Clustern von Eye-Trackingdaten zur Unterstützung bei der Früherkennung von Dyslexie

Mario Kaulmann,[1] Herval Nganya,[1]

[1]University of Applied Sciences, Technische Hochschule Brandenburg,
Magdeburger Straße 50, 14770 Brandenburg an der Havel, Deutschland

In dieser Arbeit werden Versuchspersonen mittels Eye-Trackingdaten geclustert. Diese Versuchspersonen sollten bei der Datenerfassung drei Versuche nacheinander durchführen. Bei diesen Versuchen sollte ein Punkt mit dem Blick verfolgt werden. Bei der Clusterung soll sich herausbilden, wie gut die Versuchspersonen diese Aufgabe gelöst haben. Die Bemessung der Güte der Cluster erfolgt mit Hilfe des Silhouettenkoeffizienten. Die Cluster sollen bei der Früherkennung von Dyslexie zum Einsatz kommen.

Einführung

Dyslexie zeigt sich durch signikante Schwierigkeiten Wörter schnell und korrekt zu lesen und zu verstehen (*Handler and Fierson, 2011,Siegel, 2006*). In einem Experiment wurden Eyetracking-Daten erhoben, bei denen die Versuchspersonen drei verschiedene Versuche durchführen sollten. Dabei sollten die Versuchspersonen mit den Augen einem Punkt folgen, der eine spezielle Figur zeichnete. Diese Figuren sind eine liegende Acht und eine horizontale Linie. Für jeden Versuch wurden zwei Durchläufe gemacht. Pro Durchlauf wurde die entsprechende Figur zwei mal gezeichnet. Für die liegende Acht langsam wurde zusätzlich vorher ein Probedurchlauf gemacht, bei dem die Figur nur einmal gezeichnet wurde. Die Tabelle 1 zeigt die Versuche, die durchgeführt wurden.
Die Versuchspersonen sollen so geclustert werden, dass die entstehenden Cluster zur Klassizierung neuer Daten genutzt werden können. Diese Cluster sollen zur Entscheidungsfindung beitragen, ob ein Kind an Dyslexie leidet.

Table 1: Liste der Versuche: Hier wird die Reihenfolge der Versuche angegeben, sowie die Figur, die der Punkt gezeichnet hat und die Dauer eines Durchlaufs.

Reihenfolge	**Figur**	**Dauer**
1	liegende Acht langsam	8 Sekunden
2	liegende Acht schnell	4 Sekunden
3	horizontale Linie	4 Sekunden

Daten

Die erhobenen Eye-Trackingdaten sind Zeitreihen. Zu 302 Versuchsperson gibt es je eine Datei mit den Blickpunktdaten der Person und eine Datei mit den Koordinaten des Punkts, der verfolgt werden sollte. Von den Blickpunktdaten einer Person sollten nur die Blickposition des linken und des rechten Auges genutzt werden. Die anderen Daten, sollten vernachlässigt werden. Mittels dieser Zeitreihen werden personenbezogene Merkmale erzeugt, die zur Clusterung der Versuchspersonen genutzt werden können. Abbildung 1 zeigt, die Rohdaten einer Durchführung der liegenden Acht. Daraus wird deutlich, dass die Koordinatensysteme zueinander verschoben sind. Die Verschiebung beträgt 640 px in X-Richtung und 512 px in Y-Richtung. Um die Abstände zu den Targetpunkten berechnen zu können, wurden die Targetpunkte entsprechend verschoben.

Methoden

Das Vorgehen untergliedert sich in vier Phasen. Die erste ist die explorative Analyse. Dabei werden die Daten auf Besonderheiten und ihre Wertebereiche untersucht. Die zweite Phase ist die Merkmalsgenerierung, dabei werden aus den Zeitreihen Merkmale zur Beschreibung der Versuchspersonen generiert. In der dritten Phase werden die Cluster erstellt und die Güte mittels Silhouettenkoeffizient bestimmt.

Explorative Analyse

Bei der explorativen Analyse wurden die Zeitstempel der zu einer Person gehörenden Dateien untersucht, wobei festgestellt wurde, dass die Anzahl, der gleichen Zeitstempel gegen Null geht. Darum wurde das Zusammenführen der beiden Dateien so gestaltet, dass immer der Targetpunkt mit dem Blickpunkt zusammengeführt wurde, der den nächsten größeren oder gleichen Zeitstempel hat.
Außerdem wurden die Werte der Blickpunktkoordinaten untersucht. Dabei fiel auf, dass diese entweder Null sind oder einen von Null unterschiedlichen positiven Wert haben. Die Null hat die Bedeutung, dass die Augenposition nicht gemessen werden konnte. Zum Beispiel, wenn das Auge geschlossen war, oder zu weit weg war vom Messsensor. Bei drei Versuchspersonen (VP71, VP90 und VP136) viel auf, dass für diese nur ein Auge gemessen wurde. Aus diesem Grund wurden diese beiden Versuchspersonen nicht berücksichtigt. Desweiteren wurde Versuchsperson 131 nicht berücksichtigt, da diese in den Messdaten Eventeinträge enthielt, die nicht zu den Standardeventeinträgen gehören.

Merkmalsgenerierung

Die neu erzeugten Merkmale sind in Merkmale innerhalb der Zeitreihen und Merkmale zur Beschreibung einer Versuchsperson zu unterscheiden. Die Merkmale innerhalb der Zeitreihen sind die Mittelwerte der Augenpositionen und die Abstände (euklidischer Abstand) der Augenpositionen zum Targetpunkt, außerdem die Geschwindigkeiten der Augen.
Bei den Eigenschaften, die eine Versuchsperson beschreiben handelt es sich um statistische Kenngrößen, wie das arithmetische Mittel, das Maximum und die Varianz, der Zeitreihenwerte für die einzelnen Versuche einer Versuchsperson. Die weiteren Eigenschaften werden im folgenden genauer beschrieben.
In einem Versuch mit 55 Kindern mit Dyslexie und je 55 Kindern im gleichen Alter und 55 Kindern mit der selben Lesestufe, aber ohne Dyslexie, wurde gezeigt, dass Kinder mit Dyslexie Probleme bei der Fixation haben (*Tiadi et al., 2016*). Aus diesem Grund vermuten wir, dass das Merkmal der Abweichung zum Targetpunkt ein aussagekräftiges Merkmal ist.
Wichtige Merkmale für Augenbewegungen sind Fixationen und Sakkaden. Sakkaden sind dadurch gekennzeichnet, dass sich die Augen schnell bewegen und über einem Geschwindigkeitsschwellwert liegen (*Holmqvist et al., 2011, p. 152*). Die Abbildung 2 zeigt für eine Versuchsperson den Verlauf der Augenbewegungen der mittleren Augenposition. Damit wurde das Merk-

mal erzeugt, wie viele Sakkaden es pro Durchlauf und pro Auge gab. Außerdem wurde die Rate der Sakkaden für jeden Versuch und jedes Auge ermittelt.
Als weitere Eigenschaft wurde ausgewertet, wie hoch der Anteil der echten Messungen ist, also der Messungen der Augenposition, die nicht Null sind. Dieser Wert könnte Auskunft darüber geben, wie gut sich die Versuchsperson konzentrieren konnte, oder ob es von der Aufgabe überfordert war. Vorausgesetzt, dass die Technik in Ordnung war, was wir als sehr wahrscheinlich einschätzen.
Insgesamt wird eine Versuchsperson durch 117 Merkmale beschrieben, die im whrend dieser Arbeit erzeugt wurden.

Clustern

Zum Clustern wurde der Algorithmus K-Means eingesetzt. Der Algorithmus kann in verschiedenen Durchläufen mit unterschiedlichen Startparametern für die Clusterzentren, auch bei gleicher Anzahl der Clusterzentren, verschiedene Ergebnisse erzeugen (*MacQueen, 1967*). Aus diesem Grund wurde die Initialisierung mit dem Parameter Random vorgenommen, allerdings wurde die Saat des Zufallsgenerators auf einen festen Wert festgelegt. Dadurch sind die erzeugten Ergebnisse reproduzierbar.
Der Silhouettenkoeffizient wird durch die Beziehung jedes Datenpunkts zu den Datenpunkten innerhalb des eigenen Clusters, sowie seiner Beziehung zu den Datenpunkten des nächsten anderen Clusters bestimmt (*Rousseeuw, 1987*). Die Formel 1 zeigt die Berechnung für einen Datenpunkt a_i aus dem Cluster A. Wobei $dist(a)$ der Mittelwert der Abstände innerhalb des Clusters A ist und $dist(b)$ der Mittelwert der Abstände zum nächsten Cluster ist. Die Werte können im Intervall $[-1; 1]$ liegen. Je näher der Wert an 1 ist, desto besser.

$$\frac{dist(b)_i - dist(a)_i}{\max\{dist(a)_i, dist(b)_i\}} = a_i \tag{1}$$

Um die Bewertung eines Clusters zu bekommen wird der Mittelwert für die Punkte innerhalb des Clusters gebildet.

Ergebnisse

Unter Anwendung des K-Means Algorithmus in Python aus der Bibliothek *scikit-learn* wurden mittels der Parameter aus Tabelle 2 und einen für den Zufallsgenerator festgelegten Saatwert

von 1000 Ergebnisse für zwei bis neunzehn Clusterzentren erzeugt. In der Tabelle 3 sind die entsprechenden Silhouettenkoeffizienten für die Ergebnisse angegeben. Das beste Ergebnis liegt vor, wenn zwei Cluster gebildet werden. Der Wert des Silhouettenkoeffizienten beträgt 0,6294.

Table 2: Liste der Parameter für den K-Means Algorithmus aus der Pythonbibliothek *scikit-learn*. Hierbei sind nur die Parameter beschrieben, die nicht den Standardwert haben.

n_init	**max_iter**	**init**	**precompute_distance**	**algorithm**
50	10000	Random	auto	auto

Table 3: Liste der berechneten Silhouettenkoeffizienten für die Durchläufe mit zwei bis neunzehn Clusterzentren.

Anzahl Clusterzentren	**Silhouettenkoeffizient**	**Anzahl Clusterzentren**	**Silhouettenkoeffizient**
2	0.629431187658	11	0.177136064303
3	0.446572890906	12	0.18674510844
4	0.395827376457	13	0.189391132155
5	0.410286556867	14	0.175363674753
6	0.317900401115	15	0.157178218787
7	0.182706067083	16	0.160579818612
8	0.230772013484	17	0.121848966922
9	0.210190869272	18	0.172766158589
10	0.207998314761	19	0.143293077116

Zusammenfassung

Aus den Zeitreihen der Eye-Trackingdaten der Versuchspersonen wurden 117 Merkmale abgeleitet, die eine Versuchsperson beschreiben. Mit diesen Merkmalen wurden durch Anwendung des K-Means Algorthmus und der Bewertung des Ergebnisses durch den Silhouettenkoeffizient herausgefunden, dass es am besten ist wei Cluster zu bilden, denen die Versuchspersonen zugeordnet werden. Das Ergebnis könnte durch andere Clustering-Algorithmen gegebenenfalls noch verbessert werden. Die Anwendung der gefundenen Cluster zur Früherkennung von Dyslexie konnte im Rahmen der Arbeit nicht getestet werden.

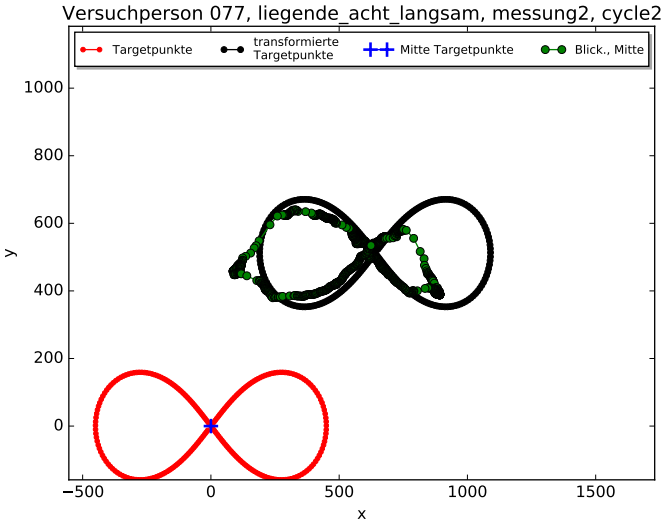

Figure 1: Beispiel liegende Acht von Versuchsperson 77: in rot sieht man die aufgezeichneten Zielpunkte, in grün die Mittleren Blickpositionen. In schwarz sind die transponierten Targetpunkte dargestellt.

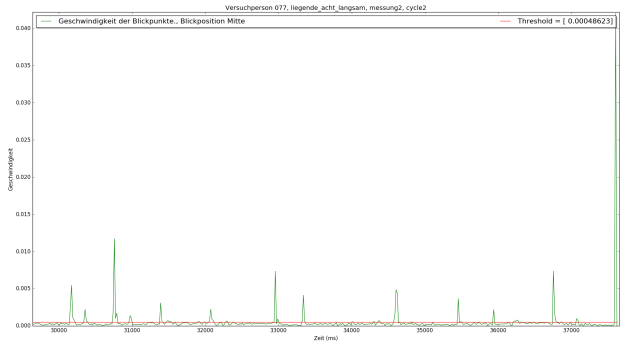

Figure 2: Beispiel Versuchsperson 77: Die rote Linie ist der Schwellwert, für die Geschwindigkeit, damit eine Augenbewegung als Sakkade gezählt wird. Die grüne Kurve ist der Verlauf der mittleren Augenposition zwischen linken und rechten Auge während eines Versuchs liegende Acht.

Literaturverzeichnis

Handler and Fierson, 2011. Handler, S. M. and Fierson, W. M. (2011). Learning disabilities, dyslexia, and vision. *Pediatrics*, 127(3):e818–e856.

Holmqvist et al., 2011. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijer, J. (2011). *Eye Tracking A Comprehensive Guide to Methods and Measures*. Oxford University Press.

MacQueen, 1967. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1:Statistics*, Fifth Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297, Berkeley, Calif. University of California Press.

Rousseeuw, 1987. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Siegel, 2006. Siegel, L. S. (2006). Perspectives on dyslexia. *Paediatrics & Child Health*, 11(9):581–587.

Tiadi et al., 2016. Tiadi, A., Grard, C.-L., Peyre, H., Bui-Quoc, E., and Bucci, M. P. (2016). Immaturity of visual fixations in dyslexic children. *Frontiers in Human Neuroscience*, 10:58.