11. November 2013

DATA MINING

- Versuch 1: Energieverbrauch und CO2-Emmission -

Teammitglieder

JanHorak

Ralf Palyov

Andreas Mayer

2.1 Datenverwaltung und Statistik

2.1.2 GPS Koordinaten

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

Interessant zu beobachten ist, dass der Wasserkraft-Anteil Canadas und Brasiliens höher ist als der der USA, obwohl die Gesamtmenge der USA um ein vielfaches höher ist. Der Kohleanteil Chinas scheint auf den ersten Blick in etwa so hoch zu sein wie der Gesamtkohleanteil aller anderen Länder. Die arabischen Länder verzichten nahezu komplett auf Kohle- Wasser- und Atomkraft.

Von implementierungstechnischer Seite aus betrachtet gestaltet sich der Zugriff auf die Geo-Daten über die Google Maps-API als sehr einfach. Auch das erzeugen eines Diagramms ist ohne großen Aufwand möglich.

2.1.3 Statistik der Daten

1 Erklären Sie sämtliche Elemente eines Boxplot (allgemein).

Der Boxplot ist eine grafische Darstellung, die die Verteilung gemessener Werte (Messniveau) hilft abzubilden.

Grundsätzlich wird bei einem Boxplot in fünf - in manchen Fällen auch sechs Bestandteile unterschieden.

Zwei davon sind die "obere" und "untere" Antenne. Diese sind in der Länge an die Menge der Daten gebunden, wobei "Aussreißer"

ausgenommen sind. Der Median (->liegt genau in der Mitte der Messdaten) des Boxplots wird von zwei anderen Bestandteilen umschlossen: dem oberen und dem unteren Quantil.

Diese dadurch enstandene Box ist ein Maß für die Streuung der gemessenen Daten.

2 Diskutieren Sie die im Boxplot angezeigte Statistik der Energieverbrauchdaten.

Öl weist eine "relativ hohe" Verbrauchsstatistik auf. Begründet wird das durch die relativ hohe Box und den hohen Verbrauchswerten.

Das passt sehr gut zu der ersten Üblerlegungen, dass Öl eine hohe Verbrauchsstatistik besitzen muss.

Es lässt sich ausserdem beobachten, dass durch die sich am oberen Quantil befindende Whisker

relativ viel Spielraum gibt. Das für den Öl- Verbrauch errechnete Maximum ist also nicht "nah" an den durchschnittlichen

Verbrauchswerten.

Der nukleare Energieverbrauch hat eine noch geringere Verbrauchsstatistik - hier ist allerdings deutlich zu sehen, dass es mehr

Aussreißer gibt (diese liegen ausserhalb des oberen / unteren Whiskers), als bei Öl.

Vielleicht ist das zu Begründen mit den in den vergangen Jahren zu beobachten "Kehrtwende" in den Atompolitiken der Welt.

Ähnlich, wie bei dem Verbrauch von nuklearer Energie ist auch der Verbrauch von Wasser zu beobachten. Dort ist auch,

beispielsweise im Vergleich zu Öl, eine niedrige Verbrauchsstatistik zu erkennen.

Kohle und Gas ähneln erwartungsgemäß dem Verbrauch von Kohle sehr stark.

Das deutet auf die Korrektheit der bestimmten Daten hin - Kohle, Öl und Gas werden weltweit am meisten verbraucht,

nukleare Energie hat zwar eine im Vergleich relativ geringe Verbrauchsstatistik, dort sind aber auffällig viele Aussreißer

erkennbar. Wasser wird im Vergleich am wenigsten verbraucht.

2.2 Anwendung von Verfahren des unüberwachten Lernens auf Energieverbrauchsdaten

2.2.1 Hierarchisches Clustering

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

Die Bildung von Clustern erfolgt durch den Vergleich aller Eigenschaften der einzelnen Objekte miteinander. Die Eigenschaften können allerdings in ihrer Skalierung stark voneinander abweichen: z.B. besitzen die Objekte folgende Eigenschaften:

- Maximale Geschwindigkeit in km/h
- Leergewicht in kg

Es sollen zwei Objekte verglichen werden:

Eigenschaft	Objekt 1	Objekt 2
Maximale Geschwindigkeit in km/h	320	160
Leergewicht in kg	1642	7490

Bei der euklidischen Metrik würde die Distanz zwischen den Werten direkt ermittelt werden. Das hat zur Folge, dass der Unterschied des Leergewichts viel mehr Gewicht hat als der Unterschied der max Geschwindigkeit.

Leergewichtsunterschied (5848 kg) > max Geschwindigkeitsunterschied (160 km/h)

Ein Unterschied in der max Geschwindigkeit würde bei einem Clustering so keinen Ausschlag geben und das Ergebnis wäre stark verfälscht.

Der Vergleich sollte also auf Daten durchgeführt werden die zuvor so angepasst wurden, dass unterschiedliche Bezugsgrößen keine Rolle mehr spielen. Dies ist die Standardisierung: Der Mittelwert einer Eigenschaft über alle Objekte wird ermittelt. Dieser Mittelwert wird dann von den einzelnen Eigenschaften abgezogen. Das Ergebnis wird abschließend durch die Standardabweichung der einzelnen Eigenschaften dividiert.

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

„metric“ gibt an welche Metrik zur Distanz zwischen zwei Objekten verwendet werden soll. Der Standard ist euklidisch, die Gefahr die sich durch diese Metrik ergibt ist in Frage 1 beschrieben. Es ist sinnvoll die selbe Metrik zu verwenden wie in der Berechnung der Objekt Distanzen ('correlation')

„linkage-method“ gibt an wie die Distanz zwischen den Clustern berechnet werden soll. D.h. wie berechnet wird, welche Objekte zu welchem Cluster gehören.

3 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?

Belgien ist Deutschland am ähnlichsten.

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

1. Cluster	Hat einen relativ ausgewogenen Energiemix, keine der Energiequellen sticht besonders hervor oder ist auf 0.
2. Cluster	Die Energiequellen Gas und Öl liegen hier deutlich im Vordergrund.
3. Cluster	Hier dominiert Kohle.
4. Cluster	Bei diesen Ländern wird die Wasserkraft sehr stark genutzt.

2.2.2 Dimensionalitätsreduktion

1 Welches Land ist nach dieser Darstellung Deutschland am ähnlichsten?

Die euklidische Distanz von Deutschland ist zu Südkorea am geringsten. Dies hängt sicherlich damit zusammen, dass sowohl das Verhältnis der Energieanteile als auch die Gesamtmenge der beiden Länder sehr ähnlich sind.

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

Bei der hier dargestellten Ähnlichkeit spielt der Gesamtenergieverbrauch eine wichtige Rolle. Länder, die zwar verhältnismäßig nahezu identisch sind, sich aber vom Gesamtverbrauch stark unterscheiden werden trotzdem weit voneinander entfernt dargestellt. Je größer der Unterscheid der Gesamtmenge ist, desto weiter sind auch die Punkte voneinander entfernt. Beim hierarchischen Clustering werden diese Ähnlichkeiten berücksichtigt und verhältnismäßig ähnliche Länder in Cluster zusammengefasst.

2.3 Überwachtes Lernen: Schätzung der CO2-Emmission

2.3.1 Feature Selection

1 Welche 3 Merkmale haben den stärksten Einfluss auf das Ausgabemerkmal CO2-Emmission? Wie groß sind die vom Programm ausgegebenen Scores?

Erwartungsgemäß haben Gas, Kohle und Öl die höchsten Scores.

Scores:

	Co2-Emissions
Hydro	3585.719875
Nuclear	5159.411960
Gas	11600.056857
Oil	15683.709785
Coal	51882.621866

2.3.2 Regression mit Epsilon-SVR

1 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?

C ≈ 0.046

epsilon ≈ 0.125

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

Oil: -3.069

Gas: -2.348

Coal: -3.961

Nuclear: -0.0007287

Hydro: -0.001065

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

Die Koeffizienten geben Aufschluss, wie stark die einzelnen Eingangsmerkmale für die Ausgabe relevant sind. Es wird deutlich, dass die Merkmale 'Nuclear' und 'Hydro' im Gegensatz zu den anderen sehr nahe am Wert 0 liegen. Dies bedeutet, dass diese zwei Merkmale einen so geringen Einfluss auf das Ausgangsmerkmal haben, dass sie vernachlässigt werden könnten.

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

Die mittlere absolute Differenz beträgt gerundet 0.129.

Dieser Wert ist sehr gut, das liegt daran, dass die SVR mit allen Daten trainiert wurde. Die Daten mit denen die Differenz ermittelt werden, sind bereits für das Training verwendet worden. Für realistischere Ergebnisse müssten neue Daten herangezogen werden.