Die Ergebnisse wurden mit LOSO-CV ermittelt, wobei über alle fünf Kategorien (Filterwechsel) gemittelt wurde.

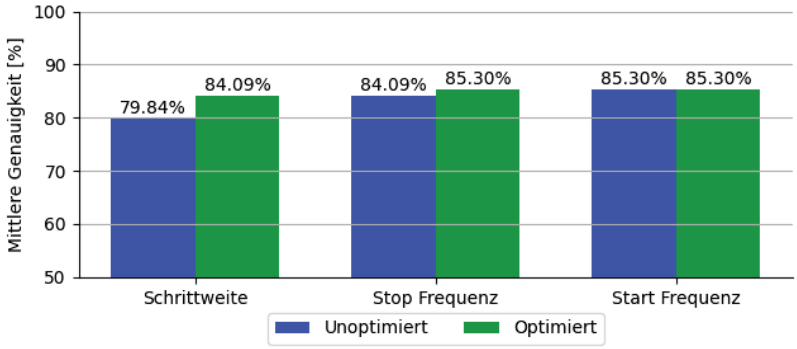

Abbildung 2: Mittlere Klassifikationsgenauigkeiten an Validierungsmengen vor und nach der empirischen Parameteroptimierung.

Die Ergebnisse der Parameteroptimierungen streuen relativ stark, wenn man auf die zuletzt genannte Mittelung verzichtet und somit für jeden individuellen Filtereinsatz die Ergebnisse analysiert (Abb. 3). Kategorie 2, 3 und 4 führen zu relativ hohen Genauigkeiten, wogegen Kategorie 1 und 5 deutlich schlechter zu klassifizieren sind. Die Parameteroptimierungen wurden für alle Kategorien einheitlich durchgeführt, sodass der Effekt auch negativ sein kann, wie man bei Kategorie 3 erkennen kann, wo durch die Optimierung die mittlere Genauigkeit um 0,16% geringer wurde.

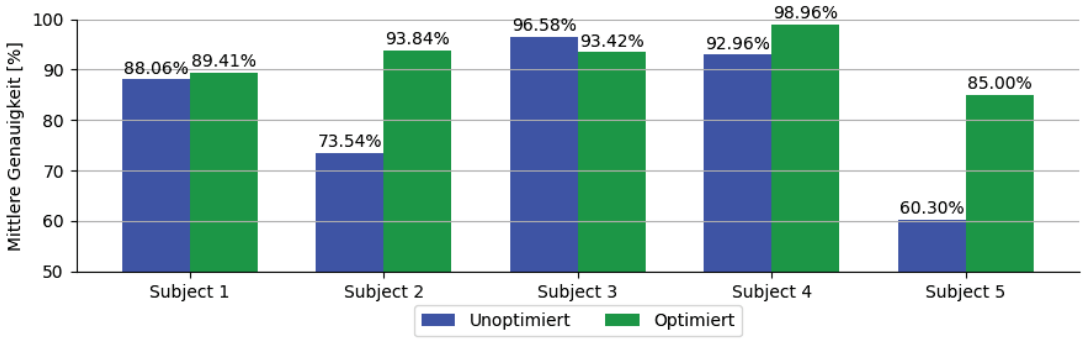

Abbildung 3: Mittlere Klassifikationsgenauigkeiten an Validierungsmengen vor und nach der empirischen Parameteroptimierung für jeden einzelnen Filterwechsel (subject).

Bei der üblichen Kreuzvalidierung der wiederholten zufälligen Partitionierung ergeben sich deutlich höhere Genauigkeiten (Tabelle 1) und niedrigere Standardabweichungen.

Kreuzvalidierung	Klassifikationsgenauigkeit
Wiederholte zufällige Partitionierung	99,95 ± 0,01 %
Leave-One-Subject-Out	85,3 ± 8,7 %

Tabelle 1: Mittelwerte und Standardabweichungen der Validierungsgenauigkeit für zwei verschiedene Kreuzvalidierungsmethoden. Die Trainingsgenauigkeit lag stets bei 100%.

4. Diskussion

Die Nutzung von Daten aller Kategorien in den Trainingsmengen führt zu einem Data-Leakage-Problem, einem der zehn größten Fehler im Data Mining [5]. Es liegt vor, wenn Informationen über die Zielvariable eines Klassifikations- oder Regressionsproblems eingebracht werden, die nicht legitim verfügbar sein sollten [6]. Zwei mögliche Quellen dieses Problems sind Merkmale und Trainingsbeispiele. Hier handelt es sich um Letzteres. Wenn das Ziel darin besteht, kategorienunabhängige Modelle abzuleiten, dann dürfen nicht alle Kategorien in der Trainingsmenge vertreten sein, weil stets die Validierungsmenge statistisch unabhängig bzgl. der Trainingsmenge sein muss. Kategorien, die während des Trainings verfügbar sind, müssen von der Validierung ausgeschlossen werden.

Bei der wiederholten zufälligen Partitionierung werden Kategorien innerhalb des Datensatzes nicht berücksichtigt. Dagegen wird bei der LOSO-CV hingegen wird kategorienbezogen eine Teilmenge aus dem Training herausgehalten und nur zum Validieren genutzt wird. Dadurch erhält man hier eine Schätzung der wahren Klassifikationsgenauigkeit, die nicht verzerrt ist durch den Einfluss dieser Kategorie.

Die erreichten hohen mittleren Klassifikationsgenauigkeiten bei der wiederholten zufälligen Partitionierung sind optimistisch verzerrt, denn sowohl in der Trainingsmenge als auch in der Validierungsmenge waren Beispiele aus allen fünf Kategorien. Die um ca. 15% geringere mittlere Klassifikationsgenauigkeit bei der LOSO-CV zeigt, dass klassenbedingte Verteilungsdichteunterschiede zwischen den Kategorien vorliegen und dass es nicht möglich ist, eine kategorienunabhängige Abbildung mit hoher mittlerer Genauigkeit zu finden, obwohl mit der Klassifikationsmethode LightGBM im Training stets eine Abbildung mit hoher Genauigkeit für die Trainingsmenge gefunden werden konnte.

Mit der LOSO-CV simuliert man, dass zukünftige unbekannte Datensätze hinzukommen, die zu bisher nicht bekannten Kategorien, hier Filtern, gehören. Der relativ hohe Genauigkeitsverlust von ca. 15% mit vergleich zur wiederholten zufälligen Kreuzvalidierung unterstreicht die hohe Bedeutung von umfassenden Validierungen. Trainings- und Validierungsmengen müssen statistisch unabhängig sein, um valide Aussagen zum Risiko der Klassifikation zu erhalten. Bei hohen inter-kategorialen Streuungen der Merkmalsausprägungen ist eine zufällige wiederholte Kreuzvalidierung optimistisch verzerrt. Nur eine Kreuzvalidierung mit Heraushalten einzelner Kategorien aus dem Trainingsprozess deckt die Verzerrung der Ergebnisse auf und führt zu einer Schätzung, wie genau ein Modell ist, das unabhängig von einzelnen Kategorien sein soll [6].

Es ist nicht auszuschließen, dass auch andere Einflüsse den Trainingsprozess beeinflusst haben. Es könnten weitere kategoriale Variablen existieren, die einen Einfluss auf

den Datenentstehungsprozess und folglich eventuell auch auf die Merkmalsausprägungen haben. Deshalb ist eine erweiterte Validierung sinnvoll, wo dieser Einfluss bezüglich der kategorialen Variable untersucht wird und mit LOSO-CV das Ausmaß des Einflusses auf die Validierungsgenauigkeit geschätzt wird. Solche Einflussvariablen sind vielfältig; sie könnten bspw. in der Jahres-, Wochen-, Tageszeit oder auch in örtlichen Abhängigkeiten gefunden werden. Dies erfordert allerdings eine umfassende Aufzeichnung aller begleitenden Umstände des zu untersuchenden Prozesses und steht damit im Einklang mit der Vision der umfassenden Datafizierung (datafication) aller wichtigen Prozesse in Wirtschaft, Behörden und weiteren gesellschaftlichen Bereichen [7].

Quellen

[1] Martini A, Troncossi M, Rivola A (2017) Vibroacoustic Measurements for Detecting Water Leaks in Buried Small-Diameter Plastic Pipes. J Pipeline Syst Engin Pract 8(4): 04017022.

[2] Siba M, Wanmahmood W, Nuawi MZ, Rasani R, Nassir M (2016) Flow-induced vibration in pipes: Challengess and solutions-A review. J Engin Scie Technol 11(3):362-382.

[3] Ke G, Meng Q, Finley T, Wang T, et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Proc Syst; 30:3146-54.

[4] Xu G, Huang JZ (2012) Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. The Annals of Statistics; 40(6):3003-30.

[5] Nisbet R, Elder J, Miner G (2009) Handbook of Statistical Analysis and Data Mining Applications. Academic Press.

[6] Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: Formulation, detection, and avoidance. ACM Trans Knowl Discov Data (TKDD); 6(4):1-21.

[7] Cukier K, Mayer-Schoenberger V (2013). The rise of big data: How it's changing the way we think about the world. Foreign Aff (92):28.