Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Umwelt, Verkehr, Energie und Kommunikation UVEK

Bundesamt für Energie BFE

Zwischenbericht 09.04.2015

Smart-Meter-Datenanalyse für automatisierte Energieberatungen

(„Smart Grid Data Analytics“)

Auftraggeber:

Bundesamt für Energie BFE
Forschungsprogramm Elektrizitätstechnologien & -anwendungen
CH-3003 Bern
www.bfe.admin.ch

Auftragnehmer:

ETH Zürich
D-MTEC
Lehrstuhl für Informationsmanagement
Weinbergstrasse 56/58
CH-8092 Zürich
www.im.ethz.ch

ETH Zürich
Distributed Systems Group
Institute for Pervasive Computing, CNB
Universitätstrasse 6
CH-8092 Zürich
www.vs.inf.ethz.ch

Autoren:

Mariya Sodenkamp, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftsinformatik, insbes. Energieeffiziente Systeme, mariya.sodenkamp@uni-bamberg.de

Ilya Kozlovskiy, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftsinformatik, insbes. Energieeffiziente Systeme, ilya.kozlovskiy@uni-bamberg.de

Christian Beckel, ETH Zürich, D-INFK, Lehrstuhl für Verteilte Systeme, beckel@inf.ethz.ch

Thorsten Staake, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Wirtschaftsinformatik, insbes. Energieeffiziente Systeme und ETH Zürich, D-MTEC, Lehrstuhl für Informationsmanagement, tstaake@ethz.ch

BFE-Bereichsleiter: Dr. Michael Moser
BFE-Programmleiter: Roland Brüniger
BFE-Vertragsnummer: SI/501053-01

Inhaltsverzeichnis

1. Zusammenfassung

Kommunikationsfähige Stromzähler („Smart Meter“) ermöglichen die Erfassung individueller Lastprofile mit hoher zeitlicher Auflösung. Projektgegenstand ist die Weiterentwicklung von Methoden des maschinellen Lernens, um aus Lastprofilen automatisiert Merkmale von Haushalten abzuleiten, welche für eine individuelle und spezifische Energieberatung von Nutzen sind. Dadurch lassen sich IT-unterstützte und skalierbare Effizienzkampagnen realisieren. In der folgenden Projektphase wird eine mehrdimensionale Klassifikation entwickelt, um mit zusätzlichen Daten aus Geo-Informationssystemen und Kundenstammdaten die Qualität der Merkmalserkennung zu verbessern sowie die Anzahl der erkennbaren Merkmale zu erhöhen. Daneben sollen die Algorithmen mit Daten des Schweizerischen Versorgers Arbon Energie AG validiert werden.

2. Summary

Smart electricity meters allow for capturing consumption data of individual households at high resolution in time (typically with 15-minute intervals). The key objective of this project is to develop further and evaluate feature extraction and machine learning techniques for automatic identification of household properties based on electricity load profiles. The gained information shall render highly targeted and scalable energy efficiency services possible. In the following project phase, a multidimensional classification using additional data from geographical information systems and core customer data will be developed. This will be done in order to elevate the classification quality as well as to increase the number of recognizable household characteristics. Furthermore, the algorithms will be validated based on the data of the Swiss utility company Arbon Energie AG.

3. Projektziele

Übergeordnetes Ziel des Projektes ist die Entwicklung von Algorithmen zur automatisierten Identifikation von Haushaltsmerkmalen aus Haushaltslastprofilen mit einer zeitlichen Auflösung von zwei oder vier Messwerten pro Stunde. Dazu werden Verfahren des maschinellen Lernens weiterentwickelt und mit bestehenden Trainings- und Validierungsdaten von über 4'000 Haushalten erprobt. Die Merkmale umfassen die Charakteristika *Single-Haushalt, Kinder, Berufsstand* (jeweils ja/nein), *Anzahl der Einwohner, Trocknernutzung, Spülmaschinennutzung, Alter des Hauses, Wohnfläche* (über- vs. unterhalb des Medians) sowie Informationen zu typischen Anwesenheitszeiten (etwa für Lastverschiebungskampagnen). Durch Kenntnis der Merkmale lässt sich die Wirkung von Feedback-Informationen zum Verbrauch und Verhaltensinterventionen deutlich steigern (z.B. durch die Wahl relevanter Vergleichsgruppen, die Selektion zielgerichteter Empfehlungen, das Setzen realistischer aber ambitionierter Ziele, etc.). Gesamthaft schätzen wir, dass durch den konsequenten Einsatz der gewonnenen Informationen ein energetisches Potenzial in der Schweiz von ca. 180 GWh entsteht (konservativ geschätzt, entsprechend etwa einem Prozentpunkt zusätzliche Einsparungen gegenüber konventioneller Smart-Meter-Nutzung in Haushalten).

Wir beabsichtigen, im Projekt insbesondere durch eine Verbesserung von bisher erst rudimentär erforschten Methoden die Prognosequalität der Merkmalserkennung zu erhöhen. Je nach Merkmal streben wir eine Treffsicherheit (Accuracy) von 70% bis 90% an. Als Datengrundlage dienen Zeitreihen von Stromverbräuchen, welche bei gängigen Smart-Meter-Infrastrukturen erfasst und übertragen werden. Dadurch kann die konventionelle Energieberatung die gewonnenen Informationen nutzen, um besonders beratungsrelevante Haushalte bevorzugt anzusprechen und so (vor allem durch eine Reduktion der Streuverluste) die Kosten je gesparter Kilowattstunde zu reduzieren.

Ein weiteres Ziel, folgend auf die Verbesserung der Methoden, ist die Steigerung der Effizienz und Skalierbarkeit der Algorithmen zur Haushaltsklassifikation. Hierbei sollen mit gängiger Hardware (z.B. mit einem Server mit Anschaffungskosten von 15'000 CHF) 10'000 Haushalte pro Stunde klassifizieren zu können. Weiter sollen die Methoden in Form eines nutzerfreundlichen Software-Pakets einem breiteren Anwenderkreis (insbesondere auch „Nicht-Programmierern“) verfügbar gemacht werden. Dieses soll einen einfachen Import der Daten und eine einfache Parametrisierung der Algorithmen ermöglichen sowie unterschiedliche Filtermethoden für Merkmale, Haushaltstypen etc. enthalten.

4. Konzept und Aufbau des Haushaltsklassifikationssystems

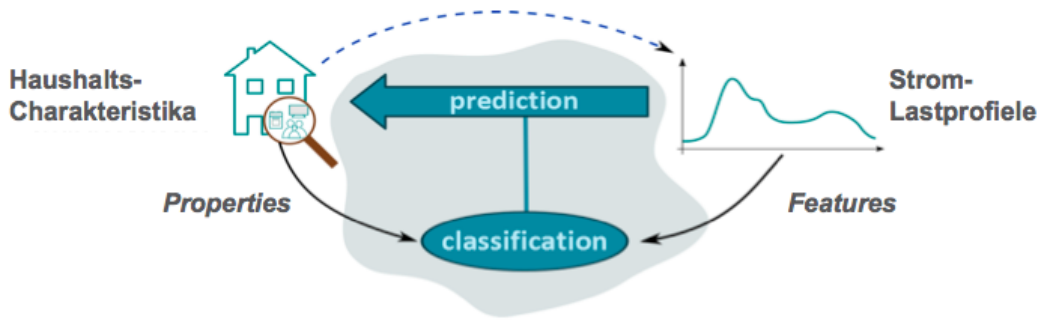

Abbildung 1: Konzept des Klassifikationssystems

4.1: Dimensionsreduktion der Lastkurven

4.1.1. Feature Extraction. Wir haben 95 Eigenschaften der Verbrauchszeitreihen empirisch identifiziert und ihre automatische Ableitung aus den Lastkurven umgesetzt. Die Features wurden in vier Kategorien unterteilt:

- Verbrauchswerte (z.B. min. Verbrauch, max. Verbrauch, Tagesverbrauch, Verbrauch am Wochenende usw.)
- Statistische Momente (z.B. Durchschnittsverbrauch am Tag, Korrelation von auf einander folgenden Tagen, Varianzen, usw.)
- Zeitliche Kennzahlten (z.B. Zeiten der Peaks, der erste Zeitpunkt mit Verbrauch > 1 kWh, usw.)
- Verhältnisse (z.B. Verbrauch in der Mittagszeit / Verbrauch am Abend, min. Verbrauch / max. Verbrauch, max. Verbrauch / Durchschnittsverbrauch, usw.)

Abbildung 2 stellt ein Beispiel für die Feature Extraktion dar.

Resultate: Bei der Klassifikation 30-minütiger Messwerte (336 Messpunkte pro Woche) führt die Extraktion von Lastkurvenmerkmalen zu fast vierfacher Reduktion der Eingabedaten und trägt zudem zur Vermeidung von Problemen mit hoher Dimensionalität / Overfitting bei. Insgesamt wurden 95 aussagekräftige Features identifiziert und erprobt. Die Gesamtliste der Features ist im Anhang 1 aufgeführt.

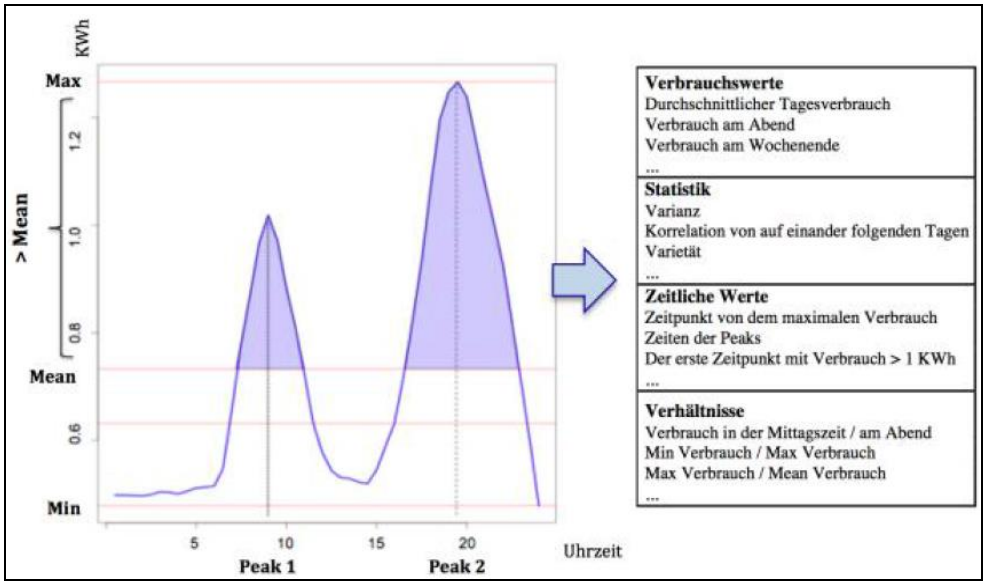

Abbildung 2: Beispiel für „Feature Extraktion“

4.1.2. Feature Selection und Filtering. Um für jede Haushaltseigenschaft die aussagekräftigsten Features auszuwählen, wurden folgende Ansätze erprobt:

(a) Feature Selection: Wrapper Methoden, die Klassifikation mit unterschiedlichen Feature Submengen mehrfach ausführt um den Gütermass direkt zu vergleichen, insb. Sequential Forward Selection und Paralell Sequential Forward Selection, sowie heuristische Wrapper Methoden wie Backward Elimination und Random Forests.

(b) Feature Filering: Etablierte Methoden der Feature-Auswertung basierend auf intrinsischen Dateneigenschaften, insb. Correlation-Based (CB) für Selektion von wenig mit einander korrelierter Features, und Kolmogorov-Smirnov-Test (KST) zum Herausfiltern der Features mit ähnlichen Distributionen in Klassen.

Statistische Ansätze zur Messung der Effektstärke, die bisher in der Literatur für FF nicht beschrieben wurden, insb. Pearson η^2-basierend (misst, wie gut ein Feature Klassen separiert) und Kombinierendes CB und KST Verfahren (misst, wie gut die Features Properties beschreiben & dann nur die wenig korrelierte Features wählen), wurden berücksichtigt.

Resultate: Feature-Selektions-Methoden sind für die Haushaltsklassifikation wegen ihrer nur marginalen Vorhersageverbesserungen sowie einem hohen Rechnungsaufwand schlecht geeignet. Filtering-Methoden haben es uns hingegen ermöglicht, die Fehlerquote um bis zu 15.2% zu reduzieren. Die Verbesserungen für einzelne Eigenschaften sind in der Tabelle 1 zusammengefasst.

Property	Accuracy ohne Feature Selection	Accuracy mit Fature Selection	Reduktion der Fehlerquote	Feature-Selektions-Methode
Single	0.80	0.83	15%	Pearson η^2
#devices	0.52	0.54	4.2%	Pearson η^2
Cooking	0.68	0.71	9.4%	Pearson η^2
Family	0.73	0.76	11.1%	CB
Children	0.63	0.68	13.5%	CB
Age_house	0.60	0.60	0%	KST
Social_class	0.51	0.52	2%	CB
Floor area	0.60	0.64	10%	CB
#residents	0.71	0.73	6.9%	Pearson η^2
#bedrooms	0.50	0.50	0%	Pearson η^2
Employment	0.70	0.72	6.7%	KST
Retirement	0.67	0.72	15.2%	Pearson η^2

Tabelle 1: Klassifikationsgenauigkeit ohne und mit Feature Filtering

4.2. Optimierung der Klassifizierer zur Steigerung der Performanz

4.2.1. Performanz Evaluierung. Die Performanz des Systems wird anhand von Fragebogen-Daten bewertet:

Schritt 1: Haushalte in zwei Gruppen einteilen: „Traings-Set“ und „Test-Set“
Schritt 2: Modell auf Basis des Trainings-Datensatzes entwickeln (Klassifikationsverfahren)
Schritt 3: Haushaltscharakteristika mit dem Test-Set schätzen
Schritt 4: Geschätzte Haushaltscharakteristika mit Fragebogen-Daten vergleichen
Schritt 5: Schritt 1 bis 4 für verschiedene Test-/Trainingsdaten wiederholen

Abbildung 3 stellt das Klassifikationssystem mit der Bewertung der Performanz dar.

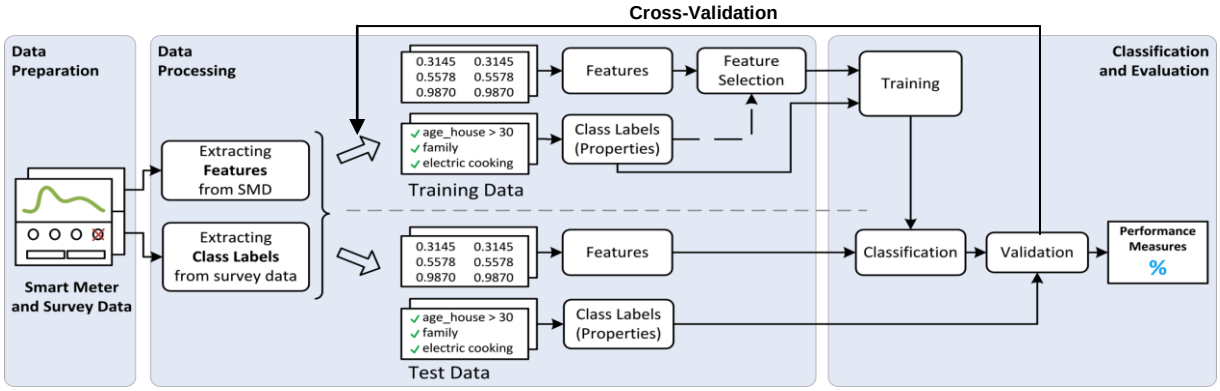

Abbildung 3: Performanz-Analyse des Systems

4.2.2. Klassifikationsverfahren. Für die genaue Haushaltsklassifikation gemäss des Energieverhaltens stützen sich unsere Lösungen auf von uns modifizierte „Supervised Machine Learning Algorithmen“:

- Klassifikation einzelner Wochen: Es wurden sieben Klassifikatoren parametrisiert, erprobt und verglichen: k Nearest Neighbors, Support Vector Machines (SVM), Linear Discriminant Analysis, Mahalanobis Distances, Adaboost, Neural Networks, und Random Forest. Zusätzlich wurden 20 Klassifikationsalgorithmen für einige Properties untersucht.

- Klassifikation mehrerer Wochen: Erweiterung des Betrachtungszeitraums um fünf repräsentative Wochen (verschiedene Monate, keine Feiertage, durchschnittliche Wetterbedingungen) mit Majority Voting für Urteilsfindung. Die Ensemble-Methoden wurden getestet, um die Klassifikationsergebnisse von mehreren Wochen zu aggregieren.

Resultate: mit der Klassifikation einzelner Wochen gelingt es, die Treffsicherheit (Accuracy) von 82% zu erreichen. Betrachtung mehrerer Wochen, Kreuzvalidierung, und Feature Filtering („Best result“ auf Abbildung 4) reduziert die Fehlerquote um durchschnittlich 18%.

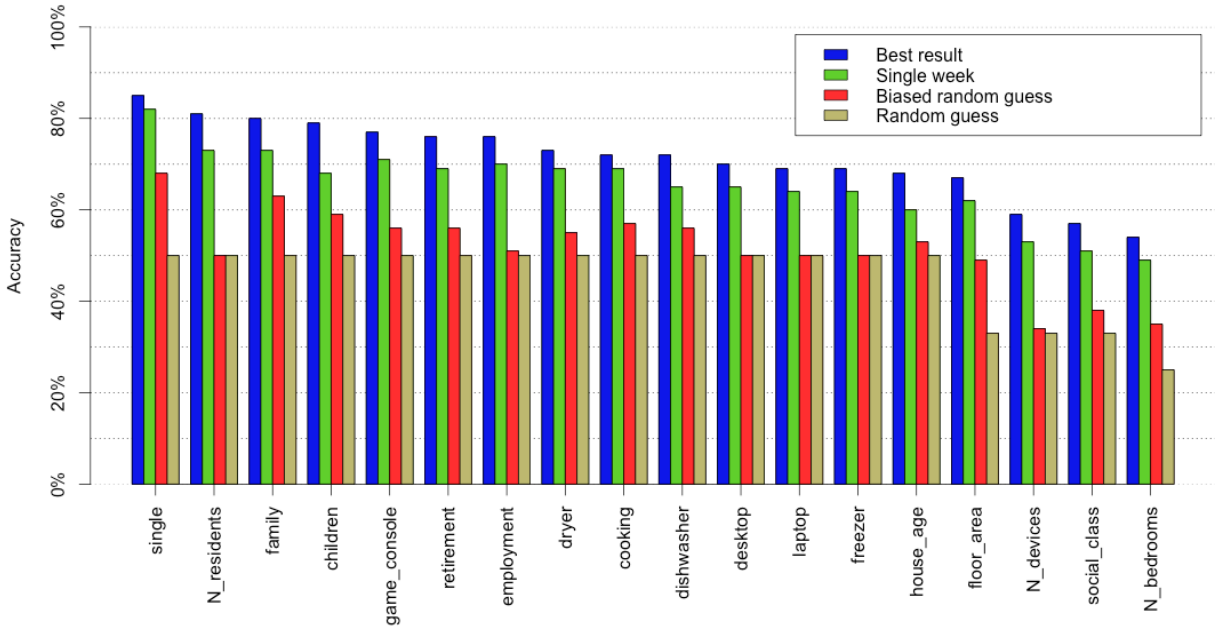

Abbildung 4: Klassifikationsgenauigkeit als Accuracy

4.2.3. Umgang mit unbalancierten Klassen

Kleine Klassen (z.B. Single 20% vs. Not Single 80%) werden mit herkömmlichen Algorithmen schlecht erkannt, wenn die Genauigkeit als Auswertungsmetrik herangezogen wird. Als Lösung wurde SMOTE (Synthetic Minority Oversampling Technique) implementiert. Hierbei werden zum Training zusätzliche Repräsentanten der kleineren Klasse generiert und als (zufällig)-gewichtetes Mittel von mehreren nahen tatsächlichen Repräsentanten generiert. Wir definieren die Klassen als unbalanciert, falls die [Grösse der kleinsten Klasse]*3 ≤ [Grösse der grössten Klasse] für das Property ist. So lässt sich die Genauigkeit als Precision für die Properties „Family“, „#bedrooms“, „Children“, „Single“ und „Floor_area“ um durchschnittlich 27% verbessern. Die Resultate sind in der Tabelle 2 zusammengefasst.

Property	Precision für einzelne Woche ohne SMOTE	Precision für mehrere Wochen mit Feature Selection und SMOTE
Family	0.61	0.72
#Bedrooms	Unbestimmt	0.43
Children	0.62	0.86
Single	0.57	0.73
Floor_area	Unbestimmt	0.29

Tabelle 2: Klassifikationsgenauigkeit als Precision bei Erkennung von unbalancierten Klassen

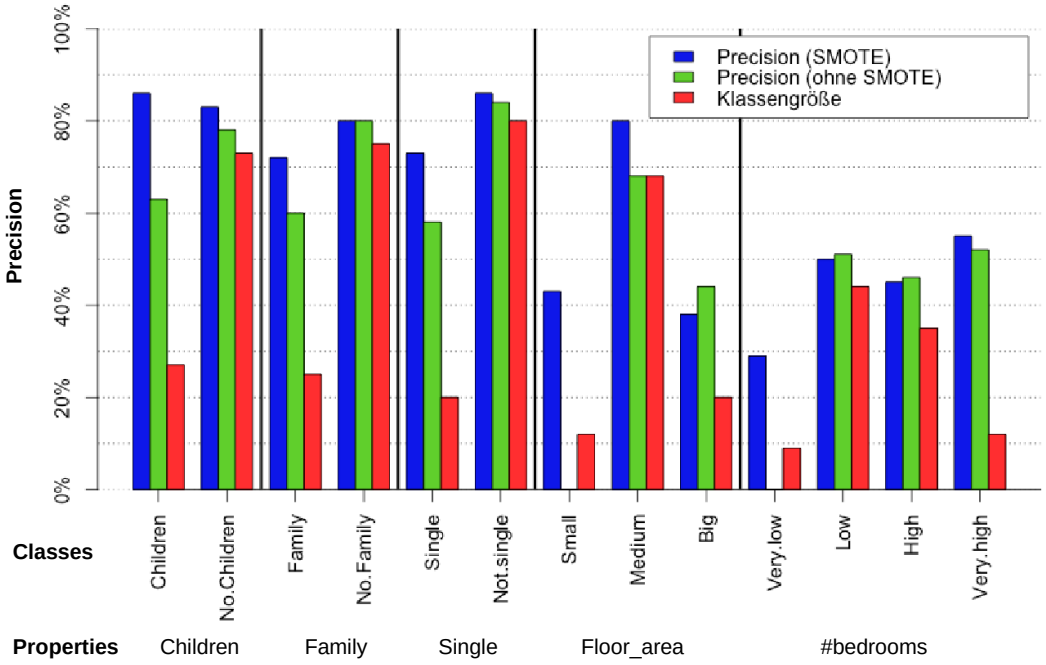

Abbildung 5: Klassifikationsgenauigkeit als Precision

4.2.4. Definition und Optimierung der Klassengrenzen

Das System erkennt 12 energieeffizienzrelevante Haushaltscharakteristika, die in diskrete Gruppen (Klassen) unterteilt sind. Die Ziele der Optimierung von Klassengrenzen sind: (1) die Performanz der Algorithmen steigern, und (2) es EVU ermöglichen, zielgruppenspezifische Interventionen zu wählen (z.B. Ökostromprodukte für Familien mit Kindern und hohem Einkommen).

Resultate: Unterteilung der Charakteristika ist in der Tabelle 3 gegeben.

Household property	Classes and their labels
Number of appliances and entertainment devices (`#devices`)	Low (`#devices` ≤ 8) Medium ($8 <$ `#devices` ≤ 11) High ($11 <$ `#devices`)
Number of bedrooms (`#bedrooms`)	Very low (`#bedrooms` ≤ 2) Low (`#bedrooms` $= 3$) High (`#bedrooms` $= 4$) Very high ($4 <$ `#bedrooms`)
Type of cooking facility (`cooking`)	Electrical Not electrical
Employment of chief income earner (`employment`)	Employed Not employed
Family (`family`)	Family (`#adults` > 1 and `#children` > 0) No family
Floor area (`floor_area`)	Small (`floor_area` $\leq 100\ m^2$) Medium ($100\ m^2 <$ `floor_area` and `floor_area` $\leq 200\ m^2$) Big ($200\ m^2 <$ `floor_area`)
Children (`#children`)	Children (`#children` ≥ 1) No children (`#children` $= 0$)
Age of building (`age_house`)	Old ($30 <$ `age_house`) New (`age_house` ≤ 30)
Number of residents (`#residents`)	Few (`#residents` ≤ 2) Many ($3 \leq$ `#residents`)
Single (`single`)	Single (`#adults` $= 1$ and `#children` $= 0$) No single
Retirement status of chief income earner (`retirement`)	Retired Not retired
Social class of chief income earner according to NRS social grades (`social_class`)	A or B C1 or C2 D or E

Tabelle 3: Haushaltscharakteristika und entsprechende Klassen-Labels

4.3. Optimierung hinsichtlich Skalierbarkeit

Um gute Skalierbarkeit der Algorithmen zu ermöglichen und mehrere tausend Haushalten gleichzeitig zu klassifizieren, wurden zwei Ansätze implementiert:

a) Erkennung und Eliminierung von Datenpunkten, die nicht ausreichend zur Klassifizierung beitragen (Dimensionsreduktion und Feature Filtering). Tabelle 4 zeigt, wie häufig die 20% der repräsentativsten Features ausgewählt wurden. Tabelle 5 zeigt die Features, die gleichzeitig von sechs oder mehr Feature-Selection-Methoden ausgewählt wurden.

Kategorie (Anzahl der Features)	**20% der am häufigsten ausgewählten Features**	**Methode der Feature-Selektion**								**Σ Auswahlhäufigkeit über alle Methoden**
		Keine	**Kombiniert**	**CB**	**Pearson η²**	**KST**	**Wrapper Forward Selection**	**Wrapper Backward Selection**	**Wrapper Parallel Sequential**	
Verbrauch (28)	c_night_no_min	12	6	12	6	6	0	11	0	53
	c_min	12	9	0	7	3	0	10	8	49
	c_evening	12	8	0	7	7	0	11	2	47
	c_max	12	8	0	6	3	0	12	6	47
	c_weekday	12	6	0	5	5	0	12	6	46
	c_weekend	12	7	0	7	4	1	12	3	46
Statistiken (16)	s_num_peaks	12	10	12	9	3	0	12	0	58
	s_var_we	12	8	12	8	7	0	11	0	58
	s_cor_wd_we	12	7	12	8	6	1	12	0	58
Zeitliche Werte (21)	t_daily_max	12	10	12	9	7	0	12	3	**65**
	dist_big_v	12	10	12	6	6	0	12	0	58
	time_above_base2	12	10	12	7	5	0	11	0	57
	t_above_base	12	10	12	5	4	1	12	0	56
Verhältnisse (23)	r_var_wd_we	12	10	12	8	5	2	11	0	**60**
	r_wd_night_day	12	8	12	9	7	0	12	0	**60**
	r_evening_noon	12	10	12	7	3	0	11	4	**59**
	r_night_wd_we	12	10	12	7	6	0	12	0	**69**
	r_morning_noon_no_min	12	10	12	9	4	0	11	0	58

Tabelle 4: 20% der am häufigsten ausgewählten Features

Nr.	**Property**	**Features**
1	#devices	s_cor_wd, r_min_wd_we
2	#bedrooms	c_max, r_night_day, r_morning_noon, r_evening_noon, r_mean_max, r_min_mean, t_above_mean, t_daily_max, … (37 andere)
3	cooking	t_above_base
4	employment	r_min_mean, t_daily_max, r_morning_wd_we, r_morning_noon_no_min, c_night_no_min, percent_above_base, dist_big_v
5	family	s_cor_wd_we
6	retirement	r_morning_noon
7	children	r_evening_noon, r_night_wd_we, s_var_we, s_var_wd, r_var_wd_we, s_cor_wd_we, r_wd_night_day, t_daily_min, ts_stl_varRem, b_day_diff, b_day_weak, r_mean_max_no_min, time_above_base2, dist_big
8	age_house	t_daily_max, r_we_morning_noon, r_mean_max_no_min
9	#residents	t_daily_max, r_var_wd_we, value_min_guess
10	Single	r_evening_noon, t_daily_max, s_cor_wd, r_evening_wd_we, s_var_we, s_cor_wd_we, r_wd_morning_noon, r_we_evening_noon, t_above_base, r_morning_noon_no_min, value_min_guess
11	#household devices	r_evening_noon, r_mean_max, t_daily_max
12	#entertainment devices	r_mean_max, t_above_mean

Tabelle 5: Aussagekräftige Features für jede Property

b) Ausschuss von Wochen ohne Anwesenheit.

Die erste Möglichkeit Abwesenheiten zu erkennen, ist die Klassifikation anhand von Trainingsdaten. Zum Irischen Datensatz liegen solche Daten jedoch nicht vor.

Die empirische Lösung ist die Erstellung von Heuristiken, um zu erkennen, ob während der Woche die Bewohner tatsächlich Strom verbraucht haben, oder ob nur eine „Grundlast“ erkennbar ist. Die Grundlast wird durch folgende Eigenschaften gekennzeichnet:

- Der maximaler Stromverbrauch ist klein
- Der Verbrauch variiert stark, aber der Unterschied zwischen Maximum und Minimum ist klein
- Der gleitende Durchschnitt ist nahezu konstant
- Es gibt nur wenige Zusammenhänge zwischen Tageszeit und Verbrauch.

Parameter für die Heuristiken wurden empirisch gewählt und getestet. Erst wurden die Verteilungen der Parameterwerte für alle Haushalte untersucht. Dadurch sollen Haushalte mit An- und Abwesenheit in zwei deutlich trennbare Gruppen geteilt werden. Danach folgt die manuelle Überprüfung der Abwesenheits-Profile. Durch diese Tests konnten die Grenzen für die Abwesenheiten festgelegt werden. Aus allen Heuristiken wurden fünf ausgewählt, die alle manuell erkannten Abwesenheiten korrekt identifizieren.

Heuristik 1. Verhältnis von maximalen Verbrauch zu minimalen Verbrauch < 4 kWh.

Das heisst, es gibt keine grossen Verbrauchsspitzen. Das ist die wichtigste Heuristik, die die meisten Abwesenheitsfälle abdeckt.

Heuristik 2. Maximaler Verbrauch < 0.3 kWh.

Das absolute Maximum ist sehr klein. Insbesondere wichtig für Haushalte die einen niedrigem Gesamtverbrauch haben. Durch diese Heuristik werden nur einzelne Haushalte ausgeschlossen.

Heuristik 3. Mindestens 5 Nullwerte während der Woche.

In den meisten Stromprofilen gibt es keine Nullwerte. Einzelne Nullwerte treten dennoch manchmal auf. Eine grössere Anzahl von Nullwerten deutet auf einen Stromausfall, Gerätefehler, oder Abwesenheit hin.

Heuristik 4. Es gibt ein Tag, an dem das geglättete Verbrauchsprofil kein lokales Maximum hat.

Durch eine starke Glättung der Verbrauchsprofils wird die Variation der Grundlast minimiert. Das heisst, die lokale Maxima entsprechen einem tatsächlichen aktiven Energieverbrauch. Für die Klassifikation nehmen wir an, dass an jedem Tag Strom verbraucht wird. Durch diese Heuristiken werden die Wochen ausgeschlossen, an denen es Tage ohne aktivem Stromverbrauch gibt.

Heuristik 5. Geglättetes Verbrauchsprofil hat ≤ 6 lokale Maxima.

Durch den Ausschluss von den Wochen ohne Anwesenheit steigt die Vorhersagekraft und die Laufzeit sinkt. Das Einlesen und die Berechnung von Features für die Trainingsdaten (3‘500 Datensätze, 18 Monate) dauert 52 min und muss nur einmal ausgeführt werden. Das Einlesen und die Berechnung von Features für die Testdaten (10‘000 Datensätze, 18 Monate) dauert 85 min und muss ebenfalls nur einmal ausgeführt werden. Es müssen auch nur die Features berechnet werden die tatsächlich gewählt worden sind. Die Berechnungen wurden auf einem Rechner mit 1.7 GHz Intel Core i7 CPU und 8 GB 1600 Hz RAM durchgeführt. Beim Training werden durchschnittlich 4 Minuten zur Modellbildung verwendet. Der meiste Zeitaufwand findet bei der Ausführung von SMOTE statt.
Für die beste Konfiguration wird die folgende Ausführungszeit benötigt (Tabelle 6):

Property	Training	Klassifikation
#devices	170 min	2,3 min
#bedrooms	410 min	
cooking	300 min	
employment	120 min	
family	180 min	
retirement	160 min	
children	220 min	
age_house	240 min	
#residents	110 min	
single	320 min	
floor_area	270 min	
social_class	230 min	
dryer	160 min	
dishwasher	170 min	
freezer	270 min	
game_console	230 min	
desktop	180 min	
laptop	190 min	

Tabelle 6: Laufzeit der Klassifikation

4.4. Tool Entwicklung

Die Klassifikationsmethoden wurden als rudimentäre kommentierte Algorithmen in Statistik-Software R entwickelt (GNU General Public License Version 3). Das Klassifikations-Tool wurde modular wie folgt aufgebaut:

- Modul 1: Laden und Transformieren von Daten
  - Umwandlung von rohen Smart-Meter-Daten in Zeitreihen
  - Datenbereinigung
  - Umwandlung von Umfragedaten in Properties und die Zuordnung der passenden Klassenlabels
  - Laden der vorbereiteten Daten
- Modul 2: Datenvorbereitung
  - Auswahl von den benötigten Daten
  - Berechnung von Features
  - Aufteilung in Test- / Trainingsdaten
- Modul 3: Klassifikation mit verschiedenen Algorithmen
  - Durchführung von SMOTE Oversampling für unbalancierte Klassen
  - Auswahl der Features durch Filter oder Wrapper Methoden
  - Konstruktion des Modells mit den Trainingsdaten
  - Klassifikation der Testdaten
- Modul 4: Evaluation mit verschiedenen Metriken
  - Erstellung der Kontingenztabellen
  - Berechnung von verschiedenen Metriken (Accuracy, Precision, Recall, F1-Score, MCC)
- Modul 5: Export und Visualisierung
  - Export von den Klassifikationswerten (Klassenlabels oder zugehörige Wahrscheinlichkeiten) in die Tabelle
  - Darstellung der Klassifikationsgüte als Graphiken (z.B., Barplots)

5. Nationale Zusammenarbeit

Die Arbeiten werden eng von dem ETH-Spinoff BEN Energy AG begleitet. BEN dient als Brückenkopf zu Energieversorger, die ihre Wünsche als mögliche zukünftige Kunden einbringen. BEN implementiert einfacher Analysemethoden in einem verwandten KTI-Projekt, um die Zeit bis zu einer möglichen Produkteinführung zu reduzieren.

6. Ausblick und Folgeaktivitäten

Die gesteckten Projektziele für die Projektphase 1 wurden vollständig erreicht. Anknüpfend an dieses Projekt beabsichtigen wir, die entwickelte eindimensionale Haushaltsklassifikation anhand von Smart-Meter-Daten auf eine mehrdimensionale Klassifikation mit zusätzlichen verbrauchsrelevanten Daten in der Projektphase 2 zu erweitern. Dazu werden unter anderem Informationen wie sozio-demographische Statistiken (z.B. Altersstruktur, Wohnfläche, Familien/Kinder), Stammdaten des Versorgers (historische Rechnungen, Ortsinformationen, Anrede usw.), Satellitenbilder und Nachbarschaftsvergleiche sowie verbrauchsbeeinflussende Umgebungsparameter (z.B. die Aussentemperatur) herangezogen. Dabei soll sowohl die Qualität der Merkmalserkennung verbessert als auch die Anzahl der erkennbaren Merkmale gesteigert werden. Ein weiteres wichtiges Ziel ist die Anpassung der Algorithmen auf einen Schweizerischen Datensatz sowie – sofern der Praxispartner dies ermöglicht – die Durchführung einer Feldstudie zur Quantifizierung der Effekte der durch die Klassifikation massgeschneiderten Interventionen.

7. Referenzen

Im Rahmen der Projekts wurden drei Veröffentlichungen herausgegeben, weitere befinden sich in Vorbereitung:

Sodenkamp, M., Hopf, K., & Staake, T. (2014). Using Supervised Machine Learning to Explore Energy Consumption Data in Private Sector Housing. *Handbook of Research on Organizational Transformations Through Big Data Analytics*, 320.

Hopf, K., Sodenkamp, M., Kozlovkiy, I., & Staake, T. (2014). Feature extraction and filtering for household classification based on smart electricity meter data. *Computer Science-Research and Development*, 1-8.

Beckel, C., Sadamori, L., Staake, T., & Santini, S. (2014). Revealing household characteristics from smart meter data. *Energy*, *78*, 397-410.

Anhang

Anhang 1: Die Gesamtliste der Features

Nr.	Kategorie	Name	Beschreibung
	Kategorie	**Feature Abkürzung**	**Feature Beschreibung**
1	Verbrauch	c_week	Mittlerer Verbrauch über die gesamte Woche
2	Verbrauch	c_morning	Mittlerer Verbrauch am Morgen (6:00 - 9:59 Uhr)
3	Verbrauch	c_noon	Mittlerer Verbrauch am Mittag (10:00 - 13:59)
4	Verbrauch	c_afternoon	Mittlerer Verbrauch am Nachmittag (14:00 - 17:59)
5	Verbrauch	c_evening	Mittlerer Verbrauch am Mittag (18:00 - 21:59)
6	Verbrauch	c_night	Mittlerer Verbrauch in der Nacht (1:00 - 5:59 Uhr)
7	Verbrauch	c_max	Wochen-Maximum
8	Verbrauch	c_min	Wochen-Minimum
9	Verbrauch	c_sm_max	Gleitendes Maximum
10	Verbrauch	c_weekday	Mittlerer Verbrauch an Wochentagen (Mo-Fr)
11	Verbrauch	c_wd_min	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
12	Verbrauch	c_wd_max	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
13	Verbrauch	c_wd_morning	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
14	Verbrauch	c_wd_noon	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
15	Verbrauch	c_wd_afternoon	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
16	Verbrauch	c_wd_evening	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
17	Verbrauch	c_wd_night	Obige Definition, beschränkt auf Wochentage (Mo-Fr)
18	Verbrauch	c_weekend	Mittlerer Verbrauch am Wochenende (Sa,So)
19	Verbrauch	c_we_min	Obige Definition, beschränkt auf Wochenentage (Sa,So)
20	Verbrauch	c_we_max	Obige Definition, beschränkt auf Wochenentage (Sa,So)
21	Verbrauch	c_we_morning	Obige Definition, beschränkt auf Wochenentage (Sa,So)
22	Verbrauch	c_we_noon	Obige Definition, beschränkt auf Wochenentage (Sa,So)
23	Verbrauch	c_we_afternoon	Obige Definition, beschränkt auf Wochenentage (Sa,So)
24	Verbrauch	c_we_evening	Obige Definition, beschränkt auf Wochenentage (Sa,So)
25	Verbrauch	c_we_night	Obige Definition, beschränkt auf Wochenentage (Sa,So)
26	Verbrauch	c_evening_no_min	c_evening abzgl. Minimum
27	Verbrauch	c_morning_no_min	c_morning abzgl. Minimum
28	Verbrauch	c_night_no_min	c_night abzgl. Minimum
29	Verbrauch	c_noon_no_min	c_noon abzgl. Minimum
30	Verbrauch	c_afternoon_no_min	c_afternoon abzgl. Minimum \todo[inline]{noch implementieren
31	Verhältnisse	r_mean_max	Verhältnis c_week / c_max
32	Verhältnisse	r_min_mean	Verhältnis c_min / c_week
33	Verhältnisse	r_night_day	Verhältnis c_night / c_week
34	Verhältnisse	r_morning_noon	Verhältnis c_morning / c_noon
35	Verhältnisse	r_evening_noon	Verhältnis c_evening / c_noon
36	Verhältnisse	r_mean_max_no_min	r_mean_max (Minimum ist jeweils abgezogen)
37	Verhältnisse	r_evening_noon_no_min	r_evening_noon (Minimum ist jeweils abgezogen)
38	Verhältnisse	r_morning_noon_no_min	r_morning_noon (Minimum ist jeweils abgezogen)
39	Verhältnisse	r_day_night_no_min	r_night_day (Minimum ist jeweils abgezogen)
40	Verhältnisse	r_var_wd_we	Verhältnis der Varianz Wochentags - Wochenendtags
41	Verhältnisse	r_min_wd_we	Verhältnis des Minimums Wochentags - Wochenendtags
42	Verhältnisse	r_max_wd_we	Verhältnis des Maximums Wochentags - Wochenendtags
43	Verhältnisse	r_evening_wd_we	Verhältnis des Verbrauchs Abends - Wochentags - Wochenendtags
44	Verhältnisse	r_night_wd_we	Verhältnis des Verbrauchs Nachts - Wochentags – Wochenendtags
45	Verhältnisse	r_noon_wd_we	Verhältnis des Verbrauchs Mittags - Wochentags – Wochenendtags
46	Verhältnisse	r_morning_wd_we	Verhältnis des Verbrauchs Morgens - Wochentags – Wochenendtags
47	Verhältnisse	r_afternoon_wd_we	Verhältnis des Verbrauchs Nachmittags - Wochentags – Wochenendtags
48	Verhältnisse	r_we_night_day	Verhältnis c_we_night / c_we_weekend
49	Verhältnisse	r_we_morning_noon	Verhältnis c_we_morning / c_we_noon
50	Verhältnisse	r_we_evening_noon	Verhältnis c_we_morning / c_we_noon
51	Verhältnisse	r_wd_night_day	Verhältnis c_wd_night / c_wd_weekend

52	Verhältnisse	r_wd_morning_noon	Verhältnis c_wd_morning / c_wd_noon
53	Verhältnisse	r_wd_evening_noon	Verhältnis c_wd_morning / c_wd_noon
54	Zeit	t_above_1kw	Zeitpunkt der ersten Überschreitung der 1kW-Grenze, (Gemittelt über alle Wochentage)
55	Zeit	t_above_2kw	Zeitpunkt der ersten Überschreitung der 2kW-Grenze, (Gemittelt über alle Wochentage)
56	Zeit	t_above_mean	Anzahl der Datenpunkte über dem Wochenmittel, (In der gesamten Woche)
57	Zeit	t_daily_max	Zeitpunkt der ersten Erreichens des Tages-Maximum, (Gemittelt über alle Wochentage)
58	Zeit	t_daily_min	Zeitpunkt der ersten Erreichens des Tages-Minimum, (Gemittelt über alle Wochentage)
59	Zeit	ts_stl_varRem	Mittlerer Rest bei Saison- / Trend- Dekomposition
60	Zeit	ts_acf_mean3h	Mittlere AutokorVerhältnisse (über eine Zeitspanne von 3h)
61	Zeit	ts_acf_mean3h_weekday	Mittlere AutokorVerhältnisse (über eine Zeitspanne von 3h) an Wochentagen
62	Zeit	smart_diff	Differenz zwischen Wochentagen +- 30 min
63	Zeit	weak_diff	Differenz zwischen Wochentagen +- 30 min (schwache version)
64	Zeit	number_big_peakss	Anzahl der Peaks nach einer groben Glättung
65	Zeit	width_peaks	Durchschnittliche Ausprägung des Peaks
66	Zeit	sm_variety	20\%-Quartil der Verteilung der Abweichung zum vorhergehenden Messwert
67	Zeit	bg_variety	60\%-Quartil der Verteilung der Abweichung zum vorhergehenden Messwert
68	Zeit	const_Zeit	Geschätzte Zeit der Grundlast
69	Zeit	value_min_guess	Geschätzte Grundlast
70	Zeit	first_above_base	Erstes Überschreiten einer als Grundlast angenommenen Grenze
71	Zeit	t_above_base2	Anzahl der Messpunkte über der Grundlast-Grenze
72	Zeit	Zeit_above_base2	Anzahl der Messpunkte über der Grundlast-Grenze (Alternative Berechnung)
73	Zeit	percent_above_base	Anteil der Messpunkte über der Grundlast-Grenze
74	Zeit	value_above_base	Summe der Messpunkte über der Grundlast-Grenze
75	Zeit	b_day_diff	Abweichung der Messwerte an Wochentagen
76	Zeit	b_day_weak	Schwache Version von b_day_diff
77	Zeit	dist_big_v	Abstand zwischen hohen Werten
78	Statistische	s_variance	Varianz
79	Statistische	s_var_wd	Varianz an Wochentagen
80	Statistische	s_var_we	Varianz an Wochenendtagen
81	Statistische	s_diff	Summe der Differenzen zum Vorgänger (Betrag)
82	Statistische	s_cor	KorVerhältnisse zwischen Tag 1 und Tag 2 der Woche
83	Statistische	s_num_peaks	Anzahl der Spitze (lokales Maximum bei der Betrachtung von drei Messwerten)
84	Statistische	s_q1	Unteres Quartil
85	Statistische	s_q2	Median
86	Statistische	s_q3	Oberes Quartil
87	Statistische	c_max_avg	Mittleres Tagesmaximum
88	Statistische	c_min_avg	Mittleres Tagesminimum
89	Statistische	n_d_diff	Nacht/Tag Differenz
90	Statistische	number_zeros	Anzahl von Null-Werten
91	Statistische	c_sm_max	Maximum bei einfacher Glättung
92	Statistische	s_cor_wd	Mittlere KorVerhältnisse zwischen den Wochentagen
93	Statistische	s_cor_we	KorVerhältnisse zwischen Sa und So
94	Statistische	s_cor_wd_we	KorVerhältnisse zwischen Wochentagen und Wochenendtagen
95	Statistische	number_small_peaks	Anzahl grosser Spitzen an Wochentagen (Berechnung mittels gleitendem gewichtetem Mittelwert