Praktikum Data Mining

Energieverbrauch und CO2-Emmisionen
Vorhersage und Clustering auf Finanzdaten

Oliver Fesseler Maria Florusß Stefan Seibert
Daniel Grießhaber

8. Mai 2014

Inhaltsverzeichnis

Energieverbrauch und CO_2-Emmission

Datenverwaltung und Statistik

Einlesen der Daten, Hinzufügen der GPS Koordinaten, Abspeichern in neuer Datei

Bei der Umsetzung der Aufgaben haben wir mit verschiedenen Darstellungsformen der Daten experimentiert. Im ersten Plot fallen vor allem die Vielverbraucher einzelner Energieformen auf. Beim zweiten Plot können die verschiedenen Energiemixe pro Land direkt miteinander verglichen werden, da sie alle im selben Maßstab nebeneinander dargestellt werden.

Ausgehend von der implementierten Visualisierung des Energieverbrauchs der Länder: Nennen Sie die 3 Ihrer Meinung nach interessantesten Beobachtungen.

1. Durch die wenigen Industrieländer mit signifikant höherem Energieverbrauch, wie China oder die USA, wird der Plot so verzerrt, dass die Länder mit durchschnittlichem Verbrauch im Plot so gestaucht werden, dass sie kaum zu erkennen sind. Dies könnte behoben werden, wenn die Daten mit den Einwohnerzahlen aller Länder normalisiert werden würden. So könnte der Pro-Kopf-Verbrauch berechnet werden, was einen besseren Vergleich der einzelnen Länder bietet.

2. Dieses Prinzip wird klar bei der Betrachtung von China und Indien, die von der Einwohnerzahl her vergleichbar sind (China 1,3 Mrd., Indien 1,2 Mrd.[1]). China zeigt einen weitaus höheren Verbrauch als Indien an den in beiden Ländern häufigen Energieformen Kohle und Öl. Zusätzlich wären noch andere Normalisierungsfaktoren interessant:

    - Bruttoinlandsprodukt
    - Außenhandelsstatistik oder Export der Länder in US$

[1] Stand 2012, Quelle: Wikipedia

- Technologieindex

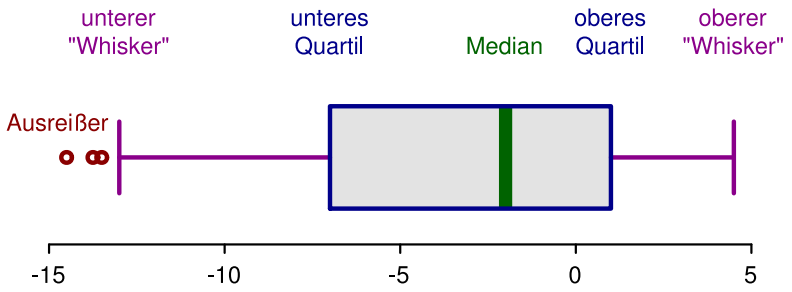
3. Die zwei Länder mit dem höchsten Energieverbrauch sind China und die USA. Dies fällt bei der Betrachtung des Gesamtenenergieverbrauchs auf. Bei der Betrachtung des Verbrauchs einzelner Energieformen wirkt es als sei China durch seinen hohen Kohleverbrauch weit vor den USA.

Abgabe: Relevante Dateien

- `energy_consumption_per_country.py` und `energy_consumption_per_country_V2.py`
  - Implementierung Aufgabe 2.1.2: 1) - 2)
- `energy_consumption_per_country.pdf`
  - Ausgabe des Skripts `energy_consumption_per_country.py`
- `energy_consumption_per_country_V2.pdf`
  - Ausgabe des Skripts `energy_consumption_per_country_V2.py`
- `appendGeoCoordinates.py`
  - Implementierung Aufgabe 2.1.2: 3) - 5)
- `EnergyMixGeo.csv`
  - Ausgabe des Skripts `appendGeoCoordinates.py`

Statistik der Daten

Erklären Sie sämtliche Elemente eines Boxplot (allgemein).



- Ausreißer (“Outliers“), in Python mit ’sym’ zu setzen.
  Daten die außerhalb der Whisker liegen und somit als Ausreißer deklariert werden.

- Whisker, Länge in Python mit 'whis' zu setzen.
  Standartmäßig die 1,5-fache Länge des entsprechenden Quartils.
- Quartil
  Die Quartile sind Bestandteile der Box, welche 50 % aller Daten enthält. Dabei enthält das obere Quartil, die 25%, die über dem Median, das untere Quartil, die 25%, die unter dem Median liegen.
- Median
  Der mittlere Wert (nicht Mittelwert oder Durchschnitt), der aus dem gesamten Datensatz ermittelt wird. Er teilt den Boxplot in zwei Hälften, die wiederum jeweils in Whisker und Quartil unterteilt werden.

Diskutieren Sie die im Boxplot angezeigte Statistik der Energieverbrauchdaten.

- Durch den signifikant höheren Verbrauch von China und den USA wird beim Anzeigen der Ausreißer der restliche Boxplot soweit gestaucht, dass vernünftigen Daten mehr abgelesen werden können. Um die Verteilung innerhalb der verschiedenen Boxplots besser vergleichen zu können, entschlossen wir uns außerdem jede Energieform in einem einzelnen Subplot darzustellen.
- Im Gesamtplot lassen sich die einzelnen Energieformen gut miteinander vergleichen, in den einzelnen Plots können die Verteilungen innerhalb der Plots besser dargestellt werden.
- Beim Boxplot der nuklearen Energieform fällt auf, dass der Median samt unterem Quartil und unterem Whisker auf 0 liegt. Dies liegt daran, dass weniger als die Hälfte aller Länder nukleare Energie verwenden.
- Im Gesamtplot kann man erkennen, dass Öl die einzige Energieform ist, deren unterer Whisker nicht auf 0 liegt. Also verwendet jedes in Betracht gezogene Land Öl. Jede andere Energieform wird von mindestens einem Land nicht verwendet.

Abgabe: Relevante Dateien

- `enegryStatistics.py`
  - Implementierung Aufgabe 2.1.3: 1) - 3)
- `energyconsumption_by_energyform_in_seperate_subboxplots.pdf` und `energyconsumption_by_energyform_in_one_plot.pdf`
  - Ausgabe des Skripts `enegryStatistics.py`

Anwendung von Verfahren des unüberwachten Lernens auf Energieverbrauchsdaten

Hierarchisches Clustering

Was wird beim Standardisieren gemacht? Welcher Effekt könnte ohne Standardisieren beim Clustering eintreten (insbesondere wenn die euklidische Metrik verwendet wird)?

Ohne Standardisieren ist die Ähnlichkeit des Gesamtverbrauchs ausschlaggebender als die Ähnlichkeit des Energiemixes. So werden eher Länder mit niedrigem Gesamtverbrauch gruppiert, als sie anhand ihres Energiemixes zu clustern. Dies kommt daher, dass die euklidische Metrik die geometrische Distanz zwischen zwei Punkten im Mehrdimensionalen Raum berücksichtigt. Zeigen zwei Vektoren in dieselbe Richtung, sind sich demnach vom Energiemix sehr ähnlich, haben aber unterschiedliche Längen, also einen unterschiedlich hohen Energieverbrauch, dann haben sie auch eine hohe euklidische Distanz und werden nicht demselben Cluster zugeordnet. Durch das Standardisieren werden die Längen der Vektoren normiert und so ein Vergleich erst möglich.

Erkären Sie die beim hierarchischen Clustering einstellbaren Parameter `linkage-method` und `metric`. Welche Metrik ist Ihrer Meinung nach für diese Anwendung geeignet? Warum?

Über die `linkage-method` wird festgelegt, wie die Cluster hierarchisch angeordnet werden. Dabei kann zwischen verschiedenen Methoden gewählt werden: Beispielsweise kann die mittlere (average), die kleinste (single) oder die größte (complete) Distanz zweier

Punkte aus beiden Clustern, oder die Distanz der beiden Clusterschwerpunkte (weighted), gemessen werden. Die `metric` bestimmt die Ähnlichkeit zwischen zwei Punkten. Hierbei kann zwischen verschiedenen Metriken gewählt werden, wobei in unserem Fall Ähnlichkeitsmaße für boolsche Werte vernachlässigt werden können. Die Ähnlichkeit kann über die euklidische Distanz gemessen werden, allerdings wird hier nur die geometrische Distanz gewertet und nicht die Richtung. Weiterhin kann Cosinus-Distanz angewandt werden, bei der die Richtung der Vektoren mehr ins Gewicht fällt als die Länge. Bei der Pearson-Ähnlichkeit wird zusätzlich die Durchschnittslänge miteinbezogen und mit der tatsächlichen Länge verrechnet. Dadurch erweist sich dieser Algorithmus als am besten geeignet.

Welches Land ist bezüglich des Verbrauchs der hier betrachteten Energiequellen Deutschland am ähnlichsten, wenn für die linkage-method `average` und die Metrik `correlation` konfiguriert wird?

Antwort: Belgien

Charakterisieren Sie die 4 Cluster. Was ist typisch für die jeweiligen Cluster?

- Cluster 0 ist einigermaßen gleichverteilt, wobei Wasserkraft den geringsten Anteil ausmacht. Im Vergleich zu den anderen Clustern ist der Anteil an aus Atomkraft gewonnener Energie sehr hoch.
- Cluster 1 zeichnet einen großen Verbrauch an fossilen Brennstoffen aus.
- Cluster 2 ist das kleinste Cluster und unterscheidet sich vor allem durch seinen hohen Kohleverbrauch von den anderen Clustern.
- Cluster 3 verzeichnet im Gegensatz zu allen anderen Clustern einen relativ hohen Wasserkraftanteil.

Fasst man den gesamten Energieverbrauch jedes jeweiligen Clusters zusammen (`individual_clusters_total.pdf`), erkennt man, dass die Charakteristik des Gesamtverbrauchs eines Clusters sehr stark von einzelnen Ländern abhängt. So diktieren China

und die USA den Gesamtverbrauch in ihrem Cluster. Gleichzeitig zeigt sich allerdings auch die Tendenz des gesamten Clusters.

Abgabe: Relevante Dateien

- `energyClustering.py`
  - Implementierung Aufgabe 2.2.1: 1) - 5)
- `dendrogram.pdf`, `individual_clusters.pdf` und `individual_clusters_total.pdf`
  - Ausgabe des Skripts `energyClustering.py`

Dimensionalitätsreduktion

Welches Land ist nach dieser Darstellung Deutschland am ähnlichsten?

Antwort: Südkorea

Warum entspricht die hier dargestellte Ähnlichkeit nicht der im oben erzeugten Dendrogram?

1. Da wir die Darstellung selbst optisch interpretiert und die Distanz zwischen den Punkten als Ähnlichkeitsmaßverwendet haben, haben wir die Ähnlichkeit nach der euklidischen Metrik bestimmt.
2. Auch im Dendrogramm waren sich Deutschland und Südkorea relativ ähnlich. Durch die Reduktion der Dimensionen von fünf auf zwei gehen zwangsweise Informationen verloren.

Abgabe: Relevante Dateien

- `energyReduceDim.py` - Implementierung Aufgabe 2.2.2: 1) - 3)
- `energyReduceDim_total.pdf`- Ausgabe des Skripts `energyReduceDim.py`
- `energyReduceDim_section.pdf`- Ausschnitt aus `energyReduceDim_total.pdf`

Überwachtes Lernen: Schätzung der CO_2-Emmission

Feature Selection

Welche 3 Merkmale haben den stärksten Einfluß auf das Ausgabemerkmal CO_2-Emmission? Wie groß sind die vom Programm ausgegebenen Scores?

Kohle	378.266881
Öl	220.010151
Wasser	79.045401
Gas	46.002230
Nuklear	34.572086

Abgabe: Relevante Dateien

- `energyFeatureSelection.py` - Implementierung Aufgabe 2.3.1: 1) - 3)

Regression mit Epsilon-SVR

Optimieren Sie die SVR-Parameter C und Epsilon so dass der Score in der Kreuzvalidierung minimal wird. Welche Werte für C und Epsilon liefern das beste Ergebnis?

Die besten Ergebnisse erhielten wir für C = 0.01 und $\varepsilon = 0.001$

Für das SVR-Objekt können die Koeffzienten der linearen Abbildung, welche durch die trainierte SVR realisiert wird, ausgegeben werden: `meineSVR.coef_`**. Notieren Sie diese Koeffzienten für die beste SVR.**

Öl	Gas	Kohle	Nuklear	Hydro
-3.0690410	-2.3485549	-3.9608432	$4.1970815e-04$	$4.1138445e-04$

Welchen Aufschluss geben diese Koeffzienten über den Einfluss der einzelnen Eingangsmerkmale auf das Ausgangsmerkmal?

Die Koeffizienten geben an, wie sehr die entsprechende Energieform Einfluss auf die CO_2-Emmission hat. Öl, Kohle und Gas haben demnach einen sehr viel größeren Einfluss als Energie aus Atom- und Wasserkraft.

Wie groß ist die mittlere absolute Differenz zwischen Soll- und Ist-Ausgabe für die beste SVR? Diskutieren Sie dieses Ergebnis.

Für die optimierten Parameter C = 0.01 und $\varepsilon = 0.001$ ergibt sich ein Mean Absolute Error (MAE) von 0.119259989310.

C	ε	MAE
1	0.01	0.119938469138
1	0.001	0.119995514827
1	0.0001	0.119986240023
1	0.1	0.124915412379
0.1	0.01	0.119776387503
0.01	0.001	0.119259989310
0.001	0.0001	64.902638179800

Der erhaltene Wert für den MAE ist für realistische Daten viel zu klein, was ein eindeutiger Hinweis darauf ist, dass die verwendeten Ausgangsdaten selbst mit einem ähnlichen Algorithmus berechnet wurden. Im Diagramm (`energyPrediction.pdf`) kann deshalb zwischen vorhergesagter und tatsächlicher Ausgabe nicht unterschieden werden, da beide Kurven genau übereinander liegen.

Abgabe: Relevante Dateien

- `energyPrediction.py` - Implementierung Aufgabe 2.3.2: 1) - 7)
- `energyPrediction.pdf` - Ausgabe des Skripts `energyPrediction.py`

Visualisierung des Clusterings in Google Maps

Abgabe: Relevante Dateien

- `cluster2GoogleMaps.py` - Implementierung Aufgabe 2.4
- `clusterMap.html` - Ausgabe des Skripts `cluster2GoogleMaps.py`