Программные средства анализа качественных данных на основе детерминационного анализа

В. В. Нестругина, email: v.nestrugina@gmail.com
И. Е. Воронина, email: irina.voronina@gmail.com

Воронежский государственный университет

Аннотация. *Рассматривается система для анализа качественных данных на основе детерминационного анализа. Программные средства осуществляют с качественными данными в их первоначальном виде, позволяют анализировать зависимости между отдельными значениями качественных переменных, которые могут быть упущены при использовании интегральных показателей.*

Ключевые слова: *детерминационный анализ, обработка качественных данных, программные средства.*

Введение

В социологических исследованиях для обработки качественных данных применяются различные способы и средства. Задача таких инструментов – изучение структуры связей между переменными, построение вторичных агрегированных показателей [2, 3].

Известные направления обработки качественных данных:

1. Квантификация значений качественной переменной. Например, шкала удовлетворенности может быть переведена в числовую. Поборники такого метода считают номинальность переменных неразвитостью техник социального измерения, которая в конце концов сойдет на нет [4]. С другой точки зрения, нет никакого основания полагать, что проблема квантификации – поиска чисел, стоящих за словами обычного человеческого языка, будет когда-либо решена [1].

2. Использование интегральных показателей, описывающих либо тесноту связи между переменными (эмпирическими индикаторами), либо расстояние (близость) между эмпирическими объектами или их разбиениями. Минусом такого подходя является получение суждений общего характера, а не видение объекта как набор свойств [1].

3. Содержательный анализ эмпирических данных, который включает изучение процентных распределений, содержащихся в таблицах сопряженности. Таблица сопряженности – тип таблицы в формате матрицы, которая показывает частоту появления каких-либо переменных. Минусами такого подхода является то, что, во-первых,

таблицы могут быть многомерными, что усложняет их понимание и обработку; во-вторых, нет возможности манипулирования конкретными свойствами (значениями качественных признаков) [1].

Детерминационный анализ – техника по построению локального, фрагментарного анализа. Он заключается в том, чтобы по одним признакам (индикаторам) предсказать наличие других. Главной характеристикой является условная частота. Манипулируя сочетаниями отдельных свойств, можно получить точные и полные детерминации, то есть предсказания [1]. Объектом исследования являются номинальные (качественные) данные, проблема квантификации значений которых в настоящее время не решена.

Рассмотрим реализацию основных задач, стоящих перед детерминационным анализом, включающих получение объяснения свойства набором других свойств, уточнение для определенного решения, анализ существенности контекст, объясняющих и объясняемых свойств, построение объясняемой и объясняющей типологии и определение их существенности.

2. Особенности реализации программных средств для основных задач детерминационного анализа

Задача: методами детерминационного анализа описать зависимости не между переменными (признаками), а между конкретными значениями этих переменных, то есть выяснить, насколько связаны переменные и конкретные значения этих переменных. Для реализации решения были разработаны программные средства, обеспечивающие методами детерминационного анализа обработку данных csv файлов в виде веера отображений (матрицы данных).

Изначально данные для социологического исследования приходят в виде таблицы (матрицы данных), где строки – объекты (E), столбцы – переменные (отображения), пересечение строки и столбца – значение.

Первичные данные исследований представляют собой веер отображений – совокупность отображений вида $E \to X_i, i \in [1, n]$, где E – множество объектов, X_i – множество значений переменной x_i.

Веер образуют компоненты-отображения. Основание веера – множество E. Математические методы анализа социально-экономических данных – методы оперирования веерами отображений. Любое множество E, X_i обладает следующими свойствами: дискретность, конечность [1].

Разработанный пакет инструментов позволяет выполнять основные функции:

1. Построение таблицы сопряженности.

Для анализа первичные данные переводят в таблицы сопряженности (рис. 1)

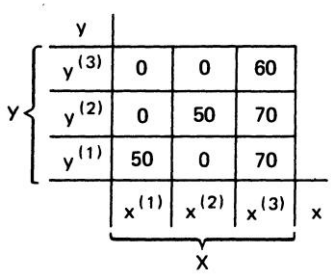

Рис. 1. Пример таблицы сопряженности между признаками (переменными) x, y

2. Вычисление интенсивности и емкости детерминации.

Оценка условной вероятности или условная эмпирическая частота является наиболее приемлемой мерой установления связи между конкретными признаками и вычисляется следующим образом:

$$P(y \mid x) = \frac{N_{x_i y_j}}{N_{x_i}} \qquad (1)$$

Жесткая детерминация – значение условной вероятности равно 1. Как правило, социальные закономерности обладают не жесткой детерминацией, а статистичностью, являющейся ограниченным нарушением детерминизма. Статистичность – мера отличности условной вероятности от 1. Чем статистичность больше, тем закономерность меньше.

В таблице сопряженности каждая клетка рассматривается как изображение прямой и обратной детерминации. ($x_i \rightarrow y_j$, $y_j \rightarrow x_i$). У каждой детерминации существуют две условные частоты – характеристики этих детерминаций – интенсивность и емкость [1].

Интенсивность детерминации $a \rightarrow b$, где $a = x_i$, $b = y_j$ вычисляется по формуле 2:

$$I(x_i \rightarrow y_j) = P(y \mid x) = \frac{N_{x_i y_j}}{N_{x_i}} \qquad (2)$$

Интенсивность детерминации отражает ее точность/истинность. Она показывает, что среди респондентов, обладающих признаком a, какая-либо доля демонстрируют поведение b [1].

Емкость вычисляется по формуле 3:

$$C(y_j \rightarrow x_i) = \frac{N_{x_i y_j}}{N_{y_j}} \quad (3)$$

Емкость показывает, сколько респондентов, или объектов, среди тех, кто демонстрирует тип поведения (признак) b, количество тех, кто является a, $a = x_i$. Иными словами, емкость – это доля реализации поведения b, которая объясняется высказыванием «Из a следует b». Полнота детерминации $a \rightarrow b$ – значение емкости, отражающее, насколько всеобъемлюще объяснение, построенное на детерминации $a \rightarrow b$.

3. Решение основного уравнения для D-функций.

Метод детерминационного анализа состоит в анализе детерминационных функций. Детерминационная функция образована детерминациями – логическими импликациями (следованиями), порожденными условными частотами. Правило «Если a, то b» – детерминация $a \rightarrow b$, то есть b предсказывается на основе a. Детерминация показывает, что одно событие оказывает влияние на другое событие [1].

Основная задача детерминационного анализа – в полном классе D-функций от x к y в контексте k найти все D-функции, которые удовлетворяют ограничениям (5):

$$\begin{cases} I(k(x \xrightarrow{\phi} y)) \geq \delta \\ C(k(x \xrightarrow{\phi} y)) \geq \sigma \end{cases} \quad (4)$$

где δ, σ – некоторые константы, устанавливающие минимальный порог интенсивности и емкости детерминации, причем $0 \leq \delta \leq 1$, $0 \leq \sigma \leq 1$ [1].

Поиск решений детерминационного уравнения осуществляется следующим образом. Для каждого непустого $x = a$ считаются I, C как функции от y, и те значения x, y, при которых названные величины превышают соответственно δ, σ, составляют детерминации-решения D-функций. Если $\delta > 0,5$, то при нахождении одного значения y (Проверяем его I, чтобы было больше данного δ) дальше можно не

искать, так как это решение единственное, поэтому переходим к следующему значению x.

4. Определение существенности различных переменных в детерминациях (D-функциях).

Уточнение, которое свойство с вносит в детерминацию $a \to b$ – наличие дополнительного детерминированного свойства c в детерминации $a \to b$. Оно обозначается как $ac \to b$.

Приращение интенсивности – мера существенности уточнения c, вносимого в детерминацию $a \to b$.

Его вычисляют по формуле 4:

$$S(ac^* \to b) = I(ac \to b) - I(a \to b) \quad (5)$$

5. Получение объяснения заданного свойства.
6. Получение уточнений для определенного решения.
7. Получение дополнений для совокупности решений задачи получения объяснения.
8. Расчет существенности контекста, объясняющих и объясняемых свойств.
9. Построение объясняющей и объясняемой типологии.

К инструментам детерминационного анализа также относится типология – классификация по существенным признакам для выявления общих закономерностей. Апостериорная типология формируется на основе изучения эмпирических данных. Первостепенным является определение наиболее важных стратегических переменных, которые будут составлять основу при создании типологии. Затем формируются группы значений X, приводящих к одному и тому же значению Y при заданных δ, σ, и на основе таких групп формулируется апостериорная типология [1].

10. Проверка объяснительных возможностей типологии.
11. Проверка объяснимости типологии.
12. Нахождение самых точных и полных детерминаций заданного и произвольного размера.

3. Сравнение известных статистических методов с методами детерминационного анализа

Детерминационный анализ удовлетворяет требованиям, вытекающим из принципов номинальности, конкретности и ограниченной статистичности. Проведение параллелей и разграничений между детерминационным анализом и иными математическими методами сводится к установлению того, каким из приведенных выше

требований тот или иной метод анализа данных не удовлетворяет. Сводные сведения представлены в таблице.

Таблица

Сравнение статистических методов с детерминационным анализом по выполнению основных принципов детерминационного анализа

Метод	**Принцип номинальности**	**Принцип конкретности**	**Принцип ограниченной статистичности**
Статистическая связь по определению класса вероятностно-статистических методов	+	+	–
Методы исследования связей на основе критерия χ^2	+	–	–
Расстояние по Хеммингу между разбиениями объектов	+	–	–
Уравнение регрессии	–	–	–
Метод главных компонент	–	–	–
Детерминационный анализ	+	+	+

4. Вычислительный эксперимент: задача исследования поведения студентов университета

С целью демонстрации использования разработанных программных средств в решении задач детерминационного анализа необходимо было провести вычислительный эксперимент на экспериментальных данных. Предметом исследования послужил набор данных, содержащий личную информацию и учебные привычки студентов инженерного и педагогического факультетов Ближневосточного Университета, расположенного на Северном Кипре [6]. Следует заметить, что, как правило, найти подходящие данные для проверки и демонстрации работы программных средств подобного рода не очень просто, чем и объясняется использование вышеупомянутых сведений, находящихся в открытом доступе. Кроме того, среди доступных наборов данных на сайте archive.ics.uci.edu используемый

набор содержал большое количество вопросов с ответами, квантификацию которых нельзя было логически обосновать. Набор с аналогичным количеством качественных данных по студентам какого-либо российского университета, к сожалению, обнаружить не удалось.

Целью исследования ставится выяснение влияния различных факторов на высокую успеваемость студента.

Вопрос 1. Кто более склонен к высокой успеваемости, мужчины или женщины?

Выбирается задача детерминационного анализа «Получение объяснений», в форму вводятся следующие данные: контекст универсальный, набор переменных X из единственной переменной – пола (код переменной 2), набор Y – переменная среднего балла за прошлый семестр со значением «3.49 из 4» (код переменной 29, код значения 5). Результат запроса представлен на рис. 2.

2	->	29	I	C	N(X&Y)	N(X)	N(Y)
2	->	5	0.218	0.760	19	87	25
1	->	5	0.103	0.240	6	58	25

Рис. 2. Полученное D-отношение для запроса получения объяснения высокому среднему баллу

Мужской пол (код значения 2) имеет большую емкость, следовательно, наиболее полно объясняет высокие оценки и является более склонным к высоким оценкам.

Вопрос 2. Остановимся на детерминации «Если это мужчина, то его средний балл будет выше 3.49/4». Можно ли объяснить их успех уровнем образования родителей?

Сначала рассматривается, какой уровень образования матери и отца вместе и по-отдельности вносит положительное уточнение в детерминацию «Если это мужчина, то его средний балл будет выше 3.49/4». Для этого выбирается пункт «Получение уточнения определенного решения», тип уточнения «Положительный», переменные Z ввести код 11 для материнского и код 12 для отцовского. Результаты представлены на рис. 3, 4 и 5.

11	I	ΔI
1	0.241	0.023
4	0.429	0.210

Рис. 3. Приращение интенсивности при уровне образования матери для мужчин

12	I	ΔI
2	0.300	0.082
3	0.267	0.048

Рис. 4. Приращение интенсивности при уровне образования отца для мужчин

11	12	I	ΔI
1	3	0.333	0.115
1	4	0.250	0.032
1	1	0.250	0.032
3	2	1.000	0.782
4	2	1.000	0.782
4	3	1.000	0.782
4	1	0.500	0.282

Рис. 5. Приращение интенсивности при учете уровня образования и матери, и отца для мужчин

Согласно рис. 5, образование отца если и вносит положительную существенность, то довольно незначительную, в то время как образование матери уровня бакалавра существенно с приращением 0.21. При учете образования обоих родителей получается даже получить случаи с максимальной точностью, то есть те, которые позволяют утверждать: «Если это мужчина и уровень образования матери – такой, а отца – такой, то у него будет высокий средний балл». Такими сочетаниями являются:

- уровень образования матери – бакалавриат (код 4), отца – средняя школа (код 2);
- уровень образования матери – бакалавриат (код 4), отца – старшая школа (код 3);
- уровень образования матери – старшая школа, отца – средняя школа.

Проверка численности этих трех групп с помощью задачи получения объяснения дает полноту 0,24, то есть точное объяснение почти четверти случаев высоких оценок.

Вопрос 3. Можно ли построить типологическое свойство на основе привычек в учебе мужчин?

Выбирается задача построения объясняющей типологии, в качестве контекста вводится переменная пола (код 2) и ее значение «мужской»

(код 2), в качестве Y переменную среднего балла (коды 1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: выше 3.49), в качестве X – набор из переменных 22, 25 и 26. Переменная 22 – частота посещения занятий (коды 1: всегда, 2: иногда, 3: никогда). Переменная 25 – ведение заметок в классе (коды 1: никогда, 2: иногда, 3: всегда). Переменная 26 – внимательное прослушивание лекции в классе (коды 1: никогда, 2: иногда, 3: всегда). Значение интенсивности 0,51, емкости – 0. Результат представлен на рис. 6.

22	25	26	->	29
2	3	1	->	3

22	25	26	->	29
1	1	2	->	4
2	3	3	->	4

22	25	26	->	29
1	2	3	->	5
1	1	3	->	5

Рис. 6. Таблицы для типологического свойства

Из полученных типологических групп можно делать следующие предположения.

Сочетания, кодирующие первую типологию – слабую группу – группу с оценками 2,50–2,99, хоть и старается все записывать, но при отсутствии вслушивания в материал это не дает им преимущества в оценках

Студенты, не дотягивающие до высоких оценок, представляют собой две группы: либо всегда посещают, но не интегрируются полностью в образовательный процесс (никогда не делают заметки, слушают иногда), либо не могут всегда посещать, но всегда полностью интегрируются в образовательный процесс (всегда и слушают, и делают заметки).

Студенты, получающие высокие оценки, всегда посещают, всегда слушают, но заметки ведут либо иногда, либо никогда.

Заключение

Программные средства на основе детерминационного анализа позволяют делать выводы о зависимости между отдельными значениями номинальных переменных путем анализа условных частот, составлять

на их основе детерминации и дифференцировать различные свойства по степени существенности их вклада в аргументы детерминации, измеряемой величинами соответствующих приращений условных частот. Реализованное приложение, обеспечивает поддержку решения основных задач детерминационного анализа и агрегированных задач на базе основных задач. В качестве возможных применений, помимо социологических исследований, можно предложить использование программных средств, например, в системе менеджмента качества или для решения различных управленческих задач.

Список литературы

1. Чесноков С. В. Детерминационный анализ / С. В. Чесноков ; под ред. Е. С. Райской. – Москва : Наука. Главная редакция физико-математической литературы, 1982. – 168 с.

2. Бородкин Ф. М. Математическое моделирование в социологии (методы и задачи) / Ф. М. Бородкин, Б. Г. Миркин. – Новосибирск : Наука, 1977. – 240 с.

3. Чесноков С. В. Взаимоотношение теоретического и эмпирического уровней описания социально-экономической реальности / С. В. Чесноков // Методология комплексного исследования социально-экономических систем. Труды ВНИИСИ ГКНТ и АН СССР. – Москва, 1980. – Вып. 1. – С. 33–42.

4. Дородницын А. А. Математика и описательные науки / А. А. Дородницын // Число и мысль. – Москва : Знание, 1977. – С. 6–15.

5. Чесноков С. В. Способ ручной обработки небольших массивов документов / С. В. Чесноков // Проблемы контент-анализа в социологии. Ротапринт ОУПЭС СО АН СССР. – Новосибирск : 1970, – С 101–137.

6. Higher education students performance evaluation dataset [Электронный ресурс] : база данных. – Режим доступа : https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset