Clusteranalyse und Mobile Health: ein Literaturreview

Johanna Wallner[1], Max-Marcel Theilig[1], Anne-Katrin Witte[1]

[1] Technical University Berlin, Chair of Information and Communication Management, Berlin, Germany
j.wallner@campus.tu-berlin.de, {m.theilig, a.witte}@tu-berlin.de

Abstract. Das Ziel der vorliegenden Arbeit ist der theoretische Vergleich von Clusteralgorithmen im Mobile Health Kontext. Dazu wurden mithilfe eines Literaturreview 16 Clusteranalyseanwendungen auf gesundheitsbezogene Daten identifiziert. Die Forschungsarbeiten wurden zunächst hinsichtlich der verwendeten Algorithmen geprüft und schließlich 8 Arbeiten ausgewählt, um sie genauer zu analysieren. Die gewonnenen Erkenntnisse wurden im Anschluss auf den Mobile Health Bereich übertragen, um eine Eignung der einzelnen Algorithmen für Mobile Health Daten zu bewerten. Der Großteil der untersuchten Arbeiten beschränkte sich auf die Verwendung der klassischen hierarchischen und partitionierenden Clusteralgorithmen. Besonders vielversprechende Ergebnisse erzielte aber die vergleichsweise moderne Technik selbstorganisierender Karten. Diese Arbeit ist für alle Leser interessant, die mehr über das Potential von Data Mining und insbesondere Clusteranalyse im Mobile Health Bereich erfahren möchten.

Keywords: cluster analysis, mobile health, data mining, literature review

1 Einleitung

Im Rahmen des Forschungsprojekts "Self-administered Psycho-Therapy-Systems“ (SELFPASS) der TU Berlin soll eine Therapie-App für Menschen mit Depression entwickelt werden. Diese App soll auf Basis von Biosignaldaten (wie z.B. der Herzfrequenz), standortabhängigen Umweltinformationen und der Selbsteinschätzung des Patienten individuelle praktische Empfehlungen zum Selbstmanagement geben. Außerdem soll die Symptomatik langfristig protokolliert und das System durch eine Feedback-Schleife weiter personalisiert werden. Es sollen zum Beispiel Warnungen generiert werden, sobald der Patient bedenkliche Verhaltensmuster zeigt. Dieser intelligente Aspekt der App soll durch Integration von Data Mining erreicht werden. Neben dem sogenannten präskriptiven Data Mining sollen auch Algorithmen zur Mustererkennung implementiert werden. Die Patienten sollen anhand der verfügbaren Daten analysiert und gruppiert werden. Diese Gruppierungen sollen neue Erkenntnisse über die Zusammenhänge zwischen individuellen physischen und psychischen Zustände liefern. Aus dieser Problemstellung ist die Idee für die vorliegende Arbeit entstanden.

Die zentrale Forschungsfrage lautet, welcher Clusteralgorithmus am besten für Mobile Health geeignet sein könnte. Mittels eines Literaturreviews soll ein Überblick über

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

aktuelle Anwendungen von Clusteranalyse auf gesundheitsbezogene Daten gegeben werden.

2 Methodik

Um die Frage zu beantworten, wie Clusteralgorithmen auf Mobile Health Daten angewendet werden können, wurde mittels einer Literaturübersicht recherchiert, wie Clusteranalysen in der Vergangenheit auf gesundheitsbezogene Daten angewendet wurden. Da bisher noch recht wenig Forschung zum Thema Clusteranalyse im Mobile Health Bereich direkt betrieben wurde, wurde das Suchfeld auf den gesundheitsbezogenen Kontext erweitert. Dies schloss somit auch Forschungsarbeiten ein, die Daten aus Umfragen analysierten, welche aber dennoch durch Smartphones und Wearables generierbar sind.

Es wurden nur Forschungsarbeiten berücksichtigt, die für den Mobile Health Bereich anwendbare Daten analysierten. Dabei wurde vor allem darauf eingegangen, welche Clusteralgorithmen verwendet worden sind und wie diese angewendet wurden. Daraus soll abgeleitet werden, welche Clusteralgorithmen sich besonders für den Mobile Health Kontext anbieten könnten. Insbesondere sollen die Unterschiede zwischen den einzelnen Algorithmen herausgestellt werden. Das Ziel ist, Handlungsempfehlungen für die Wahl des Clusteralgorithmus zu geben, die zum bestmöglichen Ergebnis bei der Analyse von Mobile Health Daten führen.

Dazu wurde ein Literaturreview nach den Vorschlägen von Levy & Ellis [1] und Webster & Watson [2] durchgeführt. Das Literaturreview gliedert sich in drei Schritte: Input, Verarbeitung und Output (s. Abbildung 1). Im Folgenden werden die einzelnen Schritte beschrieben, wobei Output mit dem Schreiben des Reviews gleichzusetzen ist und deshalb nicht explizit erwähnt wird.

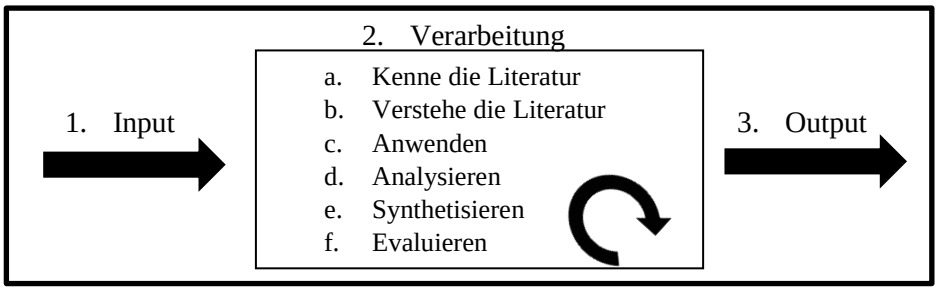

Abbildung 1. Die 3 Schritte des Literaturreviews, vgl. [1]

1. Input: Aufgrund der zeitlichen Begrenzung der Bearbeitungszeit wurde das Literaturreview gemäß Cooper's Taxonomie als „exhaustive with selective citation" durchgeführt. Das heißt, es wurden nicht alle Resultate vollständig analysiert, sondern nur eine geringe Auswahl [3].

Im ersten Schritt wurden die Datenbank PubMed und Google Scholar nach den folgenden Stichworten durchsucht: „clustering", „cluster analysis", „health data", „mental

health“, „health“, „lifestyle“, „heart rate“, „heart rate variability“, „sleep“, „pattern recognition“, „unsupervised machine learning“, „patient”, „medicine”, „healthcare”.

Anhand der Überschrift, des Abstracts und (wenn vorhanden) der Conclusion wurden relevante Arbeiten ausgewählt. Es wurden nur Forschungsarbeiten einbezogen, in denen Patienten nach Daten geclustert wurden, die in Bezug zum Mobile Health Kontext stehen. Arbeiten zum Thema Gen Clustering, Text Clustering oder Medical Image Clustering fanden keine Berücksichtigung. Studien, in denen die Daten rein statistisch untersucht wurden, also ohne den Aspekt des Data Mining, wurden ebenfalls ausgeklammert.

Außerdem soll die Literatur ausschließlich qualitativ hochwertig sein, d.h. sie soll aus Peer Review geprüften Zeitschriften stammen. Die einzige Ausnahme bildet die Arbeit von Meyer et al. aufgrund der Relevanz des Themas und der Qualität des Artikels. Die 16 resultierenden Studien wurden im nächsten Schritt weiterbearbeitet.

2. Verarbeitung: Gemäß Webster & Watson [2] wurde das Literaturreview konzeptzentriert umgesetzt. Diese Form ermöglicht im Gegensatz zur autorenzentrierten Durchführung, welche im Wesentlichen eine Aneinanderreihung von Zusammenfassungen ist, eine Synthese der einzelnen Erkenntnisse zu einer neuen Schlussfolgerung [2]. Jede Forschungsarbeit wurde nach dem gründlichen Lesen und Verstehen (Schritte 2.a und 2.b in Abbildung 1) in die sogenannte Konzeptmatrix eingetragen (Schritte 2.c). Es wurden weiterhin diejenigen Studien ausgewählt, die inhaltlich am meisten auf die Clusteranalyse eingegangen sind. Außerdem wurde versucht, eine möglichst heterogene Auswahl bezüglich der Algorithmen zu bilden. Die 8 resultierenden Studien (s. Tabelle 1) wurden hinsichtlich ihrer Daten, Distanzmaß und Validierung analysiert (Schritt 2.d). Dieser Schritt ist von der Zusammenfassung insofern abzugrenzen als dass eine Analyse nur (für die Forschungsfrage) relevante Informationen extrahiert und auswertet.

Aus den restlichen 8 Arbeiten wurden nur die verwendeten Clusteralgorithmen extrahiert (s. Tabelle 2). Da einige dieser Arbeiten dennoch interessante Erkenntnisse boten, wurden diese Informationen in die Diskussion einbezogen. Die Synthese und Evaluation (Schritte 2.e und 2.f) beinhaltet die Verbindung der einzelnen Forschungsarbeiten zu einem neuen Ganzen sowie ihre qualitative Bewertung [1].

3 Ergebnisse

Tabelle 1. Fokusarbeiten

Paper	*Inhalt*	*Algorithmus*
[4]	**Daten**: 96 kategoriale Lifestyle-Environ Variablen, depressiv (ja/nein) **Distanzmaß**: Keine Angabe **Validierung**: Externe Validierung, Multivariate logistische Regression	SOM (R), agglom. hierarch. Clustering (Complete Linkage)
[5]	**Daten**: Zeitreihen: Herzfrequenz, HRV, Movement Activity Information **Distanzmaß**: Keine Angabe **Validierung**: Partition Index, Separation Index, Xie-Beni's Index	K-Means (Matlab)
[6]	**Daten**: Zeitreihen: Activities of Daily Living **Distanzmaß**: Dynamic Time Warping **Validierung**: Externe Validierung (Accuracy, F-1 Score, Normalized Mutual Information, Cluster Purity), Average Silhouette Width	Agglom. hierarch. Clustering (Complete Linkage)
[7]	**Daten**: Kategorial: gesundheitsbezogenes Verhalten, Selbsteinschätzung der Gesundheit, Lebensqualität, soziale Schicht, psychische Gesundheit **Distanzmaß**: Log-Likelihood-Distanz **Validierung**: Bayesian Information Criterion	TwoStep Clustering (SPSS)
[8]	**Daten**: Numerische Zeitreihe: Schritte in 15-minütigen Intervallen und pro Tag, außerdem Tagestyp (Jahreszeit, Wochentag/-ende) **Distanzmaß**: Eigenes Distanzmaß (basierend auf der Gower Distanz) **Validierung**: Keine Angabe	Wards hierarch. Clustering (R)
[9]	**Daten**: Physische und ruhende Aktivität **Distanzmaß**: Euklidische Distanz **Validierung**: Mehrmalige Anwendung des Algorithmus	Agglom. Hierarch. Clustering, K-Means
[10]	**Daten**: Kategorische und kontinuierliche Variablen: Bildschirmzeit, Schlaf, akademische Leistung, sozio-demografische Daten **Distanzmaß**: Euklidische Distanz **Validierung**: Keine Angabe	SOM (Matlab), K-Means
[11]	**Daten**: Zeitreihe: Asthma Symptome und Auslöser (binär) **Distanzmaß**: Ähnlichkeitsmaß basierend auf der Häufigkeit von Symptomen in einem bestimmten Zeitfenster **Validierung**: Relative Validierung (über Simulationen)	K-Means Consensus Clustering

Tabelle 2. Weitere Arbeiten

Paper	*Algorithmus*
[12]	TwoStep-Clustering (SPSS)
[13]	Agglom. hierarch. Clustering (Ward)
[14]	K-Medoid
[15]	K-Means (SAS FASTCLUS)
[16]	K-Means (SPSS)
[17]	2 Schritte: Agglom. hierarch. Clustering (Ward), K-Means
[18]	Wahrscheinlichkeits-basiertes Clustering (EM-Algorithmus)
[19]	TwoStep-Clustering

Welcher Clusteralgorithmus für eine Aufgabe gewählt werden sollte, ist abhängig von den zu bearbeitenden Daten und vom eigentlichen Ziel der Analyse. Das Ziel der Clusteranalyse von Mobile Health Daten ist die Gruppierung der Patienten und die Entdeckung von verborgenen Zusammenhängen zwischen den Gesundheitsparametern.

Ohne einen vorliegenden realen Datensatz ist es schwierig, Annahmen über die Natur der Mobile Health Daten zu treffen. Ob ein Mobile Health Datensatz beispielsweise nicht-sphärische bzw. willkürlich geformte Cluster enthält, kann nicht eindeutig ausgeschlossen werden. Diese Besonderheit spielt aber eher in der Bilderkennung und beim Clustern von geologischen oder räumlichen Daten eine Rolle und ist deshalb im Mobile Health Bereich nicht von Priorität [20]. Auch ist nicht damit zu rechnen, dass ein Mobile Health Datensatz eine sehr große Anzahl von Dimensionen haben wird. Wenn die Mobile Health Daten als Zeitreihe aufgezeichnet werden, sind diese zwar hochdimensional, aber auch von anderen hochdimensionalen Daten, wie z.B. Genexpressionsdaten, zu unterscheiden. In diesen Datensätzen sind die Datenpunkte spärlich verteilt, d.h. es gibt sehr viele Lücken, und es sind spezielle Clusteralgorithmen notwendig [21]. Viel wichtiger ist es für Mobile Health Anwendungen, einen skalierbaren und effizienten Algorithmus zu verwenden. Außerdem sollte der Algorithmus robust gegenüber Ausreißern und Datenrauschen und bestenfalls in der Lage sein, gemischte kategoriale und numerische Daten zu verarbeiten. Bei der Clusteranalyse von Mobile Health Daten handelt es sich also um eine vergleichsweise einfache Problematik (im Vergleich zu Text-Clustering), für die eine Vielzahl von Clusteralgorithmen in Frage kommen.

Im Allgemeinen konnte festgestellt werden, dass in einem Großteil der betrachteten Arbeiten auf die in Statistiksoftware enthaltenen Standardalgorithmen zurückgegriffen wurde. Dies war jedoch vor allem bei Artikeln, die aus medizinischen Journals stammten, der Fall. In diesen Arbeiten lag der Fokus weniger auf dem maschinellen Lernen an sich als auf den dadurch ermöglichten medizinischen Erkenntnissen, und es fand allgemein wenig Auseinandersetzung mit dem Clusteralgorithmus an sich statt. Der Algorithmus wurde nur sehr knapp beschrieben und auch eine Begründung, wieso gerade dieser Algorithmus gewählt wurde, fehlte. Dipnall et al. [4] beispielsweise dokumentierten zwar sehr ausführlich die Vorgehensweise bei ihrer Clusteranalyse mit selbstorganisierenden Karten, sie gaben aber keine Begründung dafür an, wieso sie gerade ein 20 x 20 rektanguläres Gitter benutzten. Nur in einem kleinen Teil der untersuchten Arbeiten wurden mehrere Algorithmen getestet und verglichen; in keiner Arbeit wurde

Kritik am verwendeten Algorithmus geäußert. Dennoch konnten einige Informationen aus den aktuellen Anwendungen von Clusteralgorithmen auf gesundheitsbezogene Daten extrahiert werden.

3.1 Partitionierende Verfahren

Die meisten Arbeiten nutzten partitionierende Verfahren (7 von 16); davon nutzten 4 den K-Means-Algorithmus alleine, 2 K-Means in Verbindung mit hierarchischem Clustering und eine Arbeit den K-Medoid-Algorithmus. Partitionierende Clusteralgorithmen haben im Allgemeinen folgende Vorteile: Sie sind sehr effizient in Bezug auf große Datenmengen, skalierbar und einfach anzuwenden. Jedoch hat der K-Means-Algorithmus einige Nachteile. Die Daten müssen zum Beispiel messbar sein, das heißt nicht kategorial. Außerdem muss (wenigstens ungefähr) die Clusterzahl K bekannt sein. Bidargaddi et al. [5] nutzten den K-Means-Algorithmus der Matlab Toolbox. In der Arbeit selbst wurde keine Angabe zum verwendeten Distanzmaß gemacht, es handelte sich jedoch sehr wahrscheinlich um die in Matlab enthaltene Hamming Distanz, da u.a. die binäre Variable „Aktivitätszustand“ geclustert wurde und K-Means nur in Verbindung mit der Hamming Distanz dazu in der Lage ist, binäre Variablen zu clustern [22].

Wenn in dem zu clusternden Mobile Health Datensatz kategoriale Variablen enthalten sind, ist K-Means nicht anwendbar. Wenn der Datensatz nur aus Zahlen besteht, bietet sich aber dennoch eher ein auf K-Medoid basierendes Verfahren, wie z.B. CLARANS, an. Dieses Verfahren ist robuster gegenüber Ausreißern und Datenrauschen als K-Means und effizient auf große Datenmengen anwendbar.

Weiterhin sind die Ergebnisse des K-Means-Algorithmus stark von den Initialisierungswerten der Algorithmen abhängig. Bidargaddi et al. [5] wendeten den K-Means-Algorithmus mehrmals mit verschiedenen Anfangswerten an. Dies ist eine empfehlenswerte Methode, um die Konvergenz des Algorithmus bei lokalen Optima zu verhindern.

Tignor et al. [11] nutzten das sogenannte Consensus Clustering und bildeten ein Ensemble, bestehend aus 100 K-Means-Basis-Clusterlösungen. Zusammen mit einem wahrscheinlichkeitsbasierten Imputationsalgorithmus, der die nicht zufällig fehlenden Werte ersetzte, konnten sinnvolle Muster im Datensatz entdeckt werden. Zwischen der Forschungsarbeit von Tignor et al. [11] und der Fragestellung der vorliegenden Arbeit lässt sich ein direkter Bezug herstellen. Sie wandten ihre Methode erfolgreich auf einen Mobile Health Datensatz, bestehend aus Zeitreihen von Asthmasymptomen und -auslösern sowie weiteren kategoriale Patientenvariablen, an. Somit ließe sich diese Clustermethode beispielsweise auf die SELFPASS-App übertragen, indem die täglichen Angaben zu Depressionssymptomen geclustert werden.

3.2 Hierarchische Verfahren

6 der 16 untersuchten Arbeiten nutzten eine Form von hierarchischem Clustering. Der größte Nachteil dieser Clustermethode ist die quadratische Komplexität, also eine Ineffizienz für große Datenmengen. Die Arbeiten, die hierarchisches Clustering verwendeten, analysierten relativ kleine Datensätze. Für das Data Mining von Mobile Health

Datensätzen sind jedoch schlecht skalierbare Algorithmen ungeeignet. Um dennoch effektive Clusterlösungen mittels hierarchischem Clustering zu bilden, sollte deshalb eine hybride Methode gewählt werden. Marshall et al. [9] nutzte beispielsweise hybrides hierarchisches K-Means-Clustering. Im Allgemeinen bieten sich solche hybriden Methoden besonders an, da sie die verschiedenen Vorteile der Algorithmen kombinieren. Die schlechte Skalierbarkeit und Effizienz auf große Datenmengen der hierarchischen Verfahren lassen sich durch einen zweistufigen Prozess eliminieren. Dadurch können hierarchische Verfahren auch in großen Datensätze z.B. nicht-sphärische Cluster finden.

3 Arbeiten nutzten das in SPSS enthaltene hierarchische TwoStep-Clustering. Der größte Vorteil dieser Methode ist, dass sie sowohl kategoriale als auch numerische Daten verarbeiten kann und die Clusterzahl K automatisch wählt. TwoStep ist eine modifizierte Variante des BIRCH-Algorithmus (BIRCH ist nur auf numerische Daten anwendbar) und besitzt, genau wie BIRCH, eine lineare Komplexität (und keine quadratische wie herkömmliche hierarchische Algorithmen). TwoStep ist somit bestens für große Datenmengen geeignet. Wenn der Datensatz jedoch fehlende Werte enthält, werden diese Fälle von TwoStep lediglich entfernt [23]. Auch Conry et al. [7] analysierten eine große Datenmenge mit gemischten Variablen und ohne Vorwissen zur Clusterzahl. TwoStep lieferte in diesem Fall Cluster, die signifikante Zusammenhänge zwischen verschiedenen gesundheitsbezogenen Verhaltensweisen nachwiesen. Im Mobile Health Bereich kann es jedoch oft zu fehlenden Werten kommen. Würden diese Fälle einfach nur gelöscht werden, könnte das zu Informationsverlusten führen. Es wäre zwar möglich, die fehlenden Werte vor der Anwendung des Algorithmus zu ersetzen, dennoch ist der SPSS TwoStep-Algorithmus aufgrund dieser Problematik nicht empfehlenswert.

3.3 Selbstorganisierende Karten

Pieró-Velert et al. [10] und Dipnall et al. [4] setzten zweidimensionale selbstorganisierende Karten ein, weil diese Methode dazu in der Lage ist, große Datensätze zu verarbeiten und keine Abhängigkeit von Verteilungshypothesen besteht. Außerdem hat sich in einer vorgelagerten Studie gezeigt, dass selbstorganisierende Karten bessere Leistung zeigen als andere Algorithmen. Pieró-Velert et al. [10] nutzten SOM zum Vorclustern und danach K-Means für die finale Clusterlösung, während Dipnall et al. [4] dazu hierarchisches Complete-Linkage-Clustering verwendeten. Pieró-Velert et al. [10] zeigte besonders gut die visuellen Fähigkeiten von SOM. Sie clusterten die Variablen zunächst getrennt voneinander mit dem SOM-Algorithmus. Auf den resultierenden Karten war auf den ersten Blick erkennbar, dass z.B. Mädchen kaum passive Videospiele, dafür aber Mobiltelefone zur Kommunikation benutzen. Vor allem für Mobile Health Daten bzw. nutzerbezogene Daten im Allgemeinen ist eine topologieerhaltende Visualisierung von großem Nutzen, um den Einfluss von vielen verschiedenen Variablen auf die Gesundheit zu identifizieren. Die visuelle Darstellung ist außerdem für die Kommunikation an der Schnittstelle zwischen Medizinern und Datenanalysten von Vorteil. Außerdem ist SOM in der Lage, sowohl numerische als auch kategoriale Daten zu verarbeiten. Dipnall et al. [4] konnten ebenfalls Zusammenhänge zwischen einer

großen Zahl von Einflussfaktoren (insgesamt 96) mit der Variable „Depression“ sichtbar machen.

Marshall et al. [9] und Pieró-Velert et al. [10] wiesen große Ähnlichkeit in ihren zur Verfügung stehenden Daten und ihren Zielen auf. Beide untersuchten die Zusammenhänge von ruhenden Tätigkeiten und ihren möglichen Folgen bei Jugendlichen. Auch ihre Clusterlösungen zeigten ähnliche Relationen. Pieró-Velert et al. [10] schafften es jedoch, weitaus mehr Cluster und Zusammenhänge zwischen den Variablen zu identifizieren. Durch die größere Anzahl der Cluster, konnten feinere Unterschiede zwischen ihnen festgestellt werden als bei Marshall et al. [9].

Der SOM-Algorithmus kann bei falscher Wahl der Inputparameter und Lernmethode schnell zu schlechten Ergebnissen führen. Andere Algorithmen wie Fuzzy-C-Means, K-Means und Ward’s sind einfacher zu implementieren und führen ebenfalls zu guten Ergebnissen. Um die Wahl falscher Inputparameter zu verhindern, sollten, wie bei Peiró-Velert et al. [10], mehrere Variationen getestet und verglichen werden.

3.4 Weitere Verfahren

In [5, 6, 8, 11, 18] lagen die gesundheitsbezogenen Daten nicht als statische Variablen, sondern in Form von Zeitreihen vor. Auch Mobile Health Daten, wie zum Beispiel die Herzfrequenz, können als Zeitreihen aufgezeichnet und verarbeitet werden. Eine Analyse dieser Zeitreihen kann Aufschluss über zeitliche Muster in Verhalten und Gesundheit der Nutzer geben. Jedoch kommt es in mHealth-Zeitreihen oft zu fehlenden Daten, bspw. wenn der Akku des Fitnesstrackers leer ist oder die Verbindung abbricht [25]. Diese Lücken führen vor allem in Zeitreihendaten zu Problemen und müssen vor der Clusteranalyse oder währenddessen durch den Algorithmus sinnvoll gefüllt werden. Dazu verwendeten Tignor et al. [11] und Marlin et al. [18] ein wahrscheinlichkeitsbasiertes Modell. Meyer et al. [8] entfernten Tage, in denen über 90% der 15-minütigen Intervalle 0 Schritte enthielten. Bidargaddi et al. [5] analysierten die Daten nicht als Zeitreihe, sondern wandelten sie in 18202 statische Datenpunkte verteilt auf 6 Wochen und 7 Patienten um. Auf diese Daten wendeten sie dann den K-Means-Algorithmus an. Dies ergibt Sinn, wenn es wie in diesem Fall um das Clustern von Zuständen geht und vereinfacht die Analyse. Wenn Mobile Health Daten als Zeitreihe vorliegen, spielt vor allem die Wahl des Distanzmaßes eine große Rolle. Abgesehen davon können Zeitreihen mit den meisten Cluster-Algorithmen verarbeitet werden.

Yürüten et al. [6] zeigten besonders gut, inwieweit die Leistung von Clusteralgorithmen von der Vorbereitung der Daten und vom gewählten Distanzmaß abhängt. Yürüten et al. verglichen ihre Methode u.a. mit dem K-Means-Algorithmus. Während die eigene Methode das für Zeitreihen entwickelte Distanzmaß Dynamic Time Warping verwendete, wurde K-Means in Verbindung mit der euklidischen Distanz angewandt. K-Means schnitt signifikant schlechter ab als die Methoden mit DTW, obwohl K-Means im Allgemeinen gute Ergebnisse liefert.

Marlin et al. [18] nutzte zur Identifikation von Mustern in elektronischen Patientenakten modellbasiertes Clustering, genauer formuliert: den wahrscheinlichkeitsbasierten EM-Algorithmus. EHR-Daten sind in den meisten Fällen unvollständig, d.h. es fehlen

sehr viele Werte, und auch in Mobile Health Daten kann es zu fehlenden Werten kommen. Deshalb kann es, abhängig vom Anteil der fehlenden Daten, von Nutzen sein, spezielle wahrscheinlichkeitsbasierte Algorithmen für unvollständige Daten zu nutzen. Der EM-Algorithmus ist jedoch gleichzeitig sehr ineffizient für große Datensätze und deshalb für den Data Mining Kontext nicht unbedingt geeignet.

Marshall et al. [9] und Nuutinen et al. [16] führten die Clusteranalyse der Jugendlichen nach Geschlechtern getrennt durch. Die Ergebnisse der Analyse zeigten, dass diese Vorannahme korrekt war: Mädchen und Jungen zeigten unterschiedliche Verhaltensmuster. Dieser Ansatz ist auch für den Mobile Health Bereich interessant. Krankheitsverläufe, Symptome und Ursachen von beispielsweise psychischen Krankheiten können sich bei Männern und Frauen stark unterscheiden [26]. Durch eine nach Geschlecht getrennte Clusteranalyse könnten Unterschiede zwischen Männern und Frauen aufgedeckt werden, um daraus beispielsweise Gesundheitsmaßnahmen und Empfehlungen zu individualisieren.

Die verwendeten Clusteralgorithmen beschränkten sich auf hierarchische, partitionierende und modellbasierte Methoden. Bekannte Algorithmen, wie der dichtebasierte DBSCAN, fanden hingegen keine Anwendung. DBSCAN ist aufgrund seiner Robustheit gegenüber Ausreißern und seiner kleinen Anzahl an Inputparametern bestens für Mobile Health Daten geeignet. Jedoch hat DBSCAN Probleme damit, Datensätze korrekt zu analysieren, wenn die Cluster unterschiedlich dicht sind. Wenn in Mobile Health Daten ein Cluster nur wenige, weit auseinanderliegende Datenpunkte enthält, kann es sein, dass DBSCAN diese Cluster als Datenrauschen klassifiziert.

In vielen Fällen gibt es mehrere Algorithmen, die in Frage kommen und es ist a priori nicht möglich zu sagen, welcher das beste Ergebnis liefern wird. Deshalb ist es in diesen Fällen von Vorteil, mehrere Algorithmen anzuwenden und die Ergebnisse unter Zuhilfenahme von Expertenwissen über die Daten und die Algorithmen zu vergleichen [24].

Der Clusteranalyse, als Technik des unüberwachten maschinellen Lernens, fehlen objektive und allgemeingültige Bewertungskriterien. Alle Forschungsarbeiten präsentierten ihre Clusteranalyse als uneingeschränkt erfolgreich. Es fanden zwar Bewertungen über verschiedene Validierungsindizes statt, jedoch können diese nicht untereinander verglichen werden. Dieser Umstand erschwerte die kritische Diskussion und den Vergleich der Ergebnisse. Ob ein anderer Algorithmus eventuell bessere Ergebnisse geliefert hätte, kann deshalb nur vermutet werden.

4 Fazit

Es wurden 16 Forschungsarbeiten zur Clusteranalyse identifiziert, welche durch eine Kategorisierung der Patienten Zusammenhänge zwischen physiologischen Zuständen, Umweltdaten und Patienteninformationen herstellen konnten.

Um den Umfang dieser Arbeit zu begrenzen, wurde kein vollständiges Literaturreview erstellt. Eine Erweiterung der Suche hätte noch weitere informative Beispiele für Clusteranalysen von gesundheitsbezogenen Daten geliefert. Außerdem stammte der Großteil der untersuchten Arbeiten aus medizinisch orientierten Fachzeitschriften. In

diesen Arbeiten wurde wenig auf den Clusteralgorithmus eingegangen und der Informationsgehalt für die vorliegende Arbeit war deshalb teilweise relativ gering. Einige der verwendeten Methoden ließen sich dennoch auf den Mobile Health Bereich übertragen.

Im Allgemeinen konnte festgestellt werden, dass der SOM-Algorithmus eine effektive Methode für die zeitgemäße Clusteranalyse multidimensionaler Mobile Health Daten ist. Der SOM-Algorithmus ist skalierbar, anwendbar auf gemischte Variablen und liefert robuste Ergebnisse. Hierarchische Clustermethoden sind aufgrund ihrer Ineffizienz nicht für das Mining von Mobile Health Daten zu empfehlen. In Verbindung mit anderen Algorithmen, zum Beispiel mit K-Means und SOM, können sie jedoch skalierbare und besonders robuste Cluster bilden. Wenn der Mobile Health Datensatz nur aus numerischen Variablen besteht, können partitionierende Verfahren angewendet werden. Dabei ist K-Means aufgrund seiner Ausreißerempfindlichkeit nicht zu empfehlen. Die Methode von Tignor et al. zum Clustern von Zeitreihendaten aus Mobile Health stach aufgrund ihres direkten Bezugs zu Mobile Health besonders hervor und lässt sich uneingeschränkt auf die Problemstellung der vorliegenden Arbeit anwenden.

Softwarepakete wie die Matlab SOM-Toolbox bieten eine Vielzahl von verschiedenen Möglichkeiten der Anwendung von selbstorganisierenden Karten. Da die Ergebnisse des SOM-Algorithmus stark abhängig von seinen Inputparametern sind, ist es empfehlenswert, die in den Softwarepaketen enthaltenen Möglichkeiten der Variationen zu nutzen, um das bestmögliche Ergebnis zu erzielen. Auch der K-Means-Algorithmus sollte unter Verwendung von verschiedenen Inputparametern mehrfach angewandt werden. Eine gute Möglichkeit ist hier das Consensus Clustering, das verschiedene Clusterlösungen kombiniert.

Welcher Algorithmus für einen bestimmten Anwendungsfall das beste Ergebnis liefern wird, kann a priori nicht eindeutig entschieden werden. Ob ein Algorithmus die verdeckten Muster und Zusammenhänge in einem Datensatz erkennen wird, ist von vielen verschiedenen Faktoren abhängig, wie der Vorbereitung der Daten und der Wahl des Distanzmaßes und der Inputparameter. Dennoch lässt sich die Wahl eines Clusteralgorithmus im Data Mining von Mobile Health Daten auf skalierbare und robuste Methoden eingrenzen. Dazu gehören vor allem neuere Methoden, wie SOM, CLARANS und DBSCAN. Im Fall von fehlenden Werten, bieten sich modellgestützte Verfahren, wie der EM-Algorithmus an. Die Entscheidung für einen Algorithmus sollte jedoch immer im Kontext der Vor- und Nachbereitung der Clusteranalyse und der Wahl des Distanzmaßes stehen.

Referenzen

1. Levy, Y., Ellis, T.: A systems approach to conduct an effective literature review in support of information systems research. Informing Science: The International Journal of an Emerging Transdiscipline, 181–212 (2006)
2. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future. Writing a literature review. MIS Quarterly 26, xiii–xxiii (2002)

3. Cooper, H.M.: Organizing knowledge syntheses. A taxonomy of literature reviews. Knowledge in Society 1, 104 (1988)
4. Dipnall, J.F., Pasco, J.A., Berk, M., Williams, L.J., Dodd, S., Jacka, F.N., Meyer, D.: Why so GLUMM? Detecting depression clusters through graphing lifestyle-environs using machine-learning methods (GLUMM). European psychiatry : the journal of the Association of European Psychiatrists 39, 40–50 (2017)
5. Bidargaddi, N., Sarela, A., Korhonen, I.: Physiological state characterization by clustering heart rate, heart rate variability and movement activity information. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2008, 1749–1752 (2008)
6. Yürüten, O., Zhang, J., Pu, P.: Decomposing Activities of Daily Living to Discover Routine Clusters. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 1348–1354 (2014)
7. Conry, M.C., Morgan, K., Curry, P., McGee, H., Harrington, J., Ward, M., Shelley, E.: The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. BMC public health 11, 692 (2011)
8. Meyer, E.S., Tran, T., Greenwood, M.: Statistical methods for detecting groups of patterns in daily step count activity profiles. Skyline - Te Big Sky Undergraduate Journal 4 (2016)
9. Marshall, S.J., Biddle, S.J.H., Sallis, J.F., McKenzie, T.L., Conway, T.L.: Clustering of Sedentary Behaviors and Physical Activity Among Youth: A Cross-National Study. Pediatrix Exercise Science 14, 401–417 (2002)
10. Peiró-Velert, C., Valencia-Peris, A., González, L.M., García-Massó, X., Serra-Añó, P., Devís-Devís, J.: Screen media usage, sleep time and academic performance in adolescents. Clustering a self-organizing maps analysis. PloS one 9 (2014)
11. Tignor, N., Wang, P., Genes, N., Rogers, L., Hershman, S.G., Scott, E.R., Zweig, M., Yvonne Chan, Y.-F., Schadt, E.E.: Methods for Clustering Time Series Data Acquired from Mobile Health Apps. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 22, 300–311 (2016)
12. Busch, V., van Stel, H.F., Schrijvers, A.J.P., Leeuw, J.R.J. de: Clustering of health-related behaviors, health outcomes and demographics in Dutch adolescents: a cross-sectional study. BMC public health 13 (2013)
13. Verger, P., Lions, C., Ventelou, B.: Is depression associated with health risk-related behaviour clusters in adults? European journal of public health 19, 618–624 (2009)
14. Marshall, S.A., Yang, C.C., Ping, Q., Zhao, M., Avis, N.E., Ip, E.H.: Symptom clusters in women with breast cancer. An analysis of data from social media and a research study. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 25, 547–557 (2016)
15. Sewitch, M.J., Leffondré, K., Dobkin, P.L.: Clustering patients according to health perceptions. Journal of Psychosomatic Research 56, 323–332 (2004)
16. Nuutinen, T., Lehto, E., Ray, C., Roos, E., Villberg, J., Tynjälä, J.: Clustering of energy balance-related behaviours, sleep, and overweight among Finnish adolescents. International journal of public health (2017)
17. Painter, J., Trevithick, L., Hastings, R., Ingham, B., Roy, A.: The extension of a set of needs-led mental health clusters to accommodate people accessing UK intellectual disability health services. Journal of mental health (Abingdon, England), 1–9 (2017)

18. Marlin, B.M., Kale, D.C., Khemani, R.G., Wetzel, R.C.: Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 389–398 (2012)
19. Dodd, L.J., Al-Nakeeb, Y., Nevill, A., Forshaw, M.J.: Lifestyle risk factors of students: a cluster analytical approach. Preventive medicine 51, 73–77 (2010)
20. Chaoji, V., Li, G., Yildirim, H., Zaki, M.J.: ABACUS: Mining Arbitrary Shaped Clusters from Large Datasets based on Backbone Identification. In: Liu, B., Liu, H., Clifton, C.W., Washio, T., Kamath, C. (eds.) Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 295–306. [Society for Industrial and Applied Mathematics], [Philadelphia, Pennsylvania] (2011)
21. Wang, Y., Miller, D.J., Clarke, R.: Approaches to working in high-dimensional data spaces. Gene expression microarrays. British journal of cancer 98, 1023–1028 (2008)
22. MathWorks: Documentation: Statistics and Machine Learning Toolbox, https://de.mathworks.com/help/stats/kmeans.html
23. SPSS: The SPSS TwoStep Cluster Component, https://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf
24. Kaufman, L., Rousseeuw, P.J.: Finding groups in data. An introduction to cluster analysis. Wiley, New York NY u.a. (1990)
25. Dai, X., Bikdash, M.: Trend Analysis of Fragmented Time Series for mHealth Apps. Hypothesis Testing Based Adaptive Spline Filtering Method with Importance Weighting. IEEE Access, 1 (2017)
26. Afifi, M.M.: Gender differences in mental health. Singapore Medical Journal 48, 385–391 (2007)