Prognose einflussreicher User in sozialen Netzwerken

Proseminar Practical Data Mining

Florentin Wieser
Fakultät für Informatik
Technische Universität München
f.wieser@tum.de

Zusammenfassung

Um Werbekampagnen in Sozialen Netzwerken möglichst effizient zu gestalten ist es nötig, einflussreiche User in Sozialen Netzwerken zu bestimmen. In diesem Paper wird die menschliche Entscheidung in einem direkten Vergleich zweier Twitter-Accounts vorhergesagt. Dazu wird zuerst die Definition von Einfluss in Sozialen Medien betrachtet. Anschließend wird ein Generalisiertes Lineares Modell erstellt. Es wird festgestellt, dass die Variablen 'following_count', 'mentions_sent' und 'posts' eine geringe Signifikanz besitzen, während 'follower_count' und 'listed_count' sehr relevant sind.

1. Aufgabenstellung

Das Ziel der Aufgabe ist die Bestimmung des einflussreicheren Accounts in einem direkten Vergleich von zwei Twitter-Accounts. Als Grundlage hierfür gilt das menschliche Urteil. Zur Lösung des Problems soll ein maschinelles Lernmodell erstellt werden, das die Entscheidung möglichst gut vorhersagt.
Ein denkbares Anwendungsgebiet für ein solches Modell ist das Marketing in Sozialen Netzwerken.

1.1. Gegebene Daten

Es sind drei Datensätze zur Lösung der Problematik gegeben: 'Train', 'Test' und 'Sample Predictions'.

Train beinhaltet die Grundlage für das lineare Modell. Enthalten sind 5.500 Datensätze mit jeweils 23 nicht-negativen Attributen. Eines davon, genannt 'Choice', repräsentiert die binäre Entscheidung des Menschen. Dabei bedeutet der Wert '1', dass Account A einflussreicher als Account B gewertet wurde und '0' das Gegenereignis. Weitere Attribute sind: Followercount, Followingcount, Listedcount, Mentionssent, Mentionsreceived, Retweetssent, Retweetsreceived, Posts sowie Networkfeature 1, 2 und 3. Diese Attribute sind jeweils für die Accounts A und B vorhanden. Da die Networkfeatures nicht trivial zu berechnen sind, hat der Steller der Daten keine Informationen über deren Bedeutung veröffentlicht.

Test enthält 5.952 Datensätze mit, bis auf das Entscheidungsmerkmal, den gleichen Attributen wie der Train-Datensatz. Mithilfe dieser Daten kann das erstellte Modell zur Ermittlung des einflussreicheren Users getestet werden.

Sample Predictions umfasst für jede Zeile der Testdaten einen 'Choice'-Wert. Dieser wurde in Python mithilfe eines linearen Modells und den Testdaten gewonnen. Mit diesen Daten kann das eigene Modell auf Übereinstimmung mit der Lösung überprüft werden.

2. Einfluss in Social Media

Bevor versucht wird den einflussreicheren zweier User zu bestimmen, muss erst einmal klar sein was genau Einfluss heißt und wie er in sozialen Netzwerken auftritt.

> Einfluss, der
> beeinflussende, bestimmende Wirkung auf jemanden, etwas; Einwirkung [Duden]

Wenn jemand also einflussreich ist, ist er in der Lage andere in Ihrer Meinung und Wahrnehmung zu bewegen. Beispiele für Einflussnehmer gibt es unzählige: Politiker, Journalisten, bekannte Persönlichkeiten und viele mehr.

Doch wie lässt sich diese Definition auf Beobachtungen in Sozialen Netzwerken anwenden?

> [...] is the influence that these individuals wield, which is determined by the actual propagation of their content through the network. This influence is determined by many factors, such as the novelty and resonance of their messages with those of their followers and the quality and frequency of the content they generate. [Romero et al. 2011]

Ein User generiert also Einfluss, wenn seine Nachrichten im Netzwerk weiterverbreitet werden. Wichtige Faktoren für das Maß an Einfluss sind die Neuartigkeit und die Qualität der Nachricht, aber auch die Häufigkeit der Posts sowie die Reaktionen des Publikums.

Frühere Arbeiten, die sich mit der Identifizierung von Charakteristika einflussreicher Nutzer beschäftigen, kamen zu der Theorie, dass wenige informierte, respektierte und vernetze User die anderen Netzwerkteilnehmer beeinflussen. Falls die richtigen Personen adressiert werden, kann folglich mit niedrigen Marketingkosten ein großer Effekt erzielt werden [Cha et al. 2010].

Andere Arbeiten kamen zu dem Schluss, dass diese 'Influentials' oder 'Opinion Leaders' nicht so ausschlaggebend sind wie sie scheinen. So wurde herausgefunden, dass Trends eher dem Zustand einer leicht beeinflussbaren Masse geschuldet sind als dem Einfluss eines Einzelnen [Watts & Dodds 2007].

Zusammengefasst gibt es also zwei Arten von einflussreichen Social-Media-Nutzern:

- Die 'Opinion Leaders', also aktive Benutzer des Netzwerkes, welche Botschaften weiterverbreiten. Diese Art von Benutzern sind generell überzeugungsfähig, gut vernetzt und nehmen rege an der Community teil.

- Die sogenannten 'Accidential Influencers', welche zufällig Einfluss generieren. Dies geschieht indem sie zur richtigen Zeit eine Nachricht verbreiten, die die kritische Masse der User anspricht und so verbreitet wird.

3. Einfluss auf Twitter

Wird nun speziell Twitter betrachtet, wurde darauf geschlossen, dass 'Accidential Influencers' die Ausnahme der Einflussnehmer sind. Vielmehr wurde herausgefunden, dass sehr aktive User, die persönliches Engagement zeigen und ihre Tweets auf ein Thema spezialisieren, diejenigen sind die Einfluss ausüben [Quercia et al. 2011]. Zu einem ähnlichen Ergebnis kommt auch [Cha et al. 2010].

Zu dem Thema der Quantifizierung von Einfluss auf Twitter gibt es eine große Arbeit. In ihr betrachteten die Autoren [Cha et al. 2010] die drei Messpunkte Followeranzahl, Anzahl der Retweets und die Zahl der Mentions. Dabei wurden jeweils folgende Bedeutungen zugeordnet:

- **Followerzahl** beschreibt die Bekanntheit des betrachteten Users.
  Die höchsten Followerzahlen haben überwiegend Prominente und Newsseiten [Twittercounter]

- **Zahl der Retweets** verweist auf den Wert des weitergeleiteten Inhalts.
  Content-Aggregation-Services haben die meisten Retweets

- **Mentions** werden als Wert des Namens, auf den Verwiesen wird, gewertet.
  Die Mentionstärksten Accounts sind wieder bekannte Personen

Die wichtigste Erkenntnis ist, dass zwischen hohem Followercount und großem Einfluss nur eine geringe Korrelation besteht. In dieser Arbeit wurde Einfluss über die Reaktionen des Publikums definiert, also die Anzahl der Retweets und der Mentions.

Im Gegensatz zu dem kaum vorhandenen Zusammenhang zwischen Bekanntheit, also der Follower-Anzahl, und dem Einfluss des Users, gibt es allerdings einen großen Zusammenhang zwischen der Anzahl der Retweets und der Mentions.

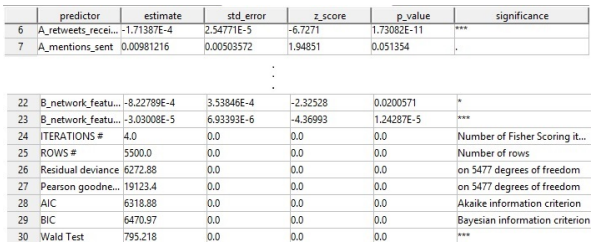
Eine weitere Beobachtung ist, dass der Retweet die größte Form des Einflusses auf Twitter ist [Qianni & Yunjing 2012]. Dies ist auch im Vergleich mit den anderen Kommunikationsformen, wie Antworten und Likes, einfach zu erkennen, da der User die Information übernimmt und weiterverbreitet.

4. Erstellen eines Linearen Modells

Nun wird ein lineares Modell zur Vorhersage des menschlichen Urteils in einem direkten Vergleich erstellt. Dabei wird das Tool 'GLM' benutzt.
Abbildung 1 zeigt den SQL-Code zur Erstellung des linearen Modells. In Columnnames wird festgelegt, dass 'Choice' die abhängige Variable ist und alle anderen Variablen, hier mit 'Vars' als Platzhalter abgekürzt, zur Vorhersage benutzt werden. Family beschreibt die Art der betrachteten Variablen. In diesem Fall fällt die Wahl auf 'Logistic', da die abhängige Variable 'Choice' nur zwei mögliche Endergebnisse hat: 0 oder 1. Mithilfe

```
SELECT * FROM GLM(
ON (SELECT 1)
PARTITION BY 1
INPUTTABLE('influencertrain')
OUTPUTTABLE('glmoutput')
COLUMNNAMES('Choice', 'Vars')
FAMILY('LOGISTIC')
WEIGHT('1')
THRESHOLD('0.01')
MAXITERNUM('10'));
```

Abbildung 1. SQL-Code zur Erstellung des Modells



	predictor	estimate	std_error	z_score	p_value	significance
6	A_retweets_recei...	-1.71387E-4	2.54771E-5	-6.7271	1.73082E-11	***
7	A_mentions_sent	0.00981216	0.00503572	1.94851	0.051354	.
⋮						
22	B_network_featu...	-8.22789E-4	3.53846E-4	-2.32528	0.0200571	*
23	B_network_featu...	-3.03008E-5	6.93393E-6	-4.36993	1.24287E-5	***
24	ITERATIONS #	4.0	0.0	0.0	0.0	Number of Fisher Scoring it...
25	ROWS #	5500.0	0.0	0.0	0.0	Number of rows
26	Residual deviance	6272.88	0.0	0.0	0.0	on 5477 degrees of freedom
27	Pearson goodne...	19123.4	0.0	0.0	0.0	on 5477 degrees of freedom
28	AIC	6318.88	0.0	0.0	0.0	Akaike information criterion
29	BIC	6470.97	0.0	0.0	0.0	Bayesian information criterion
30	Wald Test	795.218	0.0	0.0	0.0	***

Abbildung 2. Modelltabelle mit Parametern zur Modellgenauigkeit

Treshold und Maxiternum lassen sich Grenzen zum Abbruch des Algorithmus festlegen.

Die erstellte Modelltabelle (Abbildung 2) enthält für alle Variablen, die für die Prognose verwendet wurden, die Signifikanz der Variable, einen Erwartungswert und die Standardabweichung. Außerdem werden mehrere Größen zur Anpassungsgüte angegeben (Zu sehen in Abbildung 2 ab Zeile 26). Diese Variablen beschreiben die Genauigkeit des Modells.

Als nächster Schritt wird mithilfe der Daten der Testfile und des erstellten Modells prognostiziert, welcher der beiden betrachteten User einflussreicher ist. Das ganze lässt sich mithilfe der 'GLMPredict'-Funktion bewerkstelligen. Dieser wird die Modelltabelle und die Daten, auf die diese angewendet werden soll, übergeben. Weitere Parameter sind noch die in die neue Tabelle zu übernehmenden Spalten, sowie die Family mit der das Modell erstellt wurde.

Das Ergebnis dieser Funktion ist eine Schätzung der 'Choice'-Variable zu jeder Zeile aus der Testfile. Diese Schätzung befindet sich im Intervall $(0,1)$. Um nun eine Aussage über die Wahl des einflussreicheren Users treffen zu können, werden alle Werte < 0.5 dem Wert 0 und allen Werten > 0.5 dem Wert 1 zugeordnet.

Werden die Werte der Sample-Predictions nach gleichem Schema angepasst, so lässt sich die eigene Lösung und die Beispiellösung auf Übereinstimmung untersuchen. Allerdings sollte der Fokus auf ein anderes Maß gelegt werden, da die Lösung ja auch nur ein mögliches Modell beschreibt. In dieser Arbeit wird primär der BIC betrachtet, da dieser als Ziel eine möglichst genaue Theorieprüfung hat [Garnand, 2009]. Dabei gilt: je niedriger der BIC ist, desto besser passt das Modell.

Mithilfe des ersten Modells ist eine Übereinstimmung von 5.008 der 5.952 Testdaten und ein BIC von 6.470 aufgetreten. Um die Übereinstimmung und die Anpassungsgüte zu erhöhen, wird dieses Vorgehen mit verschiedenen Ausgangssituationen wiederholt.

5. Iterationen zur Verbesserung des Modells

Um eine erste Verbesserung zu erreichen, werden nicht signifikante Variablen weggelassen. Aus der Output-Tabelle des ersten linearen Modells geht hervor, dass 'A_posts','B_posts', 'A_mentions_received', 'B_mentions_received' und 'B_following_count' eine geringe Signifikanz haben. Da 'following_count' allerdings nur für B unsignifikant ist, werden für das nächste Modell die Variablen 'posts' und 'mentions_received' außer Acht gelassen.
Das Ergebnis ist eine Verringerung des BIC auf 6.440 und es werden 5.003 der Samplepredictions richtig vorhergesagt. Also wurde das Modell in Hinsicht auf den BIC verbessert, hat sich jedoch von der Beispiellösung entfernt.

Als nächste Verbesserung wird die Differenz aller Beobachtungen von A und B betrachtet. Um dies zu erreichen, wurde eine neue Tabelle mit den Differenzen der elf verschiedenen angegebenen Attributen erstellt. Dabei wurden immer die Datenpunkte von B von den jeweiligen Daten von A abgezogen. Um eine fehlerfreie Arbeit zu gewährleisten, muss die gleiche Transformation auch auf die Testdaten angewendet werden, bevor das Modell benutzt wird.
Die Auswertung dieses Modells ergibt eine Übereinstimmung mit der Musterlösung von 5.048 Datensätzen und einen BIC von 6.403. Also wurde das Modell, wenn auch nur geringfügig, verbessert.

Da in der vorherigen Iteration eine Verbesserung des Modells beobachtet wurde, wenn gering signifikante Werte weggelassen werden, werden nun wieder die Anzahl der 'posts' und die Anzahl der 'mentions_received' weggelassen.
Nun ist eine weitere Senkung des BICs zu sehen, es fällt auf 6.389. Gleichzeitig ändert sich die Zahl der gleichen Prognosen minimal auf 5.047.

Eine weitere Idee zur Verbesserung des allgemeinen linearen Modells ist, wie auf Kaggle in der Lösung

zu sehen, den Logarithmus jedes Beobachtungspunktes zu ziehen und dann aus diesen Werten die Differenz zu bilden.
Das Ergebnis dieses Durchlaufs ist eine erhebliche Verringerung des betrachteten Informationskriteriums auf 5.127. Des weiteren ist die Anzahl der Übereinstimmungen mit den Samplepredictions auf 5.891 gestiegen. Somit prognostiziert das erstellte Modell 99% der Entscheidungen gleich wie das Kaggle-Modell. Dies ist darauf zurückzuführen, dass nun die gleiche Transformation der Daten verwendet wurde.

Da das Entfernen der wenig signifikanten Variablen bisher immer einer Verbesserung erbrachte, wird diese Prozedur nun auch auf die Differenz der Logarithmen angewendet. Zunächst werden nur die Variablen 'following' und 'mentions_sent' entfernt. Da damit eine Verbesserung auf einen BIC von 5.113 erreicht wurde, wurde auch die kaum relevante Variable 'posts' entfernt. Als Ergebnis ist die betrachtete Anpassungsgröße auf 5.106 gesunken. Das weitere Gütekriterium AUC hat sich in den letzten Schritten nicht geändert.
Um ein Underfitting, also eine Verschlechterung des Modells durch das Weglassen von relevanten Variablen, zu verhindern, wird an diesem Punkt die Iteration beendet.
Bei der Durchführung der letzten Schritte sinkt natürlich die Übereinstimmung der Vorhersage des eigenen Modells mit der Vorhersage der Musterlösung. Allerdings zeigt der niedrigere BIC, dass das Modell nicht schlechter geworden ist sondern sich verbessert hat.

Werden alle Modelle der Iterationen verglichen, fällt auf, dass die Variablen 'follower_count' und 'listed_count' immer eine hohe Signifikanz haben. Gleichzeitig bleibt der Beobachtungspunkt 'following_count' immer auf einem niedrigen Relevanzniveau.
Dieses Ergebnis lässt sich auch durch einfache Plots überprüfen (Siehe Abbildung 3). In den beiden Plots wurden die relevanten Beobachtungen geplottet und zu jedem Punkt die jeweilige Entscheidung zugeordnet. Dabei entspricht eine orange Markierung der 'Choice' 1, also der Entscheidung A sei einflussreicher als B, und vice versa. Dazu sind noch die Trendlinien für die beiden Entscheidungen eingezeichnet. Da in beiden Plots die orange Trendlinie stärker wächst als die blaue, ist zu sehen, dass diese beiden Variablen starken Einfluss auf die Entscheidung haben.
Wird nun Abbildung 4 betrachtet, ist kaum ein Unterschied zwischen den beiden Geraden zu sehen. Sie verlaufen beinahe Parallel zur x-Achse. Daraus lässt sich schließen, dass das hier betrachtete Merkmal, 'following_count', minimal signifikant ist.
Diese Beobachtungen stimmen somit mit den Ergebnissen aus den Modellen überein. Im Gegensatz zu dem Fazit von [Cha et al. 2010] ist in dem erstellten Modell die Anzahl der Follower von hoher Bedeutung für die Feststellung des einflussreicheren Users. Außerdem ist die Zahl der Retweets nur mittelmäßig signifikant. Daraus lässt sich ableiten, dass der menschliche Eindruck von Einfluss sich nur begrenzt in der Definition von [Cha et al. 2010] widerspiegelt.

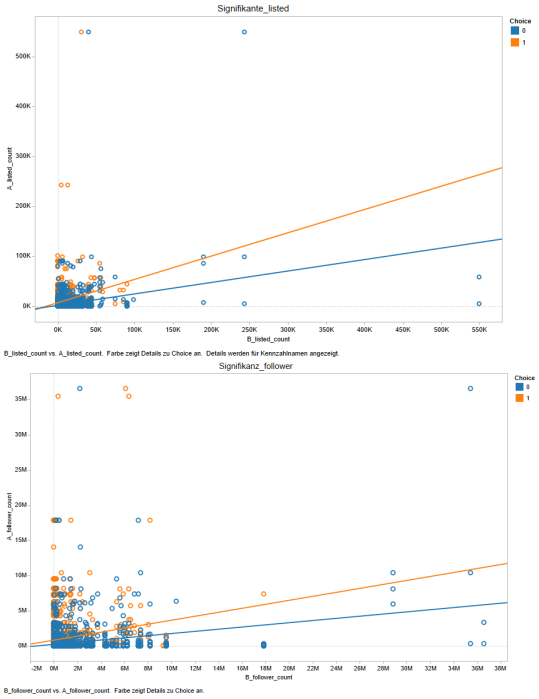


Abbildung 3. Plot und Trendlinien der relevanten Attribute Listed_count und Follower_count

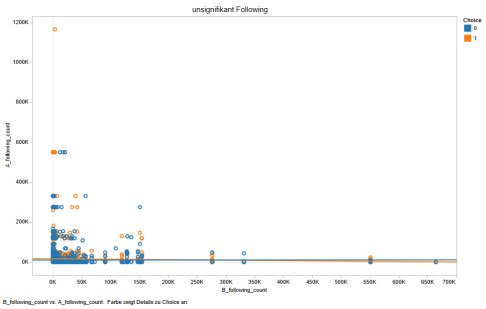


Abbildung 4. Plot und Trendlinie der unsignifikanten Variable 'posts'

6. Grenzen bei der Vorhersage von einflussreichen Usern

Allerdings gibt es bei der Vorhersage des menschlichen Urteils mit einem linearen Modell Grenzen. So wird der Mensch hinter einem Account lediglich auf einen Knoten in einem Netzwerk herunter gebrochen.
Besonders in den hier verwendeten Daten spielen viele Faktoren eines Tweets keine Rolle - es werden hauptsächlich Eckdaten von Accounts genutzt, um den Einfluss zu bestimmen. In einer anderen Arbeit wurde die Auswirkung der Sprache auf den gewonnen Einfluss untersucht. Es wurde herausgefunden, dass ein Zusammenhang zwischen dem Einfluss von Nutzern sozialer Netzwerke und deren sprachlicher Ausdrucksweise besteht [Quercia et al. 2011].

Ein weiterer Punkt ist die Passivität der Benutzer. Viele Nutzer belassen es beim Lesen einer Nachricht und Verbreiten diese nicht weiter. Diese Passivität stellt ein Maß für die Schwierigkeit Einfluss zu generieren dar und wird in diesem Modell nicht berücksichtigt [Romero et al. 2011].

7. weitere Analyseideen

Aus den Grenzen der Vorhersage lassen sich nun Ideen ableiten um das Modell zu verbessern und weiterzuentwickeln. So könnte für jeden betrachteten Benutzer einen Wert für die durchschnittliche Anzahl an Reaktionen pro Follower erstellt werden.
$AVG\_Reactions = \frac{Mentions\_received + Retweets\_received}{Follower\_count}$
Mit diesem Wert fließt die Passivität des Publikums eines Nutzers mit in die Wertung ein. Laut den Erkenntnissen von [Romero et al. 2011] ist die Passivität ein Faktor für den Einfluss eines Users. Ein hoher Wert zeugt also von reaktionsfreudigen Followern und erhöht so die Chancen des Nutzers Einfluss zu generieren.

8. Zusammenfassung

In dieser Arbeit wurde ein generalisiertes lineares Modell zur Vorhersage des einflussreicheren Twitter-Benutzers in einem direkten Vergleich erstellt. Dafür wurden zwei Arten von einflussreichen Social-Media-Nutzern untersucht. Dabei wurde festgestellt, dass auf Twitter aktive, überzeugungsfähige, gut vernetzte und aktiv an der Community teilnehmende User diejenigen sind, die Einfluss ausüben.
Beim Erstellen des Modells wurde den Logarithmus jeder Beobachtung gezogen und anschließend die Differenz der betrachteten User gebildet. Außerdem war auffällig, dass die Variablen 'following_count', 'mentions_sent' und 'posts' unsignifikant sind. Mit dem Weglassen dieser Variablen wurde das Modell verbessert.

Literatur

[Qianni & Yunjing 2012] Qianni, D, Yunjing, D 2012, *How Your Friends Influence You: Quantifying Pairwise Influences on Twitter*, 2012 International Conference on Cloud Computing and Service Computing

[Cha et al. 2010] Cha, M, Haddadi, H, Benevenuto, F & Gummandi, KP,*Measuring User Influence in Twitter: The Million Follower Fallacy*, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

[Quercia et al. 2011] Quercia, D, Ellis, J, Capra, L, & Crowcroft, J 2011, *In the Mood for Being Influential on Twitter*, 3rd IEEE International Conference on Social Computing

[Watts & Dodds 2007] Watts, DJ, Dodds, P 2007, *The Accidental Influentials*, Harvard Business Review

[Romero et al. 2011] Romero, DM, Wojciech, G, Sitaram, A & Huberman, BA 2011,*Influence and Passivity in Social Media* , WWW '11 Proceedings of the 20th international conference companion on World wide web

[Duden] http://www.duden.de/rechtschreibung/Einfluss, Aufgerufen am 5.6.16

[Twittercounter] http://twittercounter.com/pages/100, Aufgerufen am 16.6.16

[Garnand, 2009] Gernand, L, Fenske, N 2009, *Understanding AIC and BIC in Model Selection*, Handreichung zum Vortrag vom 20. Mai