Vergleich von k-Nächster-Nachbar und Random Forest für eine Klassifizierung in deutsche Städte auf Basis von Grundstücksattributen

Ben Reher
Matrikelnr.: 3318023
in Zusammenarbeit mit Konrad Schumacher

20. Januar 2022

Zusammenfassung

Dieses Projekt vergleicht die Fehlerrate des k-Nächster-Nachbar und des Random Forest Klassifikators für die Erkennung zehn deutscher Großstädte. Jedem der 60.000 Grundstücke aus dem synthetischen Datensatz wurden Werte für sechs Attribute zugewiesen. Experimentelle Ergebnisse zeigen, dass beide Klassifikatoren ähnlich gut in der Zuweisung sind. Die erfassten nur minimal Verschiedenen Ergebnisse und Fehlerstrukturen deuten darauf hin, dass der erzeugte Datensatz eine bessere Klassifizierung nicht möglich macht.

OTH Regensburg

Künstliche Intelligenz und Data Science
Anwendungsorientierte Grundlagen der Künstlichen Intelligenz

1 Empirische Ergebnisse

Das Ziel dieses Projektes ist es die beiden Klassifikatoren Random Forest (RF) und k-Nächster-Nachbar (kNN) mit Hilfe des Machine Learning Tools *Weka*[1] auf ihre Generalisierungsfähigkeit zu testen und anschließend zu vergleichen.

1.1 Daten

Bei dem Datensatz, der für den Vergleich der beiden Klassifikatoren verwendet wurde, handelt es sich um einen synthetischen Datensatz, der mithilfe eines Programms in C erstellt wurde und insgesamt 60.000 Grundstücke aus zehn verschiedenen deutschen Großstädten enthält. Den Grundstücken wurden jeweils sechs charakterisierende Attribute zugewiesen. Diese Attribute werden im Folgenden definiert:

Preis
Das Attribut Preis wird als Preis pro Quadratmeter in Euro ausgewiesen. Es handelt sich hierbei bei allen Grundstücken um die Quadratmeterpreise aus dem Jahr 2021. [2]

Internetgeschwindigkeit
Das Attribut der Internetgeschwindigkeit enthält, die maximal am jeweiligen Grundstück abrufbare Datenübertragungsrate in Megabit pro Sekunde im Jahr 2017. [3]

Flughafendistanz
Die Flughafennähe des Grundstücks beschreibt die Distanz vom Grundstück bis zum nächstliegenden Flughafen in Kilometern. Dieser Wert ist auf ist auf ganze Kilometer gerundet. [4]

Grünflächenanteil
Der Grünflächenanteil eines Grundstücks beschreibt den Anteil der Grünfläche zur Gesamtfläche im Umkreis von einem Kilometer um das Grundstück im Jahr 2016. [5]

Luftbelastung
Die Luftbelastung eines Grundstücks beschreibt den Stickstoffdioxidwert in Mikrogramm pro Kubikmeter auf dem jeweiligen Grundstück im Jahr 2020.[6]

Preisentwicklung
Die Preisentwicklung gibt die Steigerung des Preises pro Quadratmeter eines Grundstücks in der Zeit von 2020 bis 2021 in Prozent an. [7]

Da es sich bei diesem Vergleich um einen synthetischen Datensatz handelt sind die darin enthaltenen Grundstücke nicht real. Die spezifischen Werte für einen Datenpunkt wurden mittels eines Computerprogramms erstellt. Die Basis für die synthetisch generierten Werte sind Mittelwerte sowie Minimal- und Maximalwerte, die die realen Verhältnisse in den Städten widerspiegeln. Die Werte wurden auf zwei verschiedene Wege erzeugt:

Normalverteilung um den Durchschnitt
Die Internetgeschwindigkeiten, Grünflächenanteile, und die Luftbelastungen wurden mithilfe von recherchierten Mittelwerten aus den jeweiligen Städten erzeugt. Hierzu wurden die Werte mit einer festgelegten Varianz um den Mittelwert normalverteilt.

Intervall
Die Attribute des Preises und der Flughafendistanz wurden mit einer zufälligen Zuweisung innerhalb eines gegebenen Intervalls generiert. Hierzu wurde jeweils der Minimal- und Maximalwert für die jeweilige Stadt ausfindig gemacht und hieraus dann mit einem „Zufallsgenerator“ innerhalb dieses Intervalls ein Wert ausgewählt.

1.2 Experimentelles Protokoll

Zuerst wurde der Datensatz in eine Testmenge mit 10.000 Grundstücken und eine Trainings - und Validierungsmenge mit 50.000 Grundstücken geteilt. Anschließend wird mit der größeren Menge fortgefahren. Nach dem Hochladen in das Programm *Weka* muss zuerst das Klassenattribut y mit Hilfe der integrierten Filters von einem numerischen in einen nominellen Wert umgewandelt werden. Nach erfolgter Umwandlung müssen die Attributwerte x1 - x6 noch standardisiert werden, da sich diese in verschiedenen Größenordnungen befinden und damit eine Verfälschung des Ergebnisses verhindert wird. Beginnend mit einem der beiden Klassifikatoren kann mit Hilfe eines 80% Percentage Split (40.000/10.000) in einzelne Validierungs- und Trainingsmenge das beste k für den kNN mit der euklidischen Distanz bestimmt werden. Auf der Validierungsmenge wird k zwischen 1 - 1000, mit inverser Gewichtung oder ohne experimentell getestet.
Für den RF wurden die Iterationsgrößen zwischen 1 und 300 getestet und

die *Number of Features* von 0 - 10 variiert.
Die Testmenge muss ebenfalls angepasst und standardisiert werden und kann dann als Testmenge auf die zusammengefasste Trainings- und Validierungsmenge angewandt werden. Hierbei werden die besten ermittelten Einstellungen benutzt.

1.3 Resultate und Diskussion

Bei dem RF erwies sich die Iterationsgröße bei 250 und eine Number of Features von 2 am besten. Hierbei wurde ein Generalisierungsfehler von 13,76% berechnet. Der kNN schneidet etwas schlechter ab, bei $k = 75$ und inverser Gewichtung wurden 14,63% auf der Testmenge erreicht. Die Klassifikatoren erreichen damit ein nur um etwa 1% zueinander abweichendes Ergebnis. Es zeigen sich lediglich Unterschiede in der aufgewandten Zeit. Der kNN baut das Modell in unter einer Sekunde und braucht für das Testen insgesamt 25 Sekunden, der RF hingegen braucht für das Erzeugen des Modells alleine 27 Sekunden und für das Testen nur fünf, somit insgesamt 32 Sekunden.
Was sich daraus schließen lässt ist schwierig zu sagen, jedoch steht fest, dass es sich hierbei um die bestmögliche Quote auf diesem Datensatz handeln könnte, da einige Attribute sehr hohe Überschneidungen haben und somit die Klassifizierung erschweren.
Ein weiterer Faktor, der betrachtet werden kann ist die Art der Fehler die jeder Klassifikator macht. München beispielsweise wurde bei beiden Klassifikatoren zu 100% richtig Klassifiziert, auch Stuttgart und Leipzig konnten sehr gut bestimmt werden. Köln hingegen bei nur 64,7% (RF) und 61,3% (kNN) der Testgrundstücke. Hier gibt es also Unterschiede, welche auf Ähnlichkeit der Werte für bestimmte Städt zurückzuführen sind.
Ein solches Ergebnis, mit durchgehend ähnlichen Ergebnissen, welche auch beim Ausprobieren aus Testzwecken an anderen Klassifikatoren nicht merklich besser geworden sind, deuten darauf hin, dass ein besseres Ergebnis kaum erzielt werden kann. Das liegt an der Ähnlichkeit einiger Attribute. Frankfurt und Köln im Vergleich zum Beispiel haben in sämtlichen Kategorien kaum Abweichungen zueinander.
Jedoch ist ein solches Modell in jedem Fall besser, als ein Mensch, da dieser aus den reinen Zahlen wahrscheinlich nichts herauslesen könnte.

2 Zusammenfassung und Ausblick

Es lässt sich also zusammenfassend festhalten, dass sie die Klassifikatoren im Endergebnis kaum unterscheiden. Lediglich die kürzere Laufzeit des kNN gegenüber dem RF ist aufgefallen. Bei einem Projekt mit größeren Datensätzen wäre somit als Wahl des Klassifikators der einfachere kNN brauchbarer, wohingegen der RF bei kleineren Problemen mit der besseren Klassifizierung punktet. In Zukunft könnte ein solches Problem in erweiterter Form in einem realen Anwendungsfall beispielsweise bei Verkaufsplattformen im Internet auftreten, eine solches Klassifizierungsproblem wäre also denkbar. Hierfür müsste jedoch das begrenzende Attribut gefunden werden, um das Problem größer zu skalieren.

Referenzen

[1] Machine Learning Tool Weka 3. (2021). https://www.cs.waikato.ac.nz/ml/weka/

[2] Immobilienpreise 2022 - Preise abfragen. (2021). McMakler Immobilienpreise. https://www.mcmakler.de/immobilienpreise

[3] Verivox. (15. Oktober, 2018). Durchschnittliche Internetgeschwindigkeit in Großstädten in Deutschland im Jahr 2017 (in Mbit/s) [Graph]. In Statista. Zugriff am 23. Januar 2022, von https://de.statista.com/statistik/daten/studie/939412/umfrage/internetgeschwindigkeit-im-staedtevergleich-in-deutschland/

[4] Entfernungsrechner - Entfernung berechnen und darstellen. (2021). https://www.luftlinie.org/

[5] Berliner Morgenpost. (10. Mai, 2016). Anteil der Grünfläche deutscher Großstädte* im Jahr 2016 [Graph]. In Statista. Zugriff am 23. Januar 2022, von https://de.statista.com/statistik/daten/studie/417098/umfrage/deutschlands-gruenste-staedte/

[6] Umweltbundesamt. (17. Mai, 2021). Deutsche Städte mit den durchschnittlich höchsten Stickstoffdioxidwerten pro Kubikmeter Luft in den Jahren 2019 und 2020 [Graph]. In Statista. Zugriff am 23. Januar 2022, von https://de.statista.com/statistik/daten/studie/954311/umfrage/staedte-mit-den-hoechsten-stickstoffdioxidwerten-in-deutschland/

[7] Grundstückspreise 2022: Aktuelle Preise in Deutschland! (2022). experten-beraten24. https://experten-beraten24.de/grundstueckspreise/

Anhang

KNN	München	Berlin	Hamburg	Köln	Stuttgart	Frankfurt	Leipzig	Düsseldorf	Dortmund	Bremen
München	1022	0	0	0	0	0	0	0	0	0
Berlin	0	817	57	13	0	46	0	37	5	9
Hamburg	0	70	864	27	1	11	0	13	5	5
Köln	0	17	24	742	0	150	0	45	4	0
Stuttgart	0	0	0	0	971	0	0	0	2	12
Frankfurt	0	59	11	219	1	615	0	76	1	1
Leipzig	0	0	0	0	1	0	1013	0	0	0
Düsseldorf	0	69	29	127	0	133	0	679	0	0
Dortmund	0	1	1	1	4	0	2	0	920	78
Bremen	0	10	0	0	18	0	1	0	67	894
FALSCH	0	226	122	387	25	340	3	171	84	105

RF	München	Berlin	Hamburg	Köln	Stuttgart	Frankfurt	Leipzig	Düsseldorf	Dortmund	Bremen
München	1022	0	0	0	0	0	0	0	0	0
Berlin	0	805	68	9	2	53	0	40	0	7
Hamburg	0	48	889	25	1	13	0	15	2	3
Köln	0	13	19	731	1	153	0	61	4	0
Stuttgart	0	0	1	0	973	2	0	0	0	9
Frankfurt	0	52	15	201	1	631	0	81	0	2
Leipzig	0	0	0	0	1	0	1013	0	0	0
Düsseldorf	0	57	23	117	0	109	0	731	0	0
Dortmund	0	0	3	1	0	0	1	0	939	63
Bremen	0	12	2	0	12	1	0	0	73	890
FALSCH	0	182	131	353	18	331	1	197	79	84