Visuelle Evaluation von Big Data Analysen mit Heatmaps

Bachelorarbeit

zur Erlangung des akademischen Grades „Bachelor of Science (B.Sc.)“im Studiengang Wirtschaftswissenschaften der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von:

Name: Tappert [redacted] [redacted]

Vorname: Robin [redacted] [redacted]

Prüfer: Prof. Dr. Michael H. Breitner

Hannover, den 09.02.2017

Inhaltsverzeichnis

1 Einleitung

"Tell me and I forget, teach me and I may remember, involve me and I learn." - Benjamin Franklin

In der heutigen Gesellschaft befinden wir uns in einer Phase, in der Individuen geradezu permanent Daten preisgeben und wichtige Entscheidungen immer häufiger auf Informationen basieren, die durch diese freiwillige Offenlegung generiert werden können. Grund für diese Datenflut ist dabei nicht nur die Einstellung der Menschheit gegenüber Datenschutz sowie deren stark ausgeprägtes Mitteilungsbedürfnis, sondern auch die eifrige Sammellust der Unternehmen. So lassen sich immer mehr Unternehmen finden, die „Trillionen Bytes an Informationen über ihre Kunden, Zulieferer und ihren Betrieb erfassen"[Manyika et al., 2011, 1], wobei dadurch ein enormes Optimierungspotential zur Verfügung stehen könnte. Es stellt sich dabei die Frage, wie diese Datenflut, nachdem sie in einem langwierigen Prozess zunächst gesammelt, aufbereitet und an die zuständigen Bereiche verteilt wurde, in einer angemessenen und anschaulichen Art analysiert wird.

Hierbei könnten geeignete Visualisierungsmethoden erfolgsversprechende Ansätze liefern, da diese die Daten nicht nur hinsichtlich möglicher Zusammenhänge und Verteilungen analysieren sowie detaillierte Vergleiche unter Methoden und Datensätzen durchführen könnten, sondern gleichzeitig auch in der Lage sind, diese Analysen und Vergleiche in einer adäquaten Art zu veranschaulichen. Womöglich könnte es auch gelingen, Entscheidungsträgern die Problemstellung so darzulegen, sodass sie direkt in die Auswertung und gegebenenfalls auch in weitere Modifikationen des zugrunde liegenden Modells involviert werden. Ganz nach dem Leitsatz Benjamin Franklins könnte es somit gelingen, die Entscheidungsträger nicht nur dazu zu bringen, sich lediglich an das Vorgehen zu erinnern, sondern dieses Vorgehen auch durch die direkte Beteiligung am Prozess zu erlernen. Auch erste Anzeichen von Anomalien wie Nichtlinearitäten können bereits durch geeignete Visualisierungsmethoden aufgedeckt und durch die Expertise der Manager bekräftigt werden. Derartige Anomalien stehen zumeist für eine zuvor fälschlich angewandte Methode [Eilers and Breitner, 2016], sodass die Verwendung anderer maschineller Lernmethoden durchaus einen Vorteil generieren könnte.

Im Folgenden werden diese Methoden durch ein Visualisierungstool in Form einer Heatmap evaluiert, die auf einem Smoothing mittels eines Epanechnikov-Kerns basiert und dadurch Punkten in einem Gitter, das zuvor zur Illustrierung aufgebaut wurde, eine Gewichtung verleiht. Es wird zunächst versucht, die Annahme eines linearen Zusammenhangs zu bestätigen, in dem eine einfache, lineare Regression auf eine Problemstellung angewandt wird und daraufhin die resultierenden Residuen durch eine Heatmap dargestellt werden, wodurch der Anwender mittels Mustererkennung in der Heatmap auf eine vorliegende Nichtlinearität schließen könnte. Normalerweise obliegt diese Funktion konventionellen Verfahren wie dem Plotten eines 'Residual

Plots', der Residuen auf die durch das Modell angepassten Werte abträgt. Bei anderen Verfahren aus dem maschinellen Lernbereich wird für gewöhnlich ebenfalls der 'Residual Plot' betrachtet, nur dass es dabei zum Problem kommt, wenn die Daten entweder in ihren Observationen oder in ihren Parametern Dimensionen annehmen, die weit über das hinausgehen, was ursprünglich für diese Art von Methoden gedacht war. Bei ungünstigen Anwendungen kann es so dazu kommen, dass bei großen Datensätzen die Bearbeitungszeit des Rechners zu stark ausgeweitet wird [Feng et al., 2010, 986], sodass auch hier alternative Methoden, die die Datensätze vereinfachen und diese leichter interpretierbar machen, immer mehr in den Fokus rücken könnten, um die Daten letztendlich in einer angemessen Zeit bearbeiten und auswerten zu können, wobei im Idealfall nach einer Auswertung in Echtzeit eine unmittelbare Anpassung der Geschäftsprozesse ermöglicht werden könnte. Allerdings haben Berechnungen häufig Komplexitäten höherer Ordnung im Vergleich zu der Anzahl der Daten, sodass Echtzeit-Visualisierungen schwer zu realisieren sind [Choo and Park, 2013, 22].

Falls die Vermutung nach einer Evaluierung naheliegt, dass zwischen zwei Variablen ein nichtlinearer Zusammenhang besteht, ist es sinnvoller, andere maschinelle Lernmethoden wie die polynomiale Regression, Spline-Funktionen oder auch künstliche, neuronale Netzwerke anzuwenden, da diese für deutlich mehr Flexibilität stehen und dadurch in der Lage sind, komplexere Zusammenhänge besser zu approximieren, was allerdings wiederum auf Kosten der Interpretierbarkeit dieser Modelle geschieht [James et al., 2015, 25]. Die Performance der angewandten Modelle könnten per Heatmap evaluiert werden, um in Erfahrung bringen zu können, ob der Wechsel der Methode zu Fortschritten hinsichtlich der Approximation eines Testdatensatzes beigetragen hat.

Auch bei hochdimensionalen Problemstellungen, bei denen die Anzahl der Parameter zum Teil weit über die der Observationen hinausgeht, werden Methoden aus dem maschinellen Lernbereich wie lineare Modell-Selektion per 'Forward-Stepwise' Selection oder auch Variablen-Selektion per 'Shrinkage'-Methode wie dem LASSO-Verfahren verwendet, um die Parameteranzahl zu dezimieren und damit Modelle zu erzeugen, deren Interpretation deutlich einfacher sein dürfte als die der ursprünglichen Modelle. Im Anschluss daran wird auch hierbei die Performance der neuen, aus den Methoden resultierenden Modelle mit dem Modell verglichen, das Verwendung gefunden hätte, wenn der einzelne, unabhängige Parameter mittels einer multiplen Regression bzw. kleinstem p-Wert ausgesucht worden wäre, um letztendlich Kenntnisse darüber zu erhalten, ob die Methode aus dem Bereich der Variablen-Selektion einen Fortschritt in bestimmten Bereichen generieren kann und ob sie den Anwender von vornherein davor warnt, nicht die falschen Parameter bei der Dezimierung der Parameteranzahl auf Grundlage der multiplen Regression zu streichen.

Das Hauptziel der Arbeit ist es, die Erhöhung der Vorhersagegenauigkeit durch Anwendung

verschiedenster maschineller Lernmethoden herbeizuführen, indem diese nacheinander mittels Heatmaps evaluiert werden, wobei lediglich Regressionsprobleme unter die Lupe genommen werden. Zunächst wird dabei in Kapitel 2 auf die Visualisierung mittels Heatmap und deren mathematische Herleitung eingegangen, sodass diese in Kapitel 3 sogleich Anwendung auf eine lineare Regression finden können. In Kapitel 4 wird der Nichtlinearität Beachtung geschenkt, woraufhin es anschließend in Kapitel 5 um die Schwierigkeiten bei einer hochdimensionalen Datenlage geht. Kapitel 6 widmet sich dann der abschließenden Diskussion über die Performance der Heatmaps, wodurch in Kapitel 7 ein Fazit gefunden sowie ein Ausblick in die Zukunft gegeben werden kann.

7 Fazit und Ausblick

In der Arbeit wurde die Frage behandelt, ob es möglich ist, Methoden aus dem maschinellen Lernbereich mittels einer Heatmap zu evaluieren und darüber hinaus detailliert die Schwachstellen aufzudecken, die die angewandte Methode mit sich bringt. Im Fall der linearen Regression war dies sogleich erfolgreich, da deutliche Muster in den Heatmaps erkennbar wurden, wodurch auf eine Nichtlinearität geschlossen werden konnte, die durch weitere Methoden deutlich besser approximiert werden konnte. Somit trat der Vorteil von Heatmaps gegenüber anderen Evaluationskriterien wie der RSS deutlich hervor, da bei Letzterer keinerlei Informationen darüber gegeben werden kann, in welchen Bereichen das Modell schlecht approximiert bzw. ob das Modell überhaupt geeignet ist, die Daten gut zu approximieren. Auch waren Heatmaps gut dafür geeignet, Vergleiche zwischen Methoden anzustellen, sodass herausgefunden werden konnte, dass das angewandte Neuronale Netz und die lineare Regression nahezu identisch in sämtlichen Bereichen approximierte.

Es besteht viel Raum für weitere Forschungen in diesem Bereich, da neben der hier betrachteten 'Basis-Spline'-Funktion, die sich auf Basisfunktionen stützt, viele weitere 'Spline'-Funktionen wie bspw. 'Natural Splines' existieren, die weitere Eigenschaften wie robusteres Verhalten an Rändern mit sich bringen. In dieser Arbeit wurden auch lediglich Regressionsprobleme betrachtet, während auch Methoden für die Anwendung auf Klassifikationsprobleme wie die Hauptkomponentenanalyse, die 'Support Vector Machines', Diskriminanzanalysen oder wiederum künstliche neuronale Netzwerke mittels Heatmaps evaluiert werden könnten.

Visualisierungen werden so in immer mehr Bereichen verwendet [Potter et al., 2012, 240], sodass auch zukünftig eine immer größere Anzahl an Entscheidungsträgern diese verstehen müssen .