ORES

Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.8

АНАЛИЗ ПАТТЕРНОВ КОМОРБИДНОСТИ В САМООТЧЕТНЫХ МЕДИЦИНСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ АНАЛИЗ АССОЦИАТИВНЫХ ПРАВИЛ APRIORI

Сан Фелипе Гонсалес Маркос,
магистр экономики и бизнес-аналитики, Университет Иоганна Кеплера, г. Линц, Австрия

Аннотация

В данной статье применяются методы анализа ассоциативных правил для выявления паттернов коморбидности в крупном наборе самоотчетных медицинских данных, полученных в результате посещений врачей общей практики и предоставленных Университетом Иоганна Кеплера. Более конкретно, используется алгоритм Apriori с тщательно подобранными пороговыми значениями поддержки и лифта для получения надежных результатов. Анализ выявляет статистически значимые и клинически правдоподобные взаимосвязи между диагностированными заболеваниями и симптомами. Полученные результаты подчеркивают ряд устойчивых паттернов совместной встречаемости, особенно среди расстройств настроения и симптомов опорно-двигательного аппарата. Цель данного подхода заключается в демонстрации полезности методов поиска паттернов в обучении без учителя для выявления значимых структур и поведенческих закономерностей в области здоровья. Были также выявлены определенные ограничения, связанные с причинностью и возможными искажениями самоотчетных данных.

Ключевые слова: машинное обучение, анализ ассоциативных правил, Apriori, коморбидность, медицинские данные, обучение без учителя

COMORBIDITY PATTERN ANALYSIS IN SELF-REPORTED HEALTH DATA USING MACHINE LEARNING APRIORI ASSOCIATION RULE ANALYSIS

San Felipe Gonza´lez Marcos,
Master Economics and Business Analytics program, Johannes Kepler University, Linz, Austria
e-mail: marcosanfelipe4@gmail.com

ABSTRACT

This study applies association rule mining methods to identify comorbidity patterns in a large dataset of self-reported health records obtained from general practitioner visits and provided by Johannes Kepler University. More specifically, the Apriori algorithm is used with carefully selected support and lift thresholds in order to obtain reliable results. The analysis reveals statistically significant and clinically plausible relationships between diagnosed diseases and symptoms. The findings highlight several stable co-occurrence patterns, particularly among mood disorders and musculoskeletal symptoms. The purpose of this approach is to demonstrate the

usefulness of unsupervised pattern mining methods for discovering meaningful structures and behavioral regularities in healthcare data. Certain limitations related to causality and possible biases in self-reported data were also identified.

Keywords: machine learning, association rule mining, Apriori, comorbidity, healthcare data, unsupervised learning

Введение

Понимание структур коморбидности имеет фундаментальное значение не только для общественного здравоохранения, но и для клинических исследований и прогностической медицины. Хронические заболевания часто встречаются совместно вследствие общих биологических механизмов и поведенческих факторов. Алгоритм Apriori [1] предоставляет структурированный механизм для выявления подобных взаимосвязей без необходимости предварительных предположений или медицинских знаний.

Методы анализа ассоциативных правил широко обсуждаются в литературе по интеллектуальному анализу данных [2,5]. Например, в области здравоохранения аналогичные подходы применялись для изучения структур мультиморбидности среди пожилых популяций [3]. Кроме того, психиатрическая литература систематически документирует сильную коморбидность между депрессией и тревожными расстройствами [4].

В данной работе методы анализа ассоциативных правил применяются к крупному набору самоотчетных медицинских данных в рамках подхода обучения без учителя.

Данные

Набор данных был предоставлен Университетом Иоганна Кеплера (JKU) для академических исследовательских целей. Он состоит из 973 583 медицинских записей, соответствующих 50 135 отчетам от 6 307 пользователей.

Каждая запись относится к одному из пяти типов: состояние (C), симптом (S), лечение (T), заметка (N) и наблюдение за погодой (W). Данная классификация уже присутствовала в исходном наборе данных. Метки элементов были объединены с их типом (например, C_depression, S_fatigue) для сохранения семантической структуры.

Важно отметить, что данный набор данных не включает диагнозы тяжелых или онкологических заболеваний (например, рака). Выявленные ассоциации в основном отражают хронические, некритические заболевания, связанные с посещениями врачей общей практики.

После применения one-hot кодирования и предварительной обработки матрица транзакций содержала 22 615 уникальных элементов. Данный метод используется для преобразования набора данных в бинарную матрицу, где каждая строка соответствует наблюдению, а каждый столбец — возможному атрибуту (в нашем случае — объединенным атрибутам). Если человек сообщил о наличии определенного заболевания (например, C_depression), в соответствующей ячейке записывается значение “1”. Соответственно, для всех остальных заболеваний записывается “0”. Иными словами, “1” добавляется ко всем столбцам (заболеваниям), которыми страдает человек, и “0” — в остальных случаях. Такой подход используется для согласования структуры данных со способом обработки информации алгоритмами машинного обучения.

Методология

Частые наборы элементов были извлечены с использованием алгоритма Apriori с минимальным порогом поддержки 5%, а последующие ассоциативные правила были сгенерированы с использованием минимального значения лифта 3.0. Алгоритм Apriori работает путем последовательного сканирования набора данных для поиска комбинаций элементов, которые часто встречаются совместно. Он постепенно формирует более крупные комбинации из меньших частых наборов элементов, где каждая итерация соответствует увеличению размера комбинации. Кроме того, комбинации,

не удовлетворяющие минимальному порогу поддержки, удаляются, что позволяет сделать процедуру вычислительно эффективной даже для больших наборов данных.

Формальные определения

Пусть A и B обозначают наборы элементов.

$$\text{Support}(A) = \frac{\text{Количество транзакций, содержащих A}}{\text{Общее количество транзакций}} \quad (1)$$

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} = P(B \mid A) \quad (2)$$

$$\text{Lift}(A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A) \times \text{Support}(B)} \quad (3)$$

Значения лифта значительно выше 1 указывают на ненормальную (не случайную) совместную встречаемость.

Результаты

В общей сложности 698 частых наборов элементов удовлетворяли порогу поддержки 5%. Наиболее распространенными состояниями были депрессия (~18%), фибромиалгия (~18%), тревожность (~15%) и синдром хронической усталости (~11%).

В таблице 1 представлены наиболее клинически значимые ассоциативные правила, ранжированные по значению лифта.

Таблица 1. Наиболее клинически значимые ассоциативные правила, ранжированные по лифту

Правило	Поддержка	Доверие	Лифт
Усталость ∧ Депрессия → Тревожность	0.057	0.694	4.39
Депрессия → Тревожность	0.109	0.602	3.81
Боль в спине → Боль в шее	0.060	0.491	3.73
Боль в шее → Боль в спине	0.060	0.475	3.73

Связь между депрессией и тревожностью согласуется с результатами известных психиатрических исследований. Аналогично, совместное возникновение болей в спине и шее соответствует литературе по коморбидности опорно-двигательного аппарата.

Обсуждение

Важно отметить, что алгоритм Apriori выявляет совместную встречаемость, а не причинно-следственные связи. Наблюдаемые взаимосвязи могут отражать биологические механизмы, возможные искажения отчетности или общие контекстуальные факторы. Иными словами, выявленные закономерности следует интерпретировать как статистические ассоциации, а не как прямые причинные связи, и для их интерпретации необходим дополнительный медицинский анализ.

Тем не менее, ряд наиболее сильных правил соответствует хорошо задокументированным медицинским взаимосвязям, что указывает на то, что корректное применение методов машинного обучения без учителя может выявлять реальные и значимые закономерности в медицине. В данном случае согласованность между результатами алгоритма и существующей клинической литературой подтверждает внешнюю валидность полученных результатов.

С методологической точки зрения методы обучения без учителя могут дополнять медицинские исследования, основанные на гипотезах. Опираясь исключительно на статистические структуры совместной встречаемости, алгоритм Apriori позволяет исследовать сложные наборы данных без введения предварительных предположений. Следовательно, данный подход может быть использован для выявления скрытых закономерностей, которые не всегда очевидны с медицинской точки зрения, что может способствовать более эффективной диагностике, лечению и новым направлениям медицинских исследований.

В то же время важно признать, что используемый набор данных не включает тяжелые или опасные для жизни диагнозы (например, онкологические заболевания). Извлеченные ассоциативные правила в основном отражают хронические и некритические состояния, о которых сообщают сами пациенты, что ограничивает возможность обобщения результатов на более острые медицинские контексты.

Заключение

Алгоритм Apriori выявляет совместную встречаемость, а не причинно-следственные связи. Следовательно, наблюдаемые взаимосвязи могут отражать биологические механизмы, возможные искажения самоотчетных данных или контекстуальные факторы. Поэтому выявленные закономерности следует интерпретировать как статистические ассоциации, полученные с помощью алгоритма, а не как прямые причинные связи.

Тем не менее, ряд наиболее сильных правил соответствует хорошо известным медицинским взаимосвязям, особенно между депрессией и тревожностью, а также между болями в спине и шее. Такая согласованность подтверждает внешнюю валидность полученных результатов.

Список литературы:

1. Эль Халифа Р. А., Ю П. Ю., Чи С.-Л. Выявление паттернов мультиморбидности с помощью анализа ассоциативных правил у пациентов с болезнью Альцгеймера и связанными деменциями // Материалы объединённого саммита AMIA по трансляционной науке. 2024.
2. Хан Ц., Пэй Ц., Тун Х. Интеллектуальный анализ данных: концепции и методы. 4-е изд. Амстердам: Морган Кауфманн (Эльзевир), 2022. С. 145–173.
3. Чжэн Ц., Се Ю., Хуан Ц., Сун С., Чжан Р., Чэнь Л. Анализ ассоциативных правил паттернов мультиморбидности у взрослых на основе базы данных Национального обследования здоровья и питания // BMJ Open. 2022. Т. 12. № 12. DOI 10.1136/bmjopen-2022-063660.
4. Саха С., Лим С. С. У., Кэннон Д. Л., Бёртон Л., Бремнер М., Косгроув П., Хуо Ю., Макграт Дж. Дж. Коморбидность между расстройствами настроения и тревожными расстройствами: систематический обзор и мета-анализ // Depression and Anxiety. 2021. Т. 38. № 3. DOI 10.1002/da.23113.

5. Джанджи Н. З. Сравнительный анализ алгоритмов поиска частых паттернов на медицинских данных // 9-я Международная конференция IEEE по инженерным технологиям и прикладным наукам (ICETAS). 2024. DOI 10.1109/ICETAS62372.2024.11119839.

References:

1. El Khalifa R. A., Yew P. Y., Chi C.-L. Detecting Multimorbidity Patterns with Association Rule Mining in Patients with Alzheimer's Disease and Related Dementias // AMIA Joint Summits on Translational Science Proceedings. 2024.
2. Han J., Pei J., Tong H. Data Mining: Concepts and Techniques. 4th ed. Amsterdam: Morgan Kaufmann (Elsevier), 2022. pp. 145–173.
3. Zheng Z., Xie Y., Huang J., Sun X., Zhang R., Chen L. Association rules analysis on patterns of multimorbidity in adults: based on the National Health and Nutrition Examination Surveys database // BMJ Open. 2022. Vol. 12. No. 12. DOI 10.1136/bmjopen-2022-063660.
4. Saha S., Lim C. C. W., Cannon D. L., Burton L., Bremner M., Cosgrove P., Huo Y., McGrath J. J. Co-morbidity between mood and anxiety disorders: A systematic review and meta-analysis // Depression and Anxiety. 2021. Vol. 38. No. 3. DOI 10.1002/da.23113.
5. Jhanjhi N. Z. Comparative Analysis of Frequent Pattern Mining Algorithms on Healthcare Data // 2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS). 2024. DOI 10.1109/ICETAS62372.2024.11119839.