УДК 004.891.3

КЛАССИФИКАЦИЯ СИНДРОМОВ ВИРУСНОГО ГЕПАТИТА НА ОСНОВЕ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

ХИДИРОВА Ч.М.

Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Узбекистан

Аннотация. В этой статье предлагается новый метод интеллектуального анализа данных, основанный на многокритериальном ранжировании (МР), для изучения взаимосвязи между синдромами и симптомами вирусного гепатита. Поскольку МР может использовать данные симптомов с экспертной дифференциацией и данные симптомов без экспертной дифференциации в задачу классификации синдромов, клиническая информация, используемая для моделирования признаков синдрома, значительно расширена, чтобы повысить точность классификации синдромов. Предложенный метод классификации синдромов может также избежать двух недостатков предыдущих методов: линейная связь клинических данных и взаимоисключающих симптомов между различными синдромами. Это может помочь более эффективно использовать скрытую связь между синдромами и симптомами. Улучшение точность классификации синдромов может быть достигнуто в соответствии с экспериментальными результатами и клиническими экспертами.

Ключевые слова: интеллектуальный анализ данных, многокритериальное ранжирование, классификация синдромов, вирусный гепатит.

Конфликт интересов. Автор заявляют об отсутствии конфликта интересов.

CLASSIFICATION OF VIRAL HEPATITIS SYNDROMES BASED ON DATA MINING METHODS

CHAROS KHIDIROVA

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan

In this paper, a novel data mining method based on manifold ranking (MR) is proposed to explore the relation between syndromes and symptoms for viral hepatitis. Since MR could take the symptom data with expert differentiation and the symptom data without expert differentiation into the task of syndrome classification, the clinical information used for modeling the syndrome features is greatly enlarged so as to improve the precise of syndrome classification. The proposed method of syndrome classification could also avoid two disadvantages in previous methods: linear relation of the clinical data and mutually exclusive symptoms among different syndromes. It could help exploit the latent relation between syndromes and symptoms more effectively. Better performance of syndrome classification is able to be achieved according to the experimental results and the clinical experts.
Key words: data mining, manifold ranking, syndrome classification, viral hepatitis.
Conflict of interests. The author declares no conflict of interests.

Введение

В последние годы для анализа взаимосвязи между синдромами и симптомами широко применяются многие модели интеллектуального анализа данных, такие как байесовская модель, машина опорных векторов (МОВ) и Фишер [1-5]. Некоторые лучшие результаты были получены для нескольких важных заболеваний, например, ишемической болезни сердца, вирусного гепатита и диабета. Обычно общие методы классификации синдромов делится на два класса. Примеры первого класса основаны на методе обучения без учителя, например, метод k-средних [2,6], правило ассоциаций [2,7], факторный анализ [3] и метод главных компонент (МГК) [8]. Симптомы без экспертной дифференциации напрямую используются для моделирования признаков синдромов и, следовательно, классификация синдромов полностью представляет внутреннюю взаимосвязь данных симптомов. Примеры второго класса основаны на контролируемом методе обучения, например, собственный байесовский классификатор [9,10], МОВ [1], регрессионный анализ [1,3] и нейронная сеть [4,11].

Симптомы с экспертной дифференциацией используются для изучения особенностей синдрома и, следовательно, классификация синдромов полностью представляет собой внутреннее знание экспертов. Однако, с одной стороны, синдром представляет собой объективную характеристику популяции и внутренние отношения данных симптомов также полезны для представления характеристики синдромов. С другой стороны, основной силой дифференциации синдромов являются клинические эксперты и, следовательно, экспертные знания важны для классификации синдромов. Таким образом, метод полу управляемого обучения, который мог бы объединить внутреннюю взаимосвязь данных и экспертные знания, более подходит для изучения классификации синдромов. У вышеупомянутых методов есть два недостатка [1-12]. Во-первых, предполагается, что структура данных среди симптомов и синдромов является линейной зависимостью для некоторых методов, таких как МГК. Во-вторых, классификации синдромов, включающие различные симптомы, являются взаимоисключающими для некоторых методов, например, МОВ, самоорганизующейся карты, дерева решений и собственного байесовского классификатора. Предлагается новый метод классификация синдромов на основе многокритериального ранжирования (КСОМР) для исправления недостатков вышеупомянутых предыдущих методов.

В этой статье МР вводится для исследования классификации синдромов вирусного гепатита. КСОМР имеет три основных характеристики:

1) для распространения классификационных ярлыков синдрома используются внутренние отношения между симптомами;

2) потенциальные нелинейные отношения симптомов рассматриваются для достижения соответствующей вероятности, содержащейся в каждой классификации синдромов;

3) экспертные знания объединяются с внутренними взаимосвязями симптомов, чтобы выявить особенности синдромов вирусного гепатита.

Методы исследования

Общее введение. Во-первых, определение символов показано для представления КСОМР. Набор пациентов задается как $P = \{p_{ki},\ k = 1, ..., K;\ i = 1, ..., M\}$, где K - общее количество пациентов, а M - общее количество симптомов. Набор симптомов задается как $S = \{S_{ik},\ i=1,...,M;\ k =1,...,K\}$ и объ-

единены в матрицу симптомов M_s размером $M \times K$. Класс синдрома задается как $C = \{c_j, j = 1, ..., N\}$, где N - общее количество классов синдромов. Вероятность веса задается как $W = \{w_{ij}, i=1,...,M; j=1,...,N\}$ и объединяются в матрицу M_w весовой вероятности. w_{ij} обозначает вероятность симптома s_i с классом синдрома c_j. Сводная информация представлена в таблице 1.

Таблица 1. Определение символов

Table 1. Definition of symbols

Символы	Определение
$P = \{p_{ki}, \ k = 1, ..., K; \ i = 1, ..., M\}$	Набор пациентов и вектор пациентов
$S = \{s_{ik}, \ i = 1, ..., M; \ k = 1, ..., K\}$	Набор симптомов и вектор симптомов
$C = \{c_j, \ j = 1, ..., N\}$	Набор синдромов
$W = \{w_{ij}, \ i = 1, ..., M; j = 1, ..., N\}$	Набор весовой вероятности и вектор весовой вероятности

Во-вторых, с функциональной точки зрения КСОМР есть три основных модуля:

1) установка исходных векторов меток синдромов;
2) расчет матрицы сходства на основе сегментированного расстояния;
3) распространение ярлыка синдрома итеративным стилем.

Шаблон КСОМР показана на рис. 1.

Установка векторов вероятности начального веса

Вес представляет собой вероятность того, что симптом принадлежит к каждому классу синдромов. Вначале эксперты предоставляют классы некоторых симптомов. То есть классы синдромов этих симптомов изначально известны, и, следовательно, соответствующие исходные вероятности веса установлены равными 1. Аналогичным образом, классы синдромов других симптомов не приводятся экспертами. То есть классы синдрома этих симптомов изначально неизвестны, и, следовательно, соответствующие начальные вероятности веса установлены равными 0. Вероятность веса w_{ij} симптома s_i, помеченного классом синдрома c_j, устанавливается согласно уравнению (1).

$$w_{ij} = \begin{cases} 0, & \text{если симптом } s_i \text{ не относится к классу синдрома } c_j \\ 1, & \text{если симптом } s_i \text{ относится к классу синдрома } c_j \end{cases} \quad (1)$$

Вычисление матрицы подобия

Клинические данные представляют собой особенности сложности, дискретности и разнообразия. Например, некоторые симптомы соответствуют двоичному распределению. Некоторые симптомы соответствуют нормальному распределению. Некоторые симптомы соответствуют равномерному распределению. Таким образом, сходство симптомов рассчитывается с помощью сегментированного расстояния. Расстояние между сегментами определяется уравнением (2) – (4).

$$d_{Hamming} = \sum_{k=1}^{K} |S_{pk} - S_{qk}|, \quad d_{Hausdorff}(S_p, S_q) = max[h(S_p, S_q), h(S_q, S_p)] \quad (2)$$

где, $h(S_p, S_q) = max_{a \in S_p} min_{b \in S_q} \|a - b\|$, $h(S_q, S_p) = max_{b \in S_q} min_{a \in S_p} \|b - a\|$ (3)

$$d_{Euclidean}(S_p, S_q) = \sqrt{\sum_{k=1}^{K} (S_{pk} - S_{qk})^2} \quad (4)$$

Рис. 1. Шаблон классификации символов на основе многокритериального ранжирования
Fig.1. Framework of the manifold ranking based syndrome classification

Более подробно, уравнение (2) используется для вычисления сходства симптомов, соответствующих бинарному распределению. Уравнение (3) используется для вычисления сходства симптомов, соответствующих равномерному распределению. Уравнение (4) используется для расчета схожести симптомов, соответствующих нормальному распределению.

Распространение метки класса синдрома

Учитывая матрицу сходства симптомов и матрицу исходной весовой вероятности, создается граф для распространения метки класса синдрома. Все векторы симптомов построены так, чтобы быть графом $G = (V; E)$, где V обозначает множество вершин графа, а E обозначает множество ребер графа. В частности, множество вершин V - состоит из всех векторов симптомов $s_i = \{s_i, i = 1, ..., M\}$, а множество ребер E построено из всех связей двух вершин, что обозначается подобием двух вершины, т.е. $W = \{w_{ij}, i = 1, ..., M; j = 1, ..., N\}$. Распространение классов синдромов на основе теории графов реализуется посредством минимизации функции стоимости, которая показана в уравнении (5).

$$Q(r) = \frac{1}{2}\left(\sum_{i,j=1}^{K} S_{ij}\left\|\frac{r_i}{\sqrt{D_{ii}}} - \frac{r_j}{\sqrt{D_{jj}}}\right\|^2 + \mu\sum_{i,j=1}^{N}\|r_i - y_i\|^2\right) \quad (5)$$

где, r_i обозначает класс синдрома симптома s_i. Измерение гладкости информации о классе синдрома, которое получают по всему набору симптомов, представлено первой частью уравнения (5). Между тем, измерение стандартного отклонения между окончательным классом синдрома и классом сложного истинного синдрома представлено второй частью уравнения (5). Кроме того, два измерения взвешиваются и объединяются коэффициентом μ. Если задан коэффициент μ, окончательные резуль-

таты оптимизации, то есть результаты распространения класса синдрома, могут быть достигнуты в итерационном процессе.

Итерационная процедура распространения метки класса синдрома

Вход:

Матрица сходства M_s симптомов.

Матрица вероятностей начальных весов M_w всех классов синдромов.

Процедура:

Шаг 1: Сбор матрицы сходства M_s симптомов.

Шаг 2: Нормализация матрицы сходства с помощью уравнения $M_{norm} = M_D^{-\frac{1}{2}} M_S M_{D_{ii}}^{-\frac{1}{2}}$

где M_D – диагональная матрица, а M_{Dii} – сумма i-й строки весовой матрицы вероятностей M_w.

Шаг 3: Итерация уравнения (7) до получения сходящегося решения M_w^*

$$M_w^*(t+1) = aM_{norm}M_w^*(t) + (1-a)M_w(0) \quad (6)$$

где, t – номер итерации, $a \in [0,1]$ и $M_w(0)$ – исходный вес матрицы вероятности.

Шаг 4: Классификация симптомов на основе окончательного веса матрицы вероятностей M_w^*.

Выход:

Окончательный вес матрицы вероятности M_w^*.

Результаты и их обсуждение

Классификация синдромов

Всего из ретроспективных клинических данных было отобрано 1200 случаев вирусного гепатита для генерации KCOMP. Каждый случай собирает 38 показателей наблюдения, таких как осмотр, аускультация-обоняние, опрос, пальпаторный симптом, язык и пульс. Существует 5 классов синдромов: дефицит Ци и застойная желтуха, влажно-жаровая желтуха, дефицит Ян и застойная желтуха, дефицит Инь и застойная желтуха, застойный жар и желтуха. То есть вышеуказанные параметры установлены равными K = 1200, M = 38 и N = 5. Таблица 2 иллюстрирует результаты классификации синдромов и точность каждого синдрома.

Таблица 2. Точность каждого синдрома и его симптомов

Table 2. Precision of each syndrome and its symptoms

Симптомы и вероятность их веса	Стандартные симптомы синдрома	Синдром дифференцировки	Точность
Дымчато-желтый цвет лица и глаз (0,9813) Вздутие живота (0,9247) Язык со следами зубов (0,9126) Струнный импульс (0,9045) Темная моча (0,8961) Горький привкус во рту (0,8532) Анорексия (0,7854) Темно-мрачный цвет лица (0,7236) Недостаток прочности (0,6627) Грубый пульс (0,6358) Белый мех (0.6079)	Темная моча, желтая кожа и глаза, темный и мрачный цвет лица, недостаток силы, анорексия, темно-красный язык, язык со следами зубов, вздутие живота, жидкий стул, тошнота, рвота, сухость во рту, горечь во рту, ипохондрическая боль, киноварная ладонь, ангиома, толстое тело языка, белый мех, струнный пульс	Дефицит Ци и застойная желтуха	0.9311
Оранжево-желтый цвет лица (0,9765) Темная моча (0,9547) Тошнота (0,91238) Желтый мех (0,8846) Скользкий пульс (0,8438) Кислотная рвота (0,8149) Тиннитус (0,7625) Красный язык (0,7456) Горечь во рту (0,6891) Струнный импульс (0,6587)	Короткая темная моча, горечь во рту, тошнота, желтый мех, слизистая шерсть, сероватый цвет лица, жажда без предпочтения напитков, затрудненное мочеиспускание, красный язык, струнный пульс, скользкий пульс	Тепловая желтуха	0.9278

Симптомы и вероятность их веса	Стандартные симптомы синдрома	Синдром дифференцировки	Точность
Склизкий мех (0,6237) Жирная анорексия (0,6110)			
Темный и мрачный цвет лица (0,9682) Склизкий мех (0,9273) Оранжево-желтый цвет лица (0,9028) Медленный пульс (0,8637) Темно-красный язык (0,8364) Жажда без предпочтения напитков (0,8073) Частое и ясное мочеиспускание (0,7543) Затонувший пульс (0,7125) Вздутие живота (0,6513) Непереносимость холода (0,6227)	Желтуха тела и глаз, темный и мрачный цвет лица, темно-красный язык, язык со следами зубов, слизистый мех, скользкий мех, белый мех, замедленный пульс, медленный пульс, анорексия, вздутие живота, жидкий стул, тяжесть в голове и теле, жажда без предпочтения напитков	Дефицит Ян и застойная желтуха	0.8869
Темный и мрачный цвет лица (0,9658) Желтый глаз (0,9127) Грубый пульс (0,8537) Темно-красный язык (0,8333) Жажда с предпочтением напитков (0,7911) Темная моча (0,6879) Желтая кожа и глаза (0,6623) Тепло в ладонях и подошвах (0,6451) Ладони печени (0,6273) Меньше меха (0,6103)	Желтый глаз, темно-желтый цвет лица, ярко-желтая моча, слабость, слабость и истощение, ипохондрическая боль, жажда с предпочтением напитков, бессонница и мечтательность, темно-красный язык, тонкий пульс, грубый пульс	Дефицит Инь и застойная желтуха	0.7931
Жгучая кожа (0,9913) Струнный импульс (0,95432) Анорексия жирная (0,9046) Темно-фиолетовый язык (0,8876) Жажда без предпочтения напитков (0,8562) Носовое кровотечение (0,8249) Темная моча (0,8165) Горечь во рту (0,7852) Сухие два глаза (0,7002) Кислотная регургитация (0,6489) Синевато-фиолетовый язык (0.6083) Белый мех (0,6001)	Темная моча, кожный зуд, жжение кожи, темно-фиолетовый язык, экхимоз языка, петехия языка, извилистость подъязычных коллатеральных сосудов, жажда без предпочтения напитков, сухой стул, кровотечение из носа, кровотечение десен, экхимоз, опухоль в груди, меньше шерсти, тонкий мех, белый цвет мех, струнный пульс, грубый пульс	Застойный жар и желтуха	0.8156

Из изученных результатов таблицы 2 можно добиться лучшей эффективности классификации синдромов на основе КСОМР. В частности, точность дефицита Ци и застойной желтухи может быть достигнута при 0,9311. Точность тепловыделения желтухи также может быть достигнута на уровне 0,9278. Точность дефицита Инь и застойной желтухи самая низкая и достигает 0,7931. Например, синдром дефицита Ян и застойная желтуха складываются из следующих симптомов: темный и мрачный цвет лица (0,9682), слизистый мех (0,9273), оранжево-желтый цвет лица (0,9028), медленный пульс (0,8637), темно-красный язык (0,8364), жажда без предпочтения напитков (0,8073), частое и ясное мочеиспускание (0,7543), замедленный пульс (0,7125), вздутие живота (0,6513), непереносимость холода (0,6227). Соответствующие результаты достигаются на 88,69% по сравнению со стандартным соотношением симптомов и синдромов.

Сравнение с другими методами интеллектуального анализа

Чтобы представить эффективность КСОМР, к набору данных по вирусным гепатитам в настоящее время применяются несколько популярных методов, например, МГК, байесовские, ассоциативные правила, k-средних. Результаты классификации этих методов проиллюстрированы в таблице 3. Очевидно, что более высокая точность классификации синдромов может быть получена с помощью

KCOMP, потому что MP учитывает нелинейную связь симптомов в классификации синдромов. Экспертные знания также сочетаются с внутренним отношением симптомов, следовательно больше клинической информации используется для моделирования признаков синдромов. Основная корреляция между классами синдромов распространяется на все симптомы, поэтому для изучения особенностей синдромов используются более релевантные данные.

Таблица 3. Сравнение производительности различных методов

Table 3. Performance comparison of different methods

Классы синдрома	**Методы**				
	МГК	Байесовский	Правила ассоциации	k-средних	KCOMP
Дефицит Q_i и застойная желтуха	0.6278	0.7341	0.6354	0.5713	0.9311
Влажно жаровая желтуха	0.6089	0.6592	0.5913	0.4639	0.9278
Дефицит Ян и застойная желтуха	0.6001	0.6208	0.5347	0.4378	0.8869
Дефицит Инь и застойная желтуха	0.5642	0.5714	0.4923	0.3874	0.7931
Застойный жар и желтуха	0.5912	0.6108	0.5488	0.4711	0.8156

Заключение

В этой статье предлагается новый метод классификации синдромов вирусного гепатита на основе MP. Поскольку метки синдромов распространяются на все симптомы, основная взаимосвязь симптомов, соответствующая нелинейной зависимости, может быть правильно определена, чтобы повысить точность классификации синдромов. Более того, поскольку в KCOMP учитываются как экспертные знания, так и внутренняя взаимосвязь симптомов, результаты классификации не только представляют собой субъективное правило синдромов, но также представляют собой объективное правило синдромов. Результаты экспериментов показали, что KCOMP может повысить эффективность классификации синдромов и проверить клинические эксперты. KCOMP — это реальный способ повышения точности классификации синдромов, чтобы клинические врачи получали более надежные доказательства для повышения эффективности клинической терапии.

References

1. Frank Hutter, Lars Kotthoff, Joaquin Vanschoren. Automated Machine Learning (Methods, Systems, Challenges). -Switzerland, Springer Nature. 2019. -219 p.

2. Max Bramer. Principles of Data Mining. -London, Springer Nature. 2016. -526 p.

3. Charu C. Aggarwa. Data Mining (The Textbook). -New York, Springer Nature. 2015. -734 p.

4. Charos Khidirova. Comparative Analysis of Artificial Neural Network Training Algorithms // International Conference on Information Science and Communications Technologies. Tashkent, 2020.

5. You M, Zhao RW, Li GZ, Hu XH. MAPLSC: a novel multiclass classifier for medical diagnosis // Int J Data Min Bioin 2011; 5:383-401.

6. Zhou ZM, Wu ZH, Wang C, Feng Y. Mining both associated and correlated patterns // International Conference on Computational Science 2006; 468-475.

7. Zhong Y, Hu XL, Lu JF. The diagnosis analysis of gastritis based on association rules and decision tree // Chin J Inform Tradit Chin Med (Chin) 2008; 15:97-99.

8. Zhang BP, Li D, Wang NK. Computerized tongue diagnosis based on Bayesian networks // IEEE Transaction on Biomedical Engineering 2004; 51:1803-1810.

9. Wang H, Wang J. A quantitative diagnostic method based on Bayesian networks in traditional Chinese medicine // Leucture Notes in Computer Science 2006; 4234:176-183.

10. Wang XW, Qu HB, Liu P, Cheng YY. A self-learning expert system for diagnosis in traditional Chinese medicine // Expert Syst Appl 2004; 26:557-566.

11. Wu Y, Zhou CL, Zhang ZF. Optimize genetic algorithm of tongue diagnosis based on neural network // Appl Res Comput (Chin) 2007; 24(9):50-52.

Сведения об авторах

Хидирова Ч.М., PhD., доцент Ташкнтского университета Информационных технологий имени Мухаммада ал-Хоразмий.

Адрес для корреспонденции

100084, Узбекистан, г. Ташкент, ул. А. Темура 108. Ташкнтский университет Информационных технологий имени Мухаммада ал-Хоразмий.
тел. (+998) 90 425 23 46
e-mail: khcharos@gmail.com
Хидирова Чарос Муродиллоевна

Information about the authors

Khidirova Ch.M. PhD., docent at the Tashkent University of Information Technologies named after Muhammad al-Khwarizmi.

Address for correspondence

100084, Uzbekistan. Tashkent, A. Temur st. 108. Tashkent University of Information Technologies named after Muhammad al-Khwarizmi.
phone num. (+998) 90 425 23 46
e-mail: khcharos@gmail.com
Khidirova Charos Murodilloevna