Применение методов машинного обучения в задаче анализа когорт пациентов с атеросклерозом артерий конечностей

М. В. Демченко, email: mariademchenko94@gmail.ru [1]
И. Л. Каширина, email: kash.irina@mail.ru [1]

[1] Воронежский государственный университет

Аннотация. *В данной работе рассматривается применение методов понижения размерности и кластеризации с целью выделения и анализа основных когорт пациентов с установленным диагнозом атеросклероз артерий конечностей. В качестве метода понижения размерности был использован метод t-SNE, кластеризация пациентов по медицинским показателям была проведена с помощью алгоритма кластеризации k-medoids. В качестве исходного набора данных была использована выборка базы данных MIMIC-III, предоставляющая различные клинические, лабораторные и др. факторы состояния здоровья пациентов. Результатом проведенного анализа является набор основных кластеров состояния пациентов, с помощью которого возможно выделить основные закономерности, свойственные для пациентов, страдающих данным заболеванием.*

Ключевые слова: *Машинное обучение, t-SNE, k-medoids, MIMIC-III, атеросклероз артерий конечностей, кластеризация, понижение размерности.*

Введение

Имеющийся в настоящее время в свободном доступе набор информации медицинского характера предоставляет ряд возможностей для исследования хода течения и особенностей различных заболеваний.

Одним из широко используемых ресурсов электронной медицинской информации является база данных MIMIC-III [1], содержащая информацию о поступлениях пациентов в отделение интенсивной терапии медицинского центра в Бостоне с 2001 по 2012 гг. При этом одним из наиболее частых диагнозов, установленных у пациентов, является такое заболевание, как атеросклероз, что свидетельствует о том, что всестороннее исследование данного диагноза является крайне актуальной задачей современной медицины.

Целью данной работы является поиск основных кластеров пациентов с диагностированным атеросклерозом артерий конечностей, с последующим анализом и выявлением основных закономерностей течения данного заболевания.

1. Описание исходной выборки

Данная работа является продолжением решения задачи диагностики атеросклероза, поставленной в рамках исследования [2,3] ВОКБ №1. В рамках проведенного исследования была построена модель диагностики атеросклероза артерий конечностей с использованием набора данных пациентов Богучарского района Воронежской области, прошедших диспансеризацию в 2014 г. [4,5] Среди исходных показателей пациентов выделялись группы гемодинамеческих, антропометрических, социально-демографических, клинических и лабораторных признаков, с высокой вероятностью являющихся достоверными предикторами атеросклероза. Однако недостатком используемой региональной выборки являлся малый объем, т.к. изучаемый набор данных содержал данные о показателях 522 пациентов.

С целью устранения данного недостатка для продолжения исследования данного заболевания, была использована выборка данных MIMIC-III, содержащая записи измерений показателей пациентов с диагностированным атеросклерозом за весь период их госпитализации.

Данная выборка содержит 22522 записи, описывающих состояние пациента по более, чем 70 показателям, относящихся к различным категориям. Используемые признаки перечислены в таблице 1.

При этом для ряда исследуемых признаков были введены дополнительные признаки, информирующие об отклонении от нормы исходных измерений. Таким образом, фиксировались отклонения каждого из исследуемых лабораторных признаков, а также повышенные или пониженные значения частоты сердцебиения, артериального давления, частоты дыхания, насыщенности крови кислородом.

В ходе данной работы требовалось выделить основные кластеры, характеризующие состояние пациентов с диагностированным атеросклерозом, а также проанализировать основные закономерности, обнаруженные в построенных кластерах.

2. Модели и методы

Задача кластеризации представляет собой задачу машинного обучения, позволяющие выделить сходные группы объектов (кластера) в исходном наборе данных. Данный тип задач представляет собой задачу обучения без учителя и является инструментом описательной статистики, что позволяет выявить значимые закономерности в исследуемой выборке.

Записи исходного набора данных, в общем случае, представляют собой вектора высокой размерности, что зачастую усложняет решение задачи кластеризации таких объектов. Подходом, обеспечивающим

оптимальную работу алгоритмов кластеризации, является предварительное понижение размерности, в результате которого исходные объекты преобразуются в объекты малой размерности, благодаря чему становится возможным их последующая интерпретируемая кластеризация с возможностью визуализации в двухмерном или трехмерном пространствах.

Таблица 1

Основные признаки исходного набора данных

Категория признаков	**Признаки**
Гемодинамические	АД (артериальное давление), ДАД (диастолическое АД), САД (систолическое АД), ЦВД (центральное венозное давление) частота сердцебиения, частота дыхания, SpO2 % и др.
Лабораторные	Глюкоза, холестерин, кальций, гемоглобин, магний, фосфаты, тромбоциты, лейкоциты, эритроциты, натрий, калий и др.
Антропометрические	Рост, вес, пол, возраст
Социально-демографические	Семейное положение, тип страховки
Клинические	Тип оказываемой медицинской услуги, тип атеросклероза (с перемежающейся хромотой, с болью в состоянии покоя, с изъязвлением, с гангреной, без указания симптомов и др.), время пребывания в госпитале

Одним из современных алгоритмов понижения размерности является алгоритм t-SNE (Л. ван дер Маатен, Д. Хинтон, 2008 [6]).

Алгоритм понижения размерности t-SNE.

Шаг 1. Входные данные: исходная выборка $X = \{x_1, x_3, \ldots, x_n\}$, параметр функции потерь $Perp$, количество итераций T, скорость обучения η, момент $\alpha(t)$.

Шаг 2. вычислить попарное сходство $p_{j|i}$ с перплексией $Perp$ (1).

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 \Big/ 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_j\right\|^2 \Big/ 2\sigma_i^2\right)} \qquad (1)$$

Шаг 3. Установить $p_{ij} = p_{j|i} + p_{i|j} / 2n$.

Шаг 4. Инициализировать $Y(0) = \{y_1, y_3, ..., y_n\}$ точками нормального распределения.

Шаг 5. Для t от 1 до T:

Вычислить сходство точек в пространстве отображения q_{ij} (2);

$$q_{ij} = \frac{\exp\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \exp\left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}} \quad (2)$$

Вычислить градиент $\partial Cost / \partial y_i$ (3);

$$\frac{\partial Cost}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1} \quad (3)$$

Установить (4);

$$Y(t) = Y(t-1) + \eta \frac{\partial Cost}{\partial y} + \alpha(t)(Y(t-1) - Y(t-2)) \quad (4)$$

В качестве алгоритма кластеризации в данной работе был выбран метод k-medoids [7].

Алгоритм кластеризации k-medoids.

Шаг 1. Случайным образом выбрать первую центральную точку C_1.

Шаг 2. Вычислить расстояние всех точек набора данных до выбранного центроида. Расстояние от точки x_i до наиболее удаленного центроида вычисляется по формуле (5), где m - число уже выбранных центроидов.

$$d_i = max_{j:1 \mapsto m} \left\|x_i - C_j\right\|^2 \quad (5)$$

Шаг 3. Выбрать точку x_i в качестве нового центроида.

Шаг 4. Повторять шаги 2,3 до тех пор, пока не будут найдены k центроидов.

Шаг 5. Для каждой точки набора данных, вычислить евклидово расстояние между точкой и всеми центроидами. Точка будет определена к кластеру, соответствующему ближайшему центроиду.

Шаг 6. Для каждой из $m-1$ точек кластера, не являющихся центроидами, поменять местами текущий центроид и выбранную точку.

Выбрать в качестве нового центроида точку, минимизирующую функцию потерь (6).

$$M_1, M_1, \ldots, M_1 = argmin \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - M_i \right\|^2 \quad (6)$$

При этом качество алгоритма k-medoids возможно оценить с помощью метрики силуэтного анализа (7).

$$Silhoutte \quad Analysis = \frac{b^i - a^i}{max(b^i, a^i)} \quad (7)$$

3. Результаты и их обсуждение

В результате проведенных этапов понижения размерности и кластерного анализа было выделено 10 основных кластеров пациентов.

Данное число является оптимальным, согласно результатам проведенного силуэтного анализа (рис. 1).

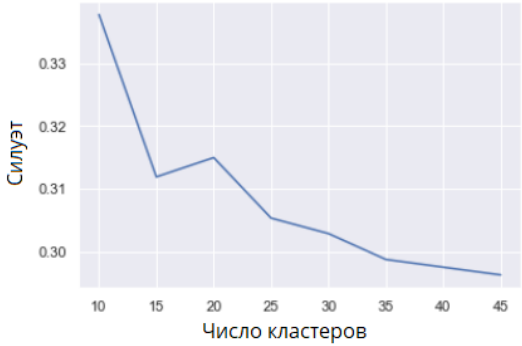

Рис. 1. Результат кластеризации состояний пациентов

Визуализация результатов кластеризации представлена на рис. 2.

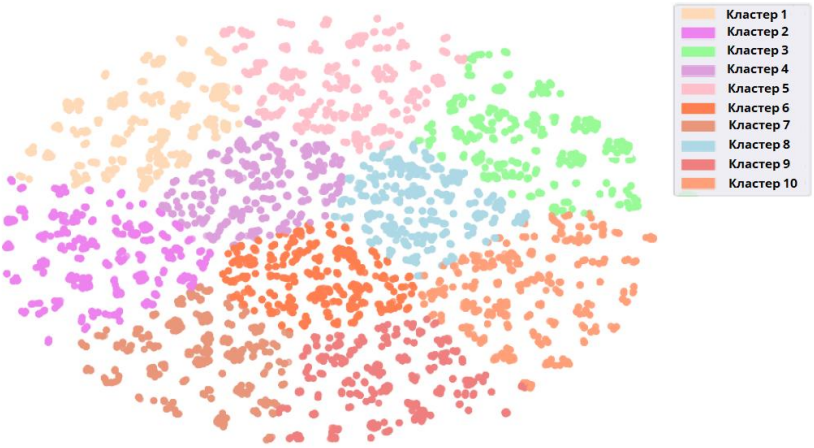

Рис. 2. Результат кластеризации состояний пациентов

Рис. 3 иллюстрирует характеристику центральных объектов полученных кластеров.

	Пол	Вес, кг	ЧСС, уд./мин	Частота дыхания, вд./мин	Время пребывания	Креатинин, мг/дл	Глюкоза, мг/дл	SpO2 %
1	М	78	66.0	11.0	4.1068	1.4	230	96.0
2	Ж	45	68.0	18.0	26.2336	0.4	120	99.0
3	М	100	105.0	27.0	26.8417	0.9	123	99.0
4	Ж	51	83.0	25.0	5.5867	1.3	240	89.0
5	М	108	74.0	24.0	25.9041	0.9	111	95.0
6	М	67	80.0	16.0	14.5210	0.8	114	99.0
7	Ж	65	96.0	20.0	4.6652	0.9	99	96.0
8	М	46	86.0	9.0	3.2958	0.5	126	100.0
9	Ж	49	96.0	14.0	0.5263	0.7	166	92.0
10	Ж	45	66.0	23.0	26.2336	0.5	137	98.0

Рис. 3. Характеристика центральных объектов полученных кластеров

Рис. 4 отражает сводную информацию о полученных кластерах, включая число умерших пациентов к кластере, число отклонений от нормы таких показателей, как креатинин, глюкоза, частота сердцебиения, артериальное давление. На основании данной информации возможно сделать ряд следующих выводов.

Выделяется группа пациентов (кластер № 6), страдающих повышенным артериальным давлением, и при этом относительно невысоким числом отклонений других показателей. Данный кластер связан с высокой смертностью пациентов.

Пониженная или пониженная частота сердцебиения ассоциированы с кластерами № 9 и № 7, соответственно, в которых наблюдается высокая смертность пациентов.

Выделяется группа пациентов с относительно низким числом отклонений показателей от нормы (кластер № 8) и высокой выживаемостью.

Повышение уровня глюкозы или креатинина в крови отрицательно влияет на выживаемость при условии того, что отклонения остальных показателей от нормы зафиксированы относительно нечасто (кластера № 1 и № 4).

	Размер кластера	Число умерших пациентов	Креатинин [Показатель вне пнормы]	Глюкоза [Показатель вне нормы]	Частота сердцебиения [Повышенная]	Частота сердцебиения [Пониженная]	Артериальное давление [Повышенное]	Артериальное давление [Пониженное]
1	2105	680	1066.0	1669.0	49	5	555	60
2	2365	1030	809.0	1183.0	226	179	407	189
3	2149	186	536.0	1307.0	13	0	423	92
4	2199	810	1223.0	1009.0	14	6	526	152
5	2451	813	876.0	1090.0	18	65	868	138
6	2282	1687	846.0	1004.0	16	0	1067	19
7	2072	901	867.0	975.0	7	292	801	284
8	2195	91	114.0	1183.0	27	0	869	28
9	2157	1740	887.0	1050.0	134	60	34	70
10	2547	1551	1013.0	1179.0	96	6	472	154

Рис. 4. Результат кластеризации состояний пациентов

Заключение

В результате данной работы был проведен кластерный анализ пациентов базы данных MIMIC-III, страдающих таким заболеванием, как атеросклероз артерий конечностей.

Данные результаты позволили выявить основные когорты пациентов и выявить основные закономерности изучаемого набора данных.

Список литературы

1. Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35

2. Demchenko, M.V., Kashirina, I.L.: The development of the atherosclerosis diagnostic models under conditions of unbalanced classes. J. Phys.: Conf. Ser. 1479 012026 (2020).

3. Lvovich Y.E., Kashirina I.L., Demchenko M.V. The Use of Machine Learning Methods to Study Markers of Atherosclerosis of the Great Arteries. Information technology 26(1), 46-55 (2020).

4. Khokhlov, R.A., Gaydashev, A.E., Akhmedzhanov N.M.: Predictors of atherosclerotic lesions of limb arteries according to cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology 11(5), 470-476 (2015).

5. Khokhlov, R.A., Gaydashev, A.E., Ostroushko N.I., Gaydashev A.E.: Multi-channel volume sphygmography in cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology 11(4), 371-379 (2015).

6. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

7. Reynolds A.P., Richards G., Rayward-Smith V.J. (2004) The Application of K-Medoids and PAM to the Clustering of Rules. In: Yang Z.R., Yin H., Everson R.M. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2004. IDEAL 2004. Lecture Notes in Computer Science, vol 3177. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28651-6_25