УДК 519.22-24 Вестник СПбГУ. Математика. Механика. Астрономия. 2021. Т. 8 (66). Вып. 3
MSC 62-07, 62B10, 62H86

Симптомно-синдромальный анализ многомерных категориальных данных на основе полиномов Жегалкина*

Н. П. Алексеева

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Алексеева Н. П.* Симптомно-синдромальный анализ многомерных категориальных данных на основе полиномов Жегалкина // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2021. Т. 8 (66). Вып. 3. С. 394–405. https://doi.org/10.21638/spbu01.2021.302

В работе изучаются распределения, энтропия и другие информационные свойства конечных проективных подпространств (синдромов), параметризуемых при помощи импульсных последовательностей с базовыми элементами в виде полиномов Жегалкина над полем характеристики два (симптомов). Доказано, что суперсиндромы, полученные при рассмотрении в качестве базовых элементов мультипликативного синдрома, замкнуты. Классы симптомов, упорядоченные по мажорированию, то есть нейтральности одного из симптомов при конъюнкции, образуют мажорированный синдром, для которого доказано свойство идентичности синдрома и суперсиндрома. Сформулированные в первой части работы утверждения используются для обоснования сходимости итерационной процедуры (ИП), в которой наиболее информативные симптомы, отобранные из частичных суперсиндромов меньшей размерности, вновь подаются на вход. Стационарное состояние ИП достигается в случае принадлежности всех элементов входного множества или одному и тому же частичному суперсиндрому, или мажорированному синдрому. Благодаря ИП удается выделять наиболее информативные симптомы из большой совокупности переменных с меньшей трудоемкостью. На примере из фтизиатрии показано, каким образом при помощи симптомного анализа можно улучшить специфичность классификации.

Ключевые слова: многомерный анализ категориальных данных, конечные геометрии, алгебраические нормальные формы, энтропия, коэффициент неопределенности, итерационная процедура, симптомно-синдромальный метод, редукция размерности, классификация, чувствительность, специфичность.

1. Введение. В данной статье речь идет об усовершенствовании специального метода изучения совокупного влияния большого количества факторов, под которыми понимаются случайные категориальные переменные[1], на результаты измерений различной природы: это могут быть или метрическая переменная с гауссовской ошибкой, или переменная типа класс, или цензурированные данные типа времени жизни и т. п. Для отдельного фактора в зависимости от типа изучаемой переменной может быть применен адекватный статистический критерий, на основе которого

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 20-01-00096-а).

[1] Случайные категориальные переменные — это дискретные случайные величины, принимающие конечное число значений с заданными вероятностями.

делается о вывод о значимом или незначимом влиянии фактора на эту переменную. Например, в случае дискретной переменной типа класс могут использоваться или статистика точного критерия Фишера, или коэффициент неопределенности, для переменной непрерывного типа — критерии однородности, в анализе данных типа времени жизни — лог-ранговый критерий или критерий Гехана — Вилкоксона. Наиболее информативными будем считать факторы, при которых соответствующие статистики достигают наибольшего значения.

Закономерно возникает вопрос о существовании других более информативных скрытых признаков, отражающих совокупное влияние исходных факторов на изучаемую переменную. Предположим, что все n факторов имеют одинаковое число градаций, которые кодируются числами $0, 1, \ldots, q-1$, где q — степень простого числа. В этом случае всевозможные значения факторов образуют конечную аффинную геометрию E_n^q, а точкам двойственной проективной геометрии P_{n-1}^q соответствуют, согласно [1], всевозможные пучки параллельных гиперплоскостей из E_n^q, определяемых линейными комбинациями n факторов над конечным полем $\mathbb{F}_q$. Основная задача заключается в поиске дискретного функционала со значениями $0, 1, \ldots, q-1$, заданного в конечном линейном пространстве точек E_n^q, который является оптимальным с точки зрения заданного критерия. Вследствие двойственности конечных геометрий, этот функционал может быть выражен в виде линейной [2] или полиномиальной комбинации дискретных случайных величин (факторов) над конечным полем $\mathbb{F}_q$. При выделении нескольких наименее связанных друг с другом функционалов с наилучшими экстремальными свойствами в соответствии с заданным критерием удается решить задачу редукции размерности, которая с точки зрения конечных геометрий означает переход к более информативным подпространствам меньшей размерности.

Такой подход оказывается несколько предпочтительнее тех средств, которые обычно используются при проверке значимости совокупного влияния нескольких факторов на результаты наблюдений. Например, при большом количестве факторов многофакторные модели дисперсионного анализа становятся настолько сложными, что для адекватной оценки параметров необходимы выборки слишком большого объема [3]. Методы многомерного шкалирования, в основе которого лежит предварительная оцифровка категориальных переменных с присвоением категориям числовых значений и последующим применением многомерных статистических методов, хуже улавливают эффекты взаимодействия факторов [4], а построение дерева решений (decision trees) не подходит для задач редукции размерности [5].

Первоначально функциональный метод [6–10] предназначался для выявления наиболее значимых сочетаний факторов только в виде их линейных комбинаций над полями $\mathbb{F}_2$ и $\mathbb{F}_3$ и применялся при решении ряда биометрических задач [11–16]. Но оказалось, что можно достичь больших результатов в плане оптимизации по заданному критерию, если по крайней мере над полем $\mathbb{F}_2$ помимо линейных комбинаций рассматривать полиномы или алгебраические нормальные формы вида [17]

$$P = a + \sum_{k=1}^{n} \sum_{i_1 \le \ldots \le i_k} a_{i_1 \ldots i_k} x_{i_1} \cdot \ldots \cdot x_{i_k} \pmod 2, \qquad \text{где } a, a_{i_1 \ldots i_k} \in \{0, 1\}.$$

Они были предложены в 1927 г. И. Жегалкиным в качестве удобного средства для представления функций булевой логики. Например, если нужно идентифицировать наличие одного признака x_1 при отсутствии другого x_2, то можно использовать

полином $x_1 + x_1x_2 \,(\mathrm{mod}\, 2)$, для описания логической суммы двух бинарных признаков — полином $x_1+x_2+x_1x_2 \,(\mathrm{mod}\, 2)$ и т. д. Для преобразования логических функций в полиномы над $\mathbb{F}_2$ используются методы, широко представленные в дискретной математике: метод суммирования, метод треугольника, использование карты Карно, быстрое преобразование Фурье [18]. Для статистического анализа категориальных данных важна идентификация значимых сочетаний факторов, поэтому актуальными оказываются другие задачи: преобразование полиномов в логические функции и параметризация полиномов Жегалкина для организации их по возможности быстрого перебора.

Для снижения вычислительной нагрузки можно было бы использовать алгоритм быстрого перечисления точек грассманиана, предложенный в работе [19], но его оказалось недостаточно. В связи с этим была предпринята попытка применить итерационную процедуру, согласно которой наиболее информативные полиномы заданного порядка подаются вновь на вход в качестве базовых. Это обеспечивает увеличение степени и разнообразия полиномов при их неполном переборе. На практике достаточно быстро эта процедура приводится к стационарному состоянию, при котором отобранные латентные признаки образуют матрицу единичного ранга.

Если кратко описывать структуру данной работы, то после введения основных понятий, описания способа построения полиномов Жегалкина по набору базовых переменных при помощи импульсных последовательностей и приведения примера, иллюстрирующего преимущества симптомно-синдромального метода в анализе биометрических данных, изучаются итерационные свойства этих полиномов, такие как замкнутость и эквивалентность, необходимые для объяснения сходимости итерационной процедуры и одноранговости матрицы стационарного решения. Наибольшее внимание уделено свойству мажорированности, означающему инвариантность одних полиномов относительно умножения на другие. Будет показано, что полиномы, упорядоченные по мажорированию, образуют конечную проективную геометрию, обеспечивающую дополнительный вариант сходимости итерационной процедуры.

2. Симптомно-синдромальный анализ. ***2.1. Основные определения.*** Обозначим через $\mathbb{X}_k = (X_0, X_1, \ldots, X_{k-1})$ случайный вектор, компоненты которого соответствуют некоторым дихотомическим переменным. В первоначальных работах автора по этой тематике случайная величина в виде конечно-линейной комбинации $\mathcal{L}(X_0, \ldots, X_{k-1}) = \alpha_0 X_0 + \ldots + \alpha_{k-1}X_{k-1} (\mathrm{mod}\, 2)$, где $\alpha_i \in \{0; 1\}$, которая очевидно принимает только значения 0 или 1, была названа симптомом.[2] Симптом $\emptyset$, принимающий значение 0 с вероятностью 1, назывался вырожденным. Совокупность факторов $\{X_0,\ X_1,\ X_{01} = X_0 + X_1 (\mathrm{mod}\, 2)\}$ определялась как синдром первого порядка. Здесь уместно провести аналогию с основными понятиями конечных проективных геометрий: симптомы — точки (пространства размерности ноль) и синдромы первого порядка — прямые (размерность один). Так же, как в конечной геометрии прямую можно провести через две точки, а каждая прямая содержит не менее трех точек, синдром первого порядка образуют любые два симптома, а третий симптом $X_{01} = X_0 + X_1 (\mathrm{mod}\, 2)$ формируется из них однозначным образом. Симптом X_{01} соответствует объединению без пересечения (исключающее или) и интерпретируется

[2]Актуальность прикладного аспекта потребовала введения новой терминологии. По настоятельной просьбе врачей, с которомы пришлось тогда работать, выражение «линейная комбинация над конечным полем характеристики два» было заменено на «симптом», означающий признак или некоторое сочетание признаков.

как диссоциированность факторов, то есть наличие какого-то одного фактора при отсутствии другого.

Для обозначения синдрома первого порядка, построенного по компонентам вектора $\mathbb{X}_2 = (X_0, X_1)$, удобно использовать вектор $S(\mathbb{X}_2) = S(X_0, X_1) = (X_0, X_1, X_{01})$, а для обозначения количества элементов в синдроме и порядка синдрома, равного количеству линейно независимых компонент без единицы, — выражения $|S(X_0, X_1)| = 3$ и $||S(X_0, X_1)|| = 1$ соответственно. Очевидно, векторы $S(X_0, X_{01}) = (X_0, X_{01}, X_1)$ и $S(X_0, X_1) = (X_0, X_1, X_{01})$ отличаются только перестановкой компонент, которые относятся к одному и тому же синдрому $\{X_0,\, X_1,\, X_{01}\}$.

Синдромы большего порядка $S(\mathbb{X}_{k+1})$, где $k > 1$, полученные из компонент вектора $\mathbb{X}_{k+1} = (X_0, X_1, \ldots, X_k)$, достаточно просто выражаются через импульсные последовательности [20], которыми называют построенные рекуррентным образом последовательности элементов конечного поля. Для этого нужно рассмотреть элемент $X_k \notin S(\mathbb{X}_k)$ и прибавить его к каждой компоненте вектора $S(\mathbb{X}_k)$ над полем $\mathbb{F}_2$. Полученный вектор обозначим через $S(\mathbb{X}_k) + X_k (\mathrm{mod}\, 2)$. Тогда $S(\mathbb{X}_{k+1})$ есть случайный дихотомический вектор размерности $2^{k+1} - 1$ вида

$$S(\mathbb{X}_{k+1}) = (S(\mathbb{X}_k), X_k, S(\mathbb{X}_k) + X_k (\mathrm{mod}\, 2))\,. \tag{1}$$

Будем использовать выражение $S(\mathbb{X}_k) \sqsubseteq S(\mathbb{X}_n)$ для обозначения ситуации, когда компоненты вектора $S(\mathbb{X}_k)$ являются компонентами вектора $S(\mathbb{X}_n)$. Очевидно, что если $\mathbb{X}_k \sqsubseteq \mathbb{X}_n$, то $S(\mathbb{X}_k) \sqsubseteq S(\mathbb{X}_n)$. Если $S(\mathbb{X}_k) \sqsubseteq S(\mathbb{X}_n)$ и $S(\mathbb{X}_k) \sqsupseteq S(\mathbb{X}_n)$, то синдромы состоят из одних и тех же элементов и могут отличаться перестановками, частотами встречаемости и наличием или отсутствием вырожденного симптома, — для этого используем обозначение подобия $S(\mathbb{X}_n) \simeq S(\mathbb{X}_k)$. Например, если в (1) использовать $X_k \in S(\mathbb{X}_k)$, то $S(\mathbb{X}_{k+1}) \simeq S(\mathbb{X}_k)$. Синдром $S(\tilde{\mathbb{X}}_k)$, где $\tilde{\mathbb{X}}_k \sqsubseteq S(\mathbb{X}_n)$, будем называть *частичным.*

Для параметризации и упорядочивания полиномов Жегалкина используется двухэтапная конструкция, в которой берется аналогичное (1) выражение, где операция сложения заменяется умножением. Тем самым получается мультипликативный синдром

$$V(\mathbb{X}_{k+1}) = (V(\mathbb{X}_k), X_k, V(\mathbb{X}_k) \cdot X_k (\mathrm{mod}\, 2)), \tag{2}$$

где $X_k \notin V(\mathbb{X}_k)\,, V(\mathbb{X}_1) = V(X_0) = X_0,\ V(\mathbb{X}_2) = V(X_0, X_1) = (X_0, X_1, X_0X_1 (\mathrm{mod}\, 2))\,.$ Если все элементы вектора $\mathbb{X}_k$ невырожденные и ни один симптом из $X_0, \ldots, X_{k-1}$ не является произведением других, то мультипликативный синдром $V(\mathbb{X}_k)$ состоит из $K = 2^k - 1$ их всевозможных произведений. При использовании в (1) в качестве базовых элементов $V(\mathbb{X}_k)$ мы получим случайный вектор размерности $2^K - 1$, который содержит всевозможные полиномы степени не выше k и называется *суперсиндромом* $SV(\mathbb{X}_k) = S(V(\mathbb{X}_k))$. В дальнейшем для краткости, чтобы не отягощать текст приставками «супер», элементы суперсиндрома будем называть *симптомами* или по необходимости *нелинейными симптомами*, а конечно-линейные комбинации над полем характеристики два, ранее определяемые как симптомы, *линейными симптомами.*

Пример 1. В легких курильщиков, по литературным данным [21], наблюдаются нарушения вентиляционной способности легких обструктивного характера. Патогенетическую основу данного типа нарушений связывают с повышенной экспрессией

матриксных металлопротеиназ (ММП) — семейства внеклеточных цинк-зависимых эндопептидаз, способных разрушать белки внеклеточного матрикса. Однако при наличии сопутствующих факторов эту зависимость иногда не удается подтвердить напрямую на основе статистических наблюдений. Например, при изучении данных о курящих и некурящих больных туберкулезом [22] можно было предположить, что курение является отягощающим фактором, тем не менее их различие по экспрессии желатиназы ММП-9 оказалось незначимым — $p = 0.08$ согласно критерию Вилкоксона. В то же время значимое почти двухкратное различие с доверительным уровнем вероятности $p = 0.0016$ наблюдалось при сравнении уровней желатиназы ММП-9 по симптому $X_0 + X_1X_2 + X_0X_1X_2 \pmod 2$, идентифицирующему составную группу риска, в которую наряду со всеми курящими (X_0) входят наиболее тяжелые больные с нарушением функции внешнего дыхания (X_1) при наличии множественных туберкулом в легких (X_2).

2.2. Замкнутость суперсиндромов. В суперсиндроме, построенном по элементам одного и того же суперсиндрома, не могут появиться какие-то иные элементы кроме элементов исходного суперсиндрома. Формально это свойство замкнутости можно выразить в виде следующих утверждений.

Лемма 1. *Пусть $\mathbb{Y}_k^m$ вектор с k компонентами, из которых m линейно независимы, $\mathbb{Y}_k^m \sqsubseteq S(\mathbb{X}_n)$ и $\mathbb{Y}_m^m \sqsubseteq \mathbb{Y}_k^m$, тогда: 1) $S(\mathbb{Y}_k^m) \simeq S(\mathbb{Y}_m^m)$, 2) $S(\mathbb{Y}_k^m) \sqsubseteq S(\mathbb{X}_n)$.*

Доказательство. Пусть $\mathbb{Y}_k^m \simeq (\mathbb{Y}_m^m, F(\mathbb{Y}_k^m \setminus \mathbb{Y}_m^m))$, где $F = F(\mathbb{Y}_k^m \setminus \mathbb{Y}_m^m) = (f_1, \ldots, f_{k-m})$ — вектор с компонентами, линейно выражаемыми через элементы $\mathbb{Y}_m^m$, то есть $f_j \in S(\mathbb{Y}_m^m)$. Тогда для любых $f_j, s_i \in S(\mathbb{Y}_m^m)$ над полем $\mathbb{F}_2$ справедливо $f_j + s_i \in S(\mathbb{Y}_m^m)$. Следовательно, $S(\mathbb{Y}_m^m) \sqsubseteq S(\mathbb{Y}_k^m) = S(\mathbb{Y}_m^m, F(\mathbb{Y}_k^m \setminus \mathbb{Y}_m^m)) \sqsubseteq S(\mathbb{Y}_m^m)$, то есть $S(\mathbb{Y}_k^m) \simeq S(\mathbb{Y}_m^m)$. Так как $S(\mathbb{Y}_m^m) \sqsubseteq S(\mathbb{X}_n)$, то и $S(\mathbb{Y}_k^m) \sqsubseteq S(\mathbb{X}_n)$. □

Следствие 1. *Пусть $\mathbb{Y}_k \sqsubseteq SV(\mathbb{X}_n)$, тогда $SV(\mathbb{Y}_k) \sqsubseteq SV(\mathbb{X}_n)$.*

2.3. Симптомы с одинаковыми значениями. Если в распределении синдрома $S(\mathbb{X}_k)$ есть нулевые вероятности, то в суперсиндроме $SV(\mathbb{X}_k)$ окажутся симптомы с одинаковыми значениями. Рассмотрим простейший случай синдрома первого порядка $S(X_0, X_1)$ с распределением $(p_{00}, p_{01}, p_{10}, p_{11})$, в котором, к примеру, $p_{10} = P\{X_0 = 1, X_1 = 0\} = 0$. Это приводит к тому, что симптом $X_0(1+X_1)$ оказывается вырожденным, то есть $P\{X_0(1+X_1) = 1\} = p_{10} = 0$ и $P\{X_0(1+X_1) = 0\} = 1$. Заметим, что сложение и умножение осуществляются над полем $\mathbb{F}_2$, в котором $x+x = 0$ и $x \cdot x = x$. Вырожденность симптома $X_0(1 + X_1)$ приводит к трем парам симптомов с одинаковыми значениями: $X_0 = X_0 + (X_0 + X_0X_1) = X_0X_1$, $X_1 = X_1 + X_0 + X_0X_1$ и $X_0 + X_1 = (X_0 + X_1) + (X_0 + X_0X_1) = X_1 + X_0X_1$. В результате в суперсиндроме $SV(X_0, X_1)$ вместо семи симптомов уникальными окажутся только три невырожденных и один вырожденный симптомы.

В общем виде аналогично, наличие нулевой вероятности в распределении означает справедливость некоторого уравнения $F(\mathbb{X}_n) = \emptyset$. При рассмотрении $M = 2^{2^n-1} - 1$ компонент $G(\mathbb{X}_n) \in SV(\mathbb{X}_n)$ можно обнаружить, что для каждой компоненты $G(\mathbb{X}_n) \neq F(\mathbb{X}_n)$ существует компонента $G_1(\mathbb{X}_n) \in SV(\mathbb{X}_n)$ такая, что $G_1(\mathbb{X}_n) = G(\mathbb{X}_n) + F(\mathbb{X}_n)$. Следовательно, при наличии одной нулевой вероятности количество неодинаковых невырожденных симптомов в суперсиндроме равно $\frac{M-1}{2} = 2^{2^n-2} - 1$.

Произведения симптомов могут также приводить к симптомам с одинаковыми значениями. Например, пусть a, b, c, d — некоторые бинарные случайные величины, $A = abc + b + bc + c$, $B = ab + abc + bc + c$, $C = b + bcd + c$, $D = b + bcd + bd + c$. Можно убедиться в том, что над $\mathbb{F}_2$ справедливо $AD = CD = D$. Рассмотрим подробнее это свойство симптомов быть нейтральными по отношению друг к другу через произведение.

2.4. Мажорированные симптомы и синдромы. Будем говорить, что невырожденный симптом X_j строго мажорирует невырожденный симптом $X_i \neq X_j$, если $X_i X_j = X_i$. Симптом X_i будем называть *мажорированным*. Для обозначения строгой мажорированности используем $X_i \prec X_j$. Если симптомы несовместны, т. е. $X_i X_j = \emptyset$, или справедливо $X_i \prec X_j$, будем говорить о нестрогом мажорировании $X_i \preceq X_j$ или просто мажорировании.

Лемма 2 (свойство транзитивности и линейной независимости). *Если $X_0 \prec X_1$ и $X_1 \preceq X_2$, то: 1) $X_0 \preceq X_2$; 2) $aX_1 + bX_2 = 0$ только при $a = b = 0$.*

Доказательство.

1) В случае строгого мажорирования $X_1 \prec X_2$ имеем, что если $X_0 X_1 = X_0$ и $X_1 X_2 = X_1$, то $X_0 X_2 = (X_0 X_1) X_2 = X_0 (X_1 X_2) = X_0 X_1 = X_0$, то есть $X_0 \prec X_2$. В случае $X_1 X_2 = \emptyset$ и $X_0 X_1 = X_0$ получаем $X_0 X_2 = X_0 X_1 X_2 = \emptyset$, и условие $X_0 \preceq X_2$ также оказывается выполненным.

2) Из $X_1 = X_1 X_2$ и $aX_1 + bX_2 = 0$ следует $aX_1 X_2 + bX_2 = 0$ и $X_2(aX_1 + b) = 0$. Поскольку предполагается невырожденность симптомов, последнее равенство справедливо только при нулевых коэффициентах. Если для невырожденных симптомов справедливо $X_1 X_2 = \emptyset$, тогда из четырех комбинаций коэффициентов (a, b) только в случае $(0, 0)$ имеет место $aX_1 + bX_2 = 0$, так как при $a = 0, b = 1$ имеем $X_2 = 0$ (при $a = 1, b = 0$ аналогично), или $a = 1, b = 1$ влечет за собой $X_1 = X_2$. □

Определение. Элементы вектора $\mathbb{X}_k$ удовлетворяют соотношению мажорирования, если его компоненты можно упорядочить таким образом, что $X_0 \preceq \ldots \preceq X_{k-1}$ верно при удалении любого промежуточного звена, $X_i \preceq X_j \ \forall\ i < j$. Прямая нумерация выбрана для определенности. Синдром $S(\mathbb{X}_n)$ с такими базовыми элементами будем называть *мажорированным*.

Лемма 3. *Пусть для элементов $\mathbb{X}_k$ справедливо соотношение мажорирования, тогда: 1) элементами мультипликативного синдрома являются элементы $\mathbb{X}_k$, то есть $V(\mathbb{X}_k) \sqsubseteq \mathbb{X}_k$, и $SV(\mathbb{X}_k) \sqsubseteq S(\mathbb{X}_k)$; 2) элементы $\mathbb{X}_k$ линейно независимы.*

Доказательство.

1) Компоненты мультипликативного синдрома $V(\mathbb{X}_k)$ имеют вид или $X_0, \ldots, X_{k-1}$, или произведений $\prod_{i=1}^{m} X_{\tau_i}$, $X_{\tau_i} \in \mathbb{X}_k$, которые благодаря мажорированию сводятся к мажорированным симптомам, то есть $V(\mathbb{X}_k) \sqsubseteq \mathbb{X}_k$. Вследствие этого $SV(\mathbb{X}_k) = S(V(\mathbb{X}_k)) \sqsubseteq S(\mathbb{X}_k)$, то есть мажорированный суперсиндром $SV(\mathbb{X}_k)$ содержит только линейные комбинации компонент $\mathbb{X}_k$.

2) Покажем, что выражение $a_0 X_0 + a_1 X_1 + \ldots + a_{n-1} X_{n-1} = 0$ имеет место только при всех $a_i = 0$. Если умножить это выражение на X_1, то в случае строгого мажорирования получаем равенство $a_0 X_0 X_1 + a_1 X_1 X_1 + \ldots + a_{n-1} X_{n-1} X_1 = a_0 X_0 + X_1(a_1 + \ldots + a_{n-1}) = 0$, которое справедливо согласно лемме 2 при $a_0 = 0$ и $a_1 + \ldots + a_{n-1} = 0$. Оставшееся выражение $a_1 X_1 + \ldots + a_{n-1} X_{n-1} = 0$ умножим на X_2. Получим $a_1 X_1 X_2 + \ldots + a_{n-1} X_{n-1} X_2 = 0$, откуда $a_1 X_1 + (a_2 + \ldots + a_{n-1}) X_2 = 0$ и

$a_1 = 0$, и так далее все $a_i = 0$. Аналогично умножением на разные компоненты можно показать справедливость этого утверждения при нестрогом мажорировании. □

Следствие 2. *При $k < n < 2^k - 1$ компоненты мажорированного вектора $\mathbb{X}_n$ не могут принадлежать какому-то одному частичному синдрому $S(\mathbb{X}_k)$, поскольку в последнем не может быть более k линейно независимых симптомов.*

Теорема. *Пусть $S(\mathbb{X}_n)$ — мажорированный синдром и $\mathbb{Y}_k \sqsubseteq S(\mathbb{X}_n)$ есть k-подмножество симптомов, по которому строится частичный суперсиндром $SV(\mathbb{Y}_k)$. Тогда: 1) $V(\mathbb{Y}_k) \sqsubseteq \mathbb{Y}_k$, 2) $SV(\mathbb{Y}_k) \sqsubseteq S(\mathbb{X}_n)$.*

Доказательство. Если $V(\mathbb{Y}_k) \sqsubseteq S(\mathbb{X}_n)$, то за счет мажорирования очевидно $V(\mathbb{Y}_k) \sqsubseteq \mathbb{Y}_k \sqsubseteq S(\mathbb{X}_n)$. Далее можно воспользоваться утверждением 2 из леммы 1, согласно которому элементы частичного синдрома являются элементами все того же синдрома $SV(\mathbb{Y}_k) = S(V(\mathbb{Y}_k)) \sqsubseteq S(\mathbb{Y}_k) \sqsubseteq S(\mathbb{X}_n)$. □

2.5. Построение упорядоченных по мажорированию симптомов. Выясним, каким образом можно получить упорядоченные по мажорированию симптомы и, как следствие, мажорированный синдром. Пусть имеются четыре бинарные переменные a, b, c, d, разбитые по тройкам: (a, b, c), (a, b, d), (a, c, d), (b, c, d), из которых соответственно формируются симптомы A, B, C, D.

Предложение. *Пусть симптомы A, B являются элементами суперсиндрома $SV(a, b, c)$, а симптомы C, D — элементами суперсиндрома $SV(b, c, d)$, и выполнены условия внутреннего мажорирования $A \prec D$, $C \prec B$. Рассмотрим $\alpha_1 \prec \alpha_2 \in SV(A, C)$ и $\beta_1 \prec \beta_2 \in SV(B, D)$. Тогда $\alpha_1(A, C) \prec \alpha_2(A, C) \prec \beta_1(B, D) \prec \beta_2(B, D)$.*

Доказательство. По условию $\alpha_1(A, C) \prec \alpha_2(A, C)$ и $\beta_1(B, D) \prec \beta_2(B, D)$. Поскольку $A \prec D$, $C \prec B$, имеем $\alpha_2(A, C) \prec \beta_1(B, D)$, отсюда по свойству транзитивности (лемма 2) получаем $\alpha_1(A, C) \prec \alpha_2(A, C) \prec \beta_1(B, D) \prec \beta_2(B, D)$. □

Пример 2. Можно непосредственно убедиться, что над $\mathbb{F}_2$ при $A = abc + c$ и $B = a + ab + abc + ac + c$ имеем $AB = A$, а при $C = ac + acd + c$ и $D = acd + ad + c$, соответственно, $CD = C$. Кроме того, $AD = A, BC = C$. Рассматривая $\alpha_1 = A$, $\alpha_2 = A + C + AC = abc + abcd + c$, $\beta_1 = BD = abcd + abd + acd + ad + c$, $\beta_2 = B$, получаем $\alpha_1\alpha_2 = A(A + C + AC) = A = \alpha_1$, $\beta_1\beta_2 = BDB = BD = \beta_1$, $\alpha_2\beta_1 = (A + C + AC)BD = A + C + AC = \alpha_2$, откуда $\alpha_1 \prec \alpha_2 \prec \beta_1 \prec \beta_2$.

3. Итерационная процедура отбора наиболее информативных симптомов. Предназначенная для того, чтобы обойти полный перебор симптомов, итерационная процедура определяется тремя параметрами: p — число исходных переменных, k — наивысшая степень полинома Жегалкина, соответственно $k - 1$ – порядок частичных синдромов, по которым ищутся наиболее информативные симптомы, и $N > k$ — количество наилучших симптомов, отобранных со всех частичных синдромов для следующего этапа.

На начальном этапе рассматривается входное множество из p дихотомических переменных. Из них выделяются C_p^k сочетаний переменных, которые принимаются в качестве базовых для частичных суперсиндромов порядка $k - 1$. Для каждого из $2^K - 1$ элементов частичного суперсиндрома, где $K = 2^k - 1$, вычисляется его

значимость — в случае дискретной зависимой переменной можно использовать коэффициенты неопределенности, а для метрических переменных или для кривых дожития — статистики соответствующих критериев однородности. Наиболее значимые N симптомов рассматриваются как элементы выходного множества на следующем этапе. Если выходное множество совпадает со входным множеством, достигается стационарное состояние.

Итерационная процедура сходится, когда элементы входного множества принадлежат одному и тому же синдрому: частичному или мажорированному. По вышеприведенной теореме и следствию 1 эти синдромы замкнуты, соответственно, экстремальные симптомы останутся теми же, что и на предыдущем шаге. В стационарном состоянии среди C_N^k сочетаний, используемых для построения частичных суперсиндромов, C_{N-1}^{k-1} сочетаний содержат первый по информативности симптом[3], следовательно, в каждом из этих частичных суперсиндромов самым информативным окажется один и тот же симптом. В результате, если не предусмотреть проверку наиболее информативных симптомов на предмет их эквивалентности, матрица, составленная из N упорядоченных по информативности суперсимптомов, при $N < C_{N-1}^{k-1}$ окажется однoранговой.

Пример 3. Применим эту процедуру для изучения структуры зависимости между морфологической активностью патологического процесса у больных с туберкуломой легких и наличием ряда факторов: бактериовыделение (a), казеозный некроз (b), специфические грануляции в стенке капсуле туберкуломы (c), неспецифические воспалительные изменения бронхиального дерева (d), табакокурение (e), нарушение функции внешнего дыхания (f). Статистические критерии, примененные к указанным характеристикам по отдельности, указывают на высокую чувствительность прогнозирования порядка 0.95, но очень низкую специфичность на уровне 0.45–0.64.

На первом этапе итерационной процедуры при $p = 6$ и $k = 3$ были выявлены факторы $U = a+ac+acf+af+c = a\bar{f}\bar{c}+c$ и $V = bcf+bf+c$, коэффициенты неопределенности для которых равны 57.85 и 46.32 % соответственно. При помощи первого фактора U (или c, или $\bar{c}a\bar{f}$) отделились больные, у которых или были специфические грануляции в капсуле, или если их не было и отсутствовали нарушения функции внешнего дыхания, но имелись бактериовыделения. В этом случае из 30 больных у 29 был активный процесс, а в противоположном случае из 10 больных активный процесс был только у двоих, значимость отличия по точному критерию Фишера [23] равна $p = 10^{-6}$. Второй фактор V (или c, или bf) означал ассоциированность активности патологического процесса или через наличие специфических грануляций, или через сочетание казеозного некроза с нарушением функции внешнего дыхания. Этот фактор встречался у 29 больных, из которых 28 имели активный процесс, в противоположном случае из 13 больных активный процесс был у четверых, значимость этого различия, согласно точному критерию Фишера, равна $p = 10^{-5}$. На втором этапе при параметре $N = 10$ был получен целый комплекс довольно сложных факторов с практически одинаковыми информационными характеристиками на уровне 70%-го коэффициента неопределенности. Проще всего оказалось выделить сочетание $W = U + V + UV$, которое означает наличие хотя бы одного из факторов U

[3] В сбалансированной блок-схеме, содержащей N элементов, распределенных по $b = C_N^k$ блокам, состоящим из k элементов, каждый элемент встречается $r = C_{N-1}^{k-1}$ раз (из соотношения баланса $Nr = bk$).

или V. При $W = 0$ и $W = 1$ соответственно вероятности активного процесса были равны 0.1 (один из 10) и 0.97 (30 из 31), $p = 10^{-7}$. Таким образом, при помощи двух симптомно-синдромальных итераций по небольшому числу переменных удалось достичь 90%-й специфичности.

4. Заключение. Проблема статистического анализа большого количества факторов актуальна для большинства медико-биологических исследований. Зачастую объективную неравновесность сочетаний факторов врачам приходится учитывать на интуитивном уровне, формируя синдромы по наитию. Параметризация всевозможных логических сочетаний полиномами Жегалкина позволяет формализовать этот подход для выявления наиболее значимых и устойчивых интегральных факторов-симптомов, агрегированных в синдромы и суперсиндромы. Для снижения трудоемкости возможен переход от полного перебора конечных подпространств к итерационной процедуре решения экстремальной задачи по пространствам меньшей размерности. Анализ сходимости этой процедуры показал, что причиной достижения стационарного состояния может быть не только снижение размерности синдромов, но и их мажорированность. Локальные экстремумы, выявленные таким образом, имеют самостоятельное значение, однако вопрос, насколько они далеки от глобального экстремума, пока остается открытым.

Симптомы позволяют установить неоднородный характер групп риска или выявить структуру латентных факторов. Варианты использования их в качестве дополнительных факторов для улучшения классификации и прогнозирования показаны на примере анализа фтизиатрических данных.

Автор благодарит старшего научного сотрудника биохимической лаборатории Санкт-Петербургского института фтизиопульмонологии кандидата биологических наук Д. С. Эсмедляеву за предоставленные данные и консультирование по вопросам терминологии и интерпретации результатов статистического анализа.

Литература

1. Холл М. *Комбинаторика*, пер. с англ. Москва, Мир (1970).

2. Алексеева Н. П. *Анализ медико-биологических систем. Реципрокность, эргодичность, синонимия.* Санкт-Петербург, Изд-во С.-Петерб. ун-та (2013).

3. Шеффе Г. *Дисперсионный анализ*, пер. с англ. Москва, Наука (1980).

4. Дэйвисон М. *Многомерное шкалирование: методы наглядного представления данных*, пер. с англ. Москва, Финансы и статистика (1988).

5. Moret B. M. E. Decision trees and diagrams. *Computing Surveys* **14**, 593–623 (1982). https://doi.org/10.1145/356893.356898

6. Алексеева Н. П., Алексеев А. О. О роли конечных геометрий в корреляционном анализе бинарных признаков. В: М. К. Чиркова (ред.) *Математические модели. Теория и приложения*, вып. 4, 102–117. Санкт-Петербург (2004).

7. Алексеева Н. П., Конради А. О., Бондаренко Б. Б. Симптомный анализ в исследовании долгосрочного клинического прогноза. *Артериальная гипертензия* **14** (1), 38–43 (2008).

8. Alexeyeva N., Smirnov I., Gracheva P., Martynov B. The finitely geometric symptom analysis in the glioma survival study. *Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics*, 2009, Tianjin, 1–4 (2009). https://doi.org/10.1109/BMEI.2009.5305560

9. Alexeyeva N., Gracheva P., Podkhalyuzina E., Usevich K., Alexeyev A. Symptom and syndrome analysis of categorical series, logical principles and forms of logic. *Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics*, 2010, Yantai, 2603–2606 (2010).

10. Alexeyeva N. P., Al-Juboori F. S., Skurat E. P. Symptom analysis of multidimensional categorical data with applications. *Periodicals of Engineering and Natural Sciences* **8** (3), 1517–1524 (2020).

11. Алексеева Н. П., Иванова Е. П., Митрофанова Л. Б., Кулешова Э. В., Енькина Т. Н., Гордеев М. Л., Ротенко М. М., Бондаренко Б. Б. О способе улучшения прогнозирования вазоспазма

лучевой артерии на основе симптомного расслоения популяций. *Ученые записки СПбГМУ им. акад. И. П. Павлова* **16** (4), 59–62 (2009).

12. Холявин А. И., Низковолос В. Б., Мартынов Б. В., Свистов Д. В., Аничков А. Д., Алексеева Н. П. Возможности использования криохирургической методики при лечении больных с глубинными опухолями головного мозга. *Вестник хирургии им. И. И. Грекова* **175** (1), 11–17 (2016).

13. Мартынов Р. С., Гайдар Б. В., Парфенов В. Е., Мартынов Б. В., Свистов Д. В., Алексеева Н. П. Осложнения раннего послеоперационного периода рецидивных глиом головного мозга супратенториальной локализации. *Нейрохирургия*, (2), 30–36 (2016).

14. Кибитов А. О., Крупицкий Е. М., Блохина Е. А., Вербицкая Е. В., Бродянский В. М., Алексеева Н. П., Бушара Н. М., Ярославцева Т. С., Палаткин В. Я., Масалов Д. В., Бураков А. М., Романова Т. Н., Сулимов Г. Ю., Гриненко А. Я., Костен Т., Ниелсен Д., Звартау Э. Э. Фармакогенетический анализ влияния генов дофаминовой и опиоидной систем на эффективность комбинированной терапии налтрексоном и гуанфацином больных опиоидной зависимостью. *Журнал неврологии и психиатрии им. С. С. Корсакова* **16** (11), 36–48 (2016).

15. Мартынов Р. С., Мартынов Б. В., Бабичев К. Н., Гаврилов Г. В., Чемодакова К. А., Свистов Д. В., Алексеева Н. П. Влияние сроков повторных оперативных вмешательств на радикальность и выживаемость у пациентов с рецидивными опухолями различной степени злокачественности супратенториальной локализации. В: *Сборник научных работ III Петербургского Международного онкологического форума «Белые ночи 2017»*. ФГБУ «НИИ онкологии им. Н. Н. Петрова» Минздрава России (2017).

16. Алексеева Н. П., Горлова И. А., Бондаренко Б. Б. Прогнозирование потребности в госпитализациях после кардиохирургического вмешательства на основе симптомно-синдромального структурирования факторов. *Трансляционная медицина* **6** (6), 14–22 (2019).

17. Яблонский С. В. *Введение в дискретную математику*. Москва, Наука (1986).

18. Супрун В. П. *Основы теории булевых функций*. Москва, Ленанд (2017).

19. Ананьевская П. В. *Исследование конечно-линейных статистических моделей. Оптимизация и избыточность*. Дисс. . . . канд. физ.-мат. наук. Санкт-Петербург (2013).

20. Лидл Р., Нидеррайтер Г. *Конечные поля*, пер. с англ. Т. 1, 2. Москва, Мир (1988).

21. Navratilova Z., Kolek V., Petrek M. Matrix Metalloproteinases and Their Inhibitors in Chronic Obstructive Pulmonary Disease. *Archivum Immunologiae et Therapiae Experimentalis* **64**, 177–193 (2016). https://doi.org/10.1007/s00005-015-0375-5

22. Эсмедляева Д. С., Алексеева Н. П., Сапожникова Н. В., Дьякова М. Е., Перова Т. Л., Кирюхина Л. Д., Журавлев В. Ю. Система матриксные металлопротеиназы-ингибиторы при инфильтративном туберкулезе легких и ее роль в оценке интенсивной фазы терапии. *Биомедицинская химия* **62** (5), 593–598 (2016).

23. Mehta C. R., Patel N. R. Exact inference in categorical data. *Biometrics* **53** (1), 112–117 (1997).

Статья поступила в редакцию 18 июля 2020 г.;
после доработки 21 октября 2020 г.;
рекомендована в печать 19 марта 2021 г.

Контактная информация:

Алексеева Нина Петровна — канд. физ.-мат. наук, доц.; nina.alekseeva@spbu.ru

The symptom-syndrome analysis of multivariate categorical data based on Zhegalkin polynomials*

N. P. Alekseeva

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Alekseeva N. P. The symptom-syndrome analysis of multivariate categorical data based on Zhegalkin polynomials. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2021, vol. 8 (66), issue 3, pp. 394–405.
https://doi.org/10.21638/spbu01.2021.302 (In Russian)

*This work is supported by Russian Foundation for Basic Research (grant No. 20-01-00096-a).

In this article, we study the distribution, entropy and other informational properties of finite projective subspaces (syndromes) parameterized by impulse sequences with basic elements in the form of symptoms — polynomials over the field $\mathbb{F}_2$ which are known as Zhegalkin polynomials. It has been proven that the super syndrome, which is a linear syndrome with basic elements in the form of a multiplicative syndrome, is closed. If in the multiplication of two symptoms one is neutral, then we are talking about its majorization. The ordered by majorization symptoms form a majorized syndrome. Is proved that the majorized syndrome is closed and coincides with the super syndrome. The statements formulated in the first part of the paper are used to justify the convergence of the iterative procedure (PI), in which the most informative symptoms selected from partial super syndromes are again used in the next step. The stationary state of PI is obtained if all elements of the input set belong to either the same partial super syndrome or to the majorized syndrome. Thanks IP it is possible to quickly find the optimal syndrome from a large set of variables. An example from phthisiology shows how the specificity of classification can be improved using symptom analysis.

Keywords: multivariate analysis of categorical data, finite geometries, algebraic normal forms, entropy, uncertainty coefficient, iterative procedure, symptom-syndromic method, dimension reduction, classification, sensitivity, specificity.

References

1. Hall M. *Combinatorial theory*. Waltham, MA, Blaisdell Publ. Co. (1967). [Russ. ed.: *Kombinatorika*, Moscow, Mir Publ. (1970)].

2. Alekseeva N. P. *Analysis of biomedical systems. Reciprocity, consistency, synonymy*. St. Petersburg, St. Petersburg University Press (2013). (In Russian)

3. Sheffe H. *The analysis of variance*. New York, Wiley (1959). [Russ. ed.: *Dispersionnyj analiz*. Moscow, Nauka Publ. (1980)].

4. Davison M. *Multidimensional scaling*. Wiley (1983). [Russ. ed.: *Mnogomernoe shkalirovanie: metody nagljadnogo predstavlenija dannyh*. Moscow, Finansy i statistika (1988)].

5. Moret B. M. E. Decision trees and diagrams. *Computing Surveys* **14**, 593–623 (1982). https://doi.org/10.1145/356893.356898

6. Alekseeva N. P., Alekseev A. O. On the role of finite geometries in the correlation analysis of binary features. In: M. K. Chirkova (ed.) *Mathematical models. Theory and applications*, iss. 4. St. Petersburg, 102–117 (2004). (In Russian)

7. Alekseeva N. P., Konradi A. O., Bondarenko B. B. Symptom Analysis in Long-Term Clinical Prognosis Research. *Arterial hypertension* **14** (1), 38–43 (2008). (In Russian)

8. Alexeyeva N., Smirnov I., Gracheva P., Martynov B. The finitely geometric symptom analysis in the glioma survival study. *Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics*, 2009, Tianjin, 1–4 (2009). https://doi.org/10.1109/BMEI.2009.5305560

9. Alexeyeva N., Gracheva P., Podkhalyuzina E., Usevich K., Alexeyev A. Symptom and syndrome analysis of categorical series, logical principles and forms of logic. *Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics*, 2010, Yantai, 2603–2606 (2010).

10. Alexeyeva N. P., Al-Juboori F. S., Skurat E. P. Symptom analysis of multidimensional categorical data with applications. *Periodicals of Engineering and Natural Sciences* **8** (3), 1517–1524 (2020).

11. Alekseeva N. P., Ivanova E. P., Mitrofanova L. B., Kuleshova E. V., En'kina T. N., Gordeev M. L., Rotenko M. M., Bondarenko B. B. On the method of improving the prediction of radial artery vasospasm based on symptomatic stratification of populations. *Scientific notes of SPbGMU im. acad. I. P. Pavlova* **16** (4), 59–62 (2009). (In Russian)

12. Kholyavin A. I., Nizkovolos V. B., Martynov B. V., Svistov D. V., Anichkov A. D., Alekseeva N. P. Possibilities of using cryosurgical techniques in the treatment of patients with deep brain tumors. *Bulletin of surgery named after I. I. Grekov* **175** (1), 11–17 (2016). (In Russian)

13. Martynov R. S., Gaidar B. V., Parfenov V. E., Martynov B. V., Svistov D. V., Alekseeva N. P. Complications of the early postoperative period of recurrent brain gliomas of supratentorial localization. *Neurosurgery*, (2), 30–36 (2016). (In Russian)

14. Kibitov A. O., Krupitsky E. M., Blokhina E. A., Verbitskaya E. V., Brodyansky V. M., Alekseeva N. P., Bushara N. M., Yaroslavtseva T. S., Palatkin V. Ya., Masalov D. V., Burakov A. M., Romanova T. N., Sulimov G. Yu., Grinenko A. Ya., Kosten T., Nielsen D., Zvartau E. E. Pharmacogenetic

analysis of the effect of genes of the dopamine and opioid systems on the effectiveness of combination therapy with naltrexone and guanfacine in patients with opioid dependence. *Journal of Neurology and Psychiatry* **16** (11), 36–48 (2016).

15. Martynov R. S., Martynov B. V., Babichev K. N., Gavrilov G. V., Chemodakova K. A., Svistov D. V., Alekseeva N. P. Influence of the timing of repeated surgical interventions on radicality and survival in patients with recurrent tumors of varying degrees of malignancy of supratentorial localization. In: *Collection of scientific papers of the III St. Petersburg International Oncological Forum "White Nights 2017"*. N. N. Petrov Research Institute of Oncology, Ministry of Health of Russia (2017). (In Russian)

16. Alekseeva N. P., Gorlova I. A., Bondarenko B. B. Predicting the need for hospitalizations after cardiac surgery based on symptom-syndromic structuring of factors. *Translational medicine* **6** (6), 14–22 (2019). (In Russian)

17. Yablonskiy S. V. *Introduction to discrete mathematics*. Moscow, Nauka Publ. (1986). (In Russian)

18. Suprun V. P. *Foundations of the theory of Boolean functions*. Moscow, Lenand Publ. (2017). (In Russian)

19. Ananievskaya P. V. *Investigation of finite-linear statistical models. Optimization and redundancy*. PhD thesis. St. Petersburg (2013). (In Russian)

20. Lidl R., Niederreiter G. *Finite fields*. Cambrige University Press (1997). [Russ. ed.: *Konechnye polja*. T. 1, 2. Moscow, Mir Publ. (1988)].

21. Navratilova Z., Kolek V., Petrek M. Matrix Metalloproteinases and Their Inhibitors in Chronic Obstructive Pulmonary Disease. *Archivum Immunologiae et Therapiae Experimentalis* **64**, 177–193 (2016). https://doi.org/10.1007/s00005-015-0375-5

22. Esmedlyaeva D. S., Alekseeva N. P., Sapozhnikova N. V., Dyakova M. E., Perova T. L., Kiryukhina L. D., Zhuravlev V. Yu. Matrix metalloproteinase inhibitors system in infiltrative pulmonary tuberculosis and its role in assessing the intensive phase of therapy. *Biomedical Chemistry* **62** (5), 593–598 (2016). (In Russian)

23. Mehta C. R., Patel N. R. Exact inference in categorical data. *Biometrics* **53** (1), 112–117 (1997).

Received: July 18, 2020
Revised: October 21, 2020
Accepted: March 19, 2021

Author's information:

Nina P. Alekseeva — nina.alekseeva@spbu.ru