

SePaMiM - Sequential Pattern Mining und
Pattern Matching von Krankheits- und
Behandlungsverläufen für klinische Krebsregister

Gefördert vom G-BA Innovationsfond
FKZ: 01VSF20018

KONZEPTION EINES DATA MINING WERKZEUGS

13. DEZEMBER 2022

Version v1.0

KOLJA BLOHM, DAVID KORFKAMP
OFFIS – Institut für Informatik
Escherweg 22, 26121 Oldenburg

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

Inhalt

1	Präambel	2
2	Data Mining im Kontext von SePaMiM	2
3	Sequentielles Clustering.....	2
3.1	Ähnlichkeitsbasiertes sequentielles Clustering	3
3.2	Indirektes sequentielles Clustering	3
3.3	Statistisches sequentielles Clustering.....	4
3.4	Umgang mit komplexen sequentiellen Daten	4
4	Konzept Data Mining Werkzeug	5
	Literatur	6

1 Präambel

Dieses Dokument beschreibt die Konzeption des Data Mining Werkzeugs, die im Rahmen des Projekts SePaMiM erstellt wurde. In Abschnitt 2 wird ein allgemeiner Überblick zum Data Mining im Kontext des SePaMiM-Projekts gegeben. In Abschnitt 3 werden verschiedene Techniken zum Clustering von sequentiellen Daten präsentiert. In Abschnitt 4 wird das geplante Vorgehen zur Erstellung des dazugehörigen Data Mining Werkzeugs skizziert.

2 Data Mining im Kontext von SePaMiM

Data Mining umfasst Techniken, um interessante Muster in großen Datenbeständen aufzudecken [1]. Data Mining Verfahren lassen sich in zwei grundlegende Kategorien unterteilen [1]: Überwachte und unüberwachte Verfahren. Überwachte Verfahren lernen anhand von zuvor gelabelten Daten ein Modell, mit dem Ergebnisse für neue, ungelabelte Daten vorhergesagt werden können. Beispiele für überwachte Verfahren sind Klassifikation und Regressionsanalysen. Unüberwachte Verfahren arbeiten auf ungelabelten Daten. Diese Verfahren dienen dazu, automatisch versteckte Muster in den Daten zu entdecken. Hierzu lernen die Verfahren selbstständig die Struktur der Daten, um Unterschiede oder interessante Eigenschaften eines Datensatzes, zum Beispiel anhand von Ähnlichkeitsmaßen, zu finden. Beispiele für unüberwachte Verfahren sind Clustering und Assoziationsanalysen.

Im SePaMiM Projekt sollen Data Mining Techniken genutzt werden, um in Behandlungsverläufen nach Mustern zu suchen. Das Konzept für das Data Mining Werkzeug basiert auf den vorab erarbeiteten Ergebnissen zum Stand der Technik bzgl. Data Mining auf Verlaufs- / Sequenzdaten bzw. Behandlungsverläufen. Viele Data Mining Aufgaben, wie Klassifikation, Regressionsanalysen, Clusteranalysen, etc., lassen sich auch auf sequentiellen Daten durchführen. Bei den hierzu verwendeten Verfahren handelt es sich teils um Erweiterungen klassischer Data Mining Verfahren aber auch um Neuentwicklungen [2].

Der Stand der Technik wurde dem LKR.NRW vorgestellt und fachlich diskutiert. Da es sich bei den Daten der klinischen Krebsregister um einen neuartigen und sehr komplexen Datensatz handelt, ist das Auffinden neuer Strukturen hier besonders relevant. Daher wurde zuerst beschlossen, den Fokus auf unüberwachte Verfahren zu legen. Hierdurch soll versucht werden, neue Muster in den Daten zu entdecken. Besonderes Interesse gab es bei der automatischen Aufdeckung von Patient*innen-Kohorten anhand ihrer Behandlungsverläufe. Hierbei erschien Clustering als besonders geeignet. In Folge dessen wurden noch einmal vertiefend die Eigenschaften von Clustering auf Verlaufsdaten untersucht und entschieden, den Fokus auf Clustering als Data Mining Verfahren zu setzen. Im folgenden Abschnitt wird daher ein Überblick über sequentielles Clustering gegeben.

3 Sequentielles Clustering

Das sequentielle Clustering befasst sich damit, eine Menge von Sequenzen anhand ihrer Ähnlichkeit in Gruppen einzuteilen [2]. Im Rahmen des SePaMiM Projektes soll sequentielles Clustering dazu genutzt werden, Patient*innenkohorten anhand von Behandlungsverläufe zu bilden. Verfahren für das Clustern von Sequenzen lassen sich in drei grundlegende Kategorien einteilen [3], [7]: ähnlichkeitsbasiertes sequentielles Clustering, indirektes sequentielles Clustering und statistisches sequentielles Clustering. Diese werden in den folgenden drei

Abschnitten vorgestellt. Da die Behandlungsverlaufsdaten im SePaMiM Projekt aus komplexen, multidimensionalen Ereignissen bestehen, werden im Anschluss Möglichkeiten zum Clustern solcher Daten beschrieben.

3.1 Ähnlichkeitsbasiertes sequentielles Clustering

Ähnlichkeitsbasiertes sequentielles Clustering nutzt bereits vorhandene distanzbasierte Clustering Verfahren, wie zum Beispiel hierarchische Verfahren, dichtebasierte Verfahren, wie DBScan und partitionierende Verfahren, wie K-Means [3], [7]. Die Herausforderung besteht hierbei darin, ein geeignetes Ähnlichkeitsmaß zu finden, das die Ähnlichkeit zweier Sequenzen bestimmt.

Ein einfacher Ansatz, die Ähnlichkeit zweier Sequenzen gleicher Länge zu bestimmen, besteht darin, gegenüberstehende Elemente der Sequenzen zu vergleichen und die einzelnen Ergebnisse zu aggregieren.

Sequenz-Alignment Ansätze, sind dagegen nicht auf Sequenzen gleicher Länge beschränkt. Diese versuchen, Sequenzen optimal aufeinander abzubilden. Hierdurch berücksichtigen sie außerdem zeitliche Verschiebungen innerhalb der Sequenzen. Hierunter fallen z.B. Editierdistanzen und Dynamic Time Warping (DTW).

Editierdistanzen werden insbesondere für Sequenzen bestehend aus Symbolen genutzt. Sie bestimmen die Ähnlichkeit zwischen zwei Sequenzen anhand der minimal benötigten Modifikationen (Einfügen, Löschen und Ersetzen), um eine Sequenz in die andere zu überführen [3]. Die Kosten für einzelne Modifikationen können dabei gewichtet werden, um zusätzlich Domänenwissen zu berücksichtigen.

DTW sucht die optimale Abbildung von zwei Sequenzen aufeinander. Hierbei werden jedoch keine Modifikationen vorgenommen, um eine Sequenz auf eine andere abzubilden. Stattdessen wird jedes Element einer Sequenz auf ein Element der Vergleichssequenz abgebildet. Es wird die Abbildung gesucht, die minimale Kosten verursacht, d.h. es werden möglichst ähnliche Elemente aufeinander abgebildet. Die Ähnlichkeit zwischen zwei Elementen kann dabei je nach Anwendungsfall unterschiedlich bestimmt werden. Bei der Abbildung wird außerdem eine Reihe von Einschränkungen berücksichtigt. Zum Beispiel kann nicht zurückgesprungen werden. Sobald ein Element der Eingangssequenz auf Element n der zu vergleichenden Sequenz abgebildet wurde, kann ein folgendes Element nicht mehr auf ein Element $n-1$ oder kleiner der Vergleichssequenz abgebildet werden [6].

3.2 Indirektes sequentielles Clustering

Indirekte sequentielle Clustering Methoden verfolgen den Ansatz, Sequenzen zunächst in Feature-Vektoren umzuwandeln, um diese im Anschluss mit vorhandenen Clustering-Verfahren clustern zu können [3], [7]. Entsprechend wichtig ist hierbei die Auswahl geeigneter Features. Ein möglicher Ansatz besteht zum Beispiel darin, Sequenzen auf das Vorhandensein bzw. Nichtvorhandensein einzelner Elemente zu reduzieren. Alternativ kann auch die Häufigkeit der Elemente genutzt werden. Bei diesen Ansätzen geht jedoch die Reihenfolge der Elemente verloren. Eine Möglichkeit, die Reihenfolge der Elemente zu berücksichtigen, ist die Verwendung von kurzen Teilsequenzen. Eine Teilsequenz besteht aus k aufeinanderfolgenden

Elementen und wird als Feature genutzt, das angibt ob bzw. wie oft eine Teilsequenz vorhanden ist [7]. Bei numerischen Sequenzen können beispielsweise bekannte Verfahren aus der Signalverarbeitung, wie die Haar Wavelet Transformation genutzt werden, um die Sequenz in einen Feature-Vektor abzubilden [7], [10].

3.3 Statistisches sequentielles Clustering

Beim statistischen sequentiellen Clustering werden statistische Modelle für das Clustering genutzt. Hierbei wird die Annahme getroffen, dass jede Sequenz durch ein statistisches Modell bzw. durch Wahrscheinlichkeitsverteilungen erstellt wurde [7]. Die verschiedenen Ansätze lassen sich weiter anhand der Verwendung der statistischen Modelle in zwei Kategorien einteilen.

In der ersten Kategorie wird davon ausgegangen, dass die den Sequenzen zugrundeliegenden Modelle jeweils ein Cluster darstellen. Entsprechend werden hierbei die Parameter der Modelle erlernt, beispielsweise mit Hilfe des EM-Algorithmus (Expectation–maximization algorithm) [3]. Die zweite Kategorie beinhaltet Methoden, die statistische Modelle zur Berechnung der Ähnlichkeit zwischen Sequenzen nutzen [7]. Hierbei wird zunächst jede Sequenz auf ein Modell abgebildet und im Anschluss der Abstand zwischen zwei Sequenzen mittels Ähnlichkeitsmaßen für Wahrscheinlichkeitsverteilungen, wie z.B. der Kullback-Leibler-Divergenz ermittelt [7]. Zum Clustern werden dann auf Abstandsmaßen basierende Verfahren genutzt. In beiden Ansätzen werden u.A. Markov Ketten, Hidden Markov Modelle oder Autoregressive Moving Average (ARMA) Modelle als statistische Modelle genutzt [3], [7].

3.4 Umgang mit komplexen sequentiellen Daten

Viele Ansätze zum Clustern von sequentiellen Daten sind zunächst darauf ausgelegt, auf einfachen Sequenzen bestehend aus numerischen oder kategorialen Werten angewendet zu werden. Die Behandlungsverläufe, die im SePaMiM Projekt als Eingabe für das Clustering dienen, bestehen jedoch aus komplexen, multidimensionalen Ereignissen, die sowohl numerische als auch kategoriale Attribute besitzen. Zudem unterscheidet sich die Struktur der Ereignisse je Ereignistyp. Beispielsweise beinhalten Operationsereignisse u.a. eine Intention, eine Liste von Komplikationen, sowie eine Liste von OPS Codes. Strahlentherapien hingegen beinhalten u.a. ebenfalls eine Intention und zusätzlich eine Liste von Bestrahlungen und eine Liste von Nebenwirkungen.

Eine Möglichkeit, mit Sequenzen komplexer Daten umzugehen, besteht darin, die einzelnen Elemente zu kategorisieren und auf einfache Symbole abzubilden. Dieses Vorgehen wird zum Beispiel in [9] beschrieben. Die Autor*innen untersuchen die Versorgungspfade unterschiedlicher Patient*innen mit Herzinsuffizienz. Datengrundlage sind hierbei Abrechnungsdaten von Krankenkassen aus einer Periode von 24 Monaten. Aus diesen Daten werden einfache, aus Symbolen bestehende Sequenzen gebildet. Da die Abrechnungsdaten lediglich quartalsweise erhoben wurden, repräsentiert jedes Symbol einer Sequenz ein Abrechnungsquartal. Für jeden der drei Aspekte *Spezialisierung des Arztes*, *Prozedur* und *Medikation*, wird jeweils ein Sequenztyp gebildet. So wird beispielsweise für jede untersuchte Person eine Sequenz mit der Reihenfolge der Medikation erstellt, deren Symbole entweder die Ausprägung *ACE Hemmer* oder *Betablocker* haben. Die einzelnen Sequenztypen werden

dann individuell geclustert. Hierzu wird das k-Medoids Clustering Verfahren mit der Editierdistanz *Longest Common Subsequence* (LCS) genutzt.

Insbesondere im Bereich der auf Abstandsmaßen basierenden Clusteringverfahren konnten bereits verschiedene Beispiele gefunden werden, die auf komplexeren Daten ausgeführt werden. In [6] wird ein Clustering von Schneeprofilen beschrieben. Ein Schneeprofil besteht aus mehreren Schichten, die sich über die Zeit entwickeln. Jede Schneesicht enthält Attribute, sowohl numerische als auch kategoriale, wie beispielsweise Schneekörnigkeit und Schneehärte. Es werden verschiedene Abstandsmaße vorgestellt, um die einzelnen Attribute der Schneesichten zu vergleichen. Hierauf aufbauend wird ein auf Dynamic Time Warping basierendes Clustering vorgestellt, bei dem die Attribute außerdem unterschiedlich gewichtet werden können. In [8] befassen sich die Autoren mit dem Clustern von multidimensionalen, semantischen Sequenzen. Als Anwendungsbeispiel dienen hier Tagesausflüge, die anhand der unternommenen Aktivitäten in Gruppen eingeteilt werden sollen. Eine Aktivität wird durch drei Attribute beschrieben: besuchter Ort, Art der Unternehmung und architektonischer Stil. Jedes Attribut ist ein Element einer Ontologie. Deren hierarchische Struktur wird zur Bestimmung des Abstandes von Attributen aus zwei unterschiedlichen Sequenzen herangezogen. Beispielsweise ähnelt der Besuch einer Bar dem Besuch eines Restaurants, da beide in der Ontologie als Orte, an denen etwas gegessen werden kann, aufgeführt werden. Zum Vergleich von zwei Sequenzen wird eine Editierdistanz verwendet. Die Autoren untersuchen unterschiedliche Clusteringverfahren und kombinieren diese zusätzlich mit dem Dimensionsreduktionsverfahren UMAP (Uniform manifold approximation and projection for dimension reduction).

4 Konzept Data Mining Werkzeug

Wie im vorherigen Abschnitt dargestellt, gibt es eine Vielzahl von vielversprechenden Ansätzen, um Sequenzen zu clustern. Im SePaMiM Projekt sollen unterschiedliche Ansätze zum Clustering verfolgt werden. Die Datengrundlage ist hierbei das in Zusammenarbeit mit dem LKR.NRW entwickelte Analysemodell, aus dem die für die jeweilige Analyse benötigten Daten exportiert und in ein geeignetes Format transformiert werden. Die Entwicklung des Data Mining Werkzeugs erfolgt iterativ mit regelmäßigen Feedbackrunden zwischen OFFIS und dem LKR.NRW.

Da sich Krebsarten im Verhalten erheblich untereinander unterscheiden können und daher verschiedene Behandlungsmaßnahmen erfordern, ist das Clustering anhand von Behandlungsverläufen insbesondere innerhalb einzelner Krebsentitäten sinnvoll. Daher wird der Fokus auf das Clustern von Brustkrebspatient*innen gelegt. Es wurde beschlossen, sich auf ähnlichkeitsbasierte sequentielle Clusteringverfahren zu fokussieren, da diese es besonders einfach machen, Expertenwissen in die Berechnung der Ähnlichkeit und somit in das Ergebnis des Clustering einfließen zu lassen. Die Ergebnisse des Clustering werden dann mit dem LKR.NRW diskutiert. Die hierbei gewonnenen Erkenntnisse sollen dann in die weitere Entwicklung des Data Mining Werkzeuges einfließen. Es ist daher davon auszugehen, dass hierdurch Anpassungen an dem im Folgenden skizzierten Konzept notwendig sein werden.

Ein erster Ansatz, der untersucht werden soll, ist die Kategorisierung von Ereignissen auf Symbole. Ein sehr einfacher Ansatz hierfür ist die Abbildung der Ereignisse des Analysemodells

auf ihren Ereignistyp, wie z.B. Operation oder Strahlentherapie. Dabei gehen jedoch viele Informationen verloren. Deshalb sollen die Sequenzen auf analyse-spezifisch abgeleitete Symbole abgebildet werden. Beispielsweise kann eine Operation im Kontext einer Brustkrebserkrankung anhand ihres OPS-Codes auf die Symbole BET (Brusterhaltene Therapie), Mastektomie bzw. sonstige Operation abgebildet werden. Ähnliche Abbildungen sind auch für andere Ereignistypen bzw. Krebsentitäten denkbar und werden im Verlauf der Entwicklung mit dem LKR.NRW besprochen und umgesetzt. Diese Sequenzen von Symbolen können dann im Anschluss geclustert werden. Als Abstandsmaße können hierzu beispielsweise Editierdistanzen genutzt werden. Ein solches Vorgehen soll unter Verwendung der Levenshtein Editierdistanz und einem hierarchischen Clusteringverfahren als erster Proof-of-Concept umgesetzt.

In einem weiteren Ansatz soll ein Ähnlichkeitsmaß entwickelt werden, das sowohl die Ähnlichkeit einzelner Ereignisse als auch die Reihenfolge der Ereignisse berücksichtigt. Dazu sollen Operationen, Strahlentherapien, systemischen Therapien und Krebsdiagnosen eines Patienten herangezogen werden. Die Ähnlichkeit zwischen verschiedenen Behandlungsereignissen soll anhand ausgewählter Attribute bestimmt werden. Hierzu wurden bereits erste Vorarbeiten vom LKR.NRW durchgeführt. Bei Operationen sollen zum Beispiel der OPS-Code, der Residualstatus und die Komplikationen zur Bestimmung des Abstandes herangezogen werden. Die berechneten Abstände zwischen einzelnen Ereignissen sollen dann als Kosten einer Levenshtein-Distanz genutzt werden.

Das Ergebnis des Clusterings ist die Zuordnung einzelner Patient*innen zu einer Kohorte anhand ihres Behandlungsverlaufs. Diese Ergebnisse sollen durch Domänenexpert*innen ausgewertet und interpretiert werden können. Neben der Möglichkeit, einzelne Patient*innen aus den Clustern zu betrachten, was insbesondere bei größeren Clustern sehr aufwändig werden kann, ist es hierzu sinnvoll eine Zusammenfassung der einzelnen Cluster zu erstellen. Zum einen können die Cluster über Verfahren der deskriptiven Statistiken beschrieben werden. Beispiele hierfür sind die Größe des Clusters oder die Geschlechtsverteilung innerhalb eines Clusters. Eine weitere Möglichkeit ist es, jedes Cluster durch einen repräsentativen Behandlungsverlauf darzustellen, beispielsweise durch den medoiden Behandlungsverlauf. Ebenso ist es denkbar, die Überlebenszeit der einzelnen Cluster zu berechnen, um die Cluster über diese Metrik miteinander zu vergleichen. Auch die Verwendung weiterer Data Mining Verfahren wie Sequential Pattern Mining ist eine Option. Hierbei werden in einer Menge von Sequenzen häufig auftretende Teilsequenzen gesucht [11]. Hiermit könnten häufig auftretende Teilbehandlungsverläufe eines Clusters ermittelt werden.

Literatur

1. Han J, Kamber M, Pei J (2011) Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
2. Laxman S, Sastry P (2006) A Survey of Temporal Data Mining. *Sadhana* 31:173–198
3. Xu R, Wunsch D (2005) Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions on* 16:645–678

4. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26:43–49. <https://doi.org/10.1109/TASSP.1978.1163055>
5. Xing Zz, Pei J, Keogh E (2010) A Brief Survey on Sequence Classification. SIGKDD Explorations 12:40–48. <https://doi.org/10.1145/1882471.1882478>
6. Herla F, Horton S, Mair P, Haegeli P (2021) Snow profile alignment and similarity assessment for aggregating, clustering, and evaluating snowpack model output for avalanche forecasting. Geoscientific Model Development 14:239–258
7. Warren Liao T (2005) Clustering of Time Series Data-a Survey. Pattern Recogn. 38:1857–1874
8. Moreau C, Chanson A, Peralta V, De Runz C (2021) Clustering sequences of multi-dimensional sets of semantic elements. pp 384–391
9. Vogt V, Scholz S, Sundmacher L (2018) Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. The European Journal of Public Health 28:214–219
10. Rani S, Sikka G (2012) Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications 52:1–9. <https://doi.org/10.5120/8282-1278>
11. Fournier-Viger P, Chun J, Lin W, Kiran RU, Koh YS, Thomas R. A Survey of Sequential Pattern Mining. Vol. 1. 2017.