

# Выявление причинно-следственных связей в данных психологического тестирования логическими методами<sup>1</sup>

**Аннотация.** Работа посвящена совместному использованию AQ-обучения и ДСМ-метода в задаче выявления причинно-следственных связей в данных психологического тестирования. AQ-обучение применяется для построения описания группы испытуемых с помощью правил. Описание группы служит основой построения базы фактов ДСМ-метода. Для выдвижения гипотез о наличии причинно-следственных связей в базе фактов используется первый этап ДСМ-метода. Описан полный алгоритм разработанного метода, предлагаются рекомендации по его применению. Проводится сравнительный анализ результатов, полученных с помощью предложенного метода, с результатами статистической обработки.

**Ключевые слова:** интеллектуальный анализ данных, причинно-следственные связи, ДСМ-метод, AQ-обучение, статистический анализ данных, психологическое тестирование.

## Введение

В практике психологического исследования часто возникает задача объяснения проявления тех или иных характеристик группы испытуемых. Например, требуется объяснить наличие у респондентов с высоким уровнем враждебности по шкале Басса-Перри убежденности в возможности контролировать неприятные ситуации в их жизни. Одним из этапов интерпретации того или иного факта является выявление других фактов, которые служат причиной его проявления. В данном случае под фактом понимается значение некоторого признака или шкалы опросника, с помощью которого описывается группа респондентов. Таким образом, возникает задача описания направленных каузальных связей между значениями признаков или, в более простой и частой постановке, между самими признаками. Следует отметить, что корреляция между признаками хоть и свидетельствует о наличии прямой зависимости между описываемыми ими характеристиками испытуемых, т.е. о соответствии изменчивости одной характеристики изменчивости другой, но

ничего не говорит о том, что из них является причиной, а что следствием. То же самое можно сказать о факторном анализе. Выявляемые в его ходе линейные комбинации признаков являются лишь сборными факторами, по которым хорошо разделяются объекты исследования, что также не может свидетельствовать о характере причинно-следственных связей.

Проблема интерпретации результатов статистического анализа данных связана именно с тем, что их можно использовать лишь для подтверждения с той или иной степенью достоверности уже выдвинутых гипотез о каузальной зависимости. Классическим примером несостоятельности интерпретации корреляций в качестве каузальных зависимостей служит случай трех признаков, предположим,  $a$ ,  $b$  и  $c$ , для которых определены причинно-следственные связи  $a \rightarrow b$  ( $a$  детерминирует проявление  $b$ ) и  $a \rightarrow c$  ( $a$  детерминирует проявление  $c$ ). При уменьшении или увеличении проявления признака « $a$ » будет в меньшей или большей степени соответственно проявляться и признаки « $b$ » и « $c$ », т.е. « $b$ » и « $c$ » будут изменяться согласованно. Между ними будет определен высокий уровень

<sup>1</sup> Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты №12-07-00611 и № 12-07-31057) и Российского гуманитарного научного фонда (грант №10-06-00834).

корреляции, что на самом деле не отражает реально имеющихся каузальных связей.

Отдельно следует выделить случай сложных объектов исследования, с которыми в основном и имеет дело психолог. Для сложных объектов характерны и сложные причинно-следственные связи, т.е. случаи, когда некоторый признак проявляется только при совместном проявлении целого ряда других признаков. Каждый из них не предопределяет проявление интересующей нас характеристики, поэтому здесь нужно говорить и о сложной причине, состоящей из совокупности нескольких признаков, вполне возможно, независимых. Выделить такие сложные связи на основе только ненаправленных корреляций обычно очень сложно.

Наряду со статистическим анализом данных в методах машинного обучения активно применяется интеллектуальный анализ данных. Под ним будет пониматься такой метод получения новой информации, в котором анализируются логические связи между структурными компонентами данных. С этой точки зрения наибольший интерес представляют индуктивные методы обучения и правдоподобные рассуждения, в частности ДСМ - метод [16], который непосредственно предназначен для индуктивного порождения гипотез о наличии причинно-следственных связей.

ДСМ-метод в основном применяется для анализа данных из предметных областей, где существуют только строго детерминистические отношения между объектами [17]. Однако существуют расширения ДСМ-метода с учетом статистических соображений [6], что позволяет применять его в тех областях, где законы взаимодействия объектов точно не известны. Имеется достаточно много работ по применению ДСМ-метода в слабоструктурированных областях знаний, таких как социология [8]. Он был успешно применен в задачах анализа текстов на естественном языке [3, 14, 15] и в задаче поиска информации [15].

Интеллектуальные методы анализа имеют то преимущество, что опираются на интуитивно понятные и очевидные исследователю аксиомы и правила вывода. Индуктивные рассуждения, являющиеся переходом от знания меньшей степени общности к знанию большей степени общности, т.е. от единичных фактов к обобщениям, не являются, вообще говоря, достоверными и относятся к правдоподобным заключе-

ниям. Однако недостоверность индуктивной гипотезы носит иной смысл, чем недостоверность статистической гипотезы. В первом случае вывод является достоверным для имеющихся данных и теряет свою достоверность только при переходе на выборку большего объема. Во втором же случае гипотеза может не подтверждаться даже для некоторых объектов первоначальных данных, не говоря уже о большей выборке.

Работа посвящена исследованию и применению интеллектуальных методов выявления причинно-следственных связей при анализе данных психологического тестирования по методике, разрабатываемой автором и основные принципы которой описаны в работах [11, 12]. В предлагаемой работе особое внимание уделено условиям применимости предлагаемого метода и сравнению получаемых результатов с результатами статистического анализа. При этом использованы данные тестирования 70 испытуемых по шкале базисных установок (ШБУ), по методике субъективной оценки межличностных отношений (СОМО), по шкале агрессивности Басса-Перри (БП), по опроснику легитимации агрессии (ЛА) и общей шкале конструктивного мышления (ОКМ).

## 1. Метод

### 1.1. Требования к данным

Задача выявления причинно-следственных связей, характерных для группы испытуемых, предполагает наличие множества объектов, разбитых на классы. В психологических исследованиях разбиение на классы получается следующими способами, в зависимости от условий эксперимента:

1. Перед психологом стоит задача исследования группы испытуемых, для которых заранее известен их отличительный, так называемый классифицирующий, признак. Например, проводится исследование респондентов, пострадавших в результате аварии на АЭС, или работников конкретной профессии, такой как журналистика. Для получения разбиения испытуемых на группы в этой ситуации проводится тестирование дополнительной выборки, в которой испытуемые не обладают классифицирующим признаком или этот признак имеет другое значение. Примерами дополнительной выборки могут

служить респонденты другой профессии или респонденты, не находившиеся в зоне аварии.

2. Исследователь работает с не пополняемой выборкой, например с данными тестирования работников конкретного предприятия, и стоит задача объяснения характеристик выделяемых в этой выборке групп. В этом случае разбиение получается либо проведением кластерного анализа, если классифицирующий признак заранее не известен, либо прямым выбором интересующего классифицирующего признака, если заранее известно на какие группы разбивается выборка.

Следует отметить ограничения на параметры данных, которые накладываются при использовании предлагаемого в настоящей работе индуктивного метода выявления причинно-следственных связей. Главное требование к данным – это их полнота, означающая, что среди объектов выборки не должны присутствовать такие объекты, у которых есть неопределенные или недостоверные значения признаков. Если такие объекты присутствуют, необходимо удалить из выборки либо сами эти объекты, либо признаки, надежные значения которых отсутствуют у этих объектов. На индуктивные методы не оказывает влияния закон вероятностного распределения значений признаков в выборке. В то же время он важен для параметрических критериев проверки статистических гипотез (например, для критерия корреляции Пирса), для дискриминантного и факторного анализов. Выборки, соответствующие разным группам испытуемых, для индуктивных методов не обязательно должны быть получены из некоторой генеральной совокупности, поэтому предварительный подтверждающий дисперсионный анализ обычно не проводится. Еще одно частое требование для многих типов статистического анализа – нормализация значений признаков, также является необязательным для логического вывода. Признаки могут быть измерены с применением любого типа шкалы, как номинативного, так и интервального.

Тестовые данные, используемые в данной работе, разбиваются на три класса по уровню значения шкалы «Враждебность» (БП). К классу 1 относятся испытуемые с низким уровнем враждебности, к классу 2 – со средним уровнем, к классу 3 – с высоким.

После того как проведена проверка на то, что данные удовлетворяют всем требованиям,

каждый класс объектов получает описание в виде множества правил соотнесения объектов к данному классу. Правила строятся с использованием метода АQ-обучения [2, 4].

## 1.2. Построение описания группы испытуемых

Пусть имеются три множества:  $O=\{o_{ij}\}$  – множество из  $n$  объектов, которые представляют данные тестирования испытуемых;  $C=\{c_k\}$  – множество классов объектов, соответствующие группам, на которые были разбиты испытуемые;  $P=\{p_j\}$  – множество из  $m$  признаков, которые обычно соответствуют определенным шкалам методик и опросников. Составляется матрица значений признаков  $A_{ij}=\{a_{ij}\}$ , в которой каждому признаку соответствует столбец его значений из матрицы  $A_{ij}$   $p_j \rightarrow (a_{1j}, a_{2j}, \dots, a_{nj})$ , а каждому объекту соответствует его описание  $o_i \rightarrow (p_1=a_{i1}, p_2=a_{i2}, \dots, p_m=a_{im})$ , где пара  $p_j=a_{ij}$  называется свойством объекта.

Множество значений каждого интервального признака  $p_j$  делится на части, количество которых обычно равно трем:  $w_1$  – высокое значение признака,  $w_2$  – среднее значение и  $w_3$  – низкое значение –  $(a_{1j}, a_{2j}, \dots, a_{nj})=w_1 \cup w_2 \cup w_3$ . В качестве алгоритма разбиения интервальной шкалы используется метод  $\chi$ -слияния [1]. На подготовительном этапе алгоритма все объекты сортируются по величине интервального признака, а затем запускается итеративный процесс попарного слияния тех соседних интервалов, для которых оказалось минимальным значение  $\chi^2$ . Слияние продолжается до тех пор, пока  $\chi^2$  не превысит некоторый порог, обычно равный 0,9. Значение  $\chi^2$  вычисляется по формуле:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

где  $m$  – количество сравниваемых интервалов,  $k$  – количество классов (в нашем случае  $m=2, k=3$ ),  $A_{ij}$  – количество примеров класса  $j$  в  $i$ -ом интервале,

$$E_{ij} = \frac{\sum_{j=1}^k A_{ij} \sum_{i=1}^m A_{ij}}{\sum_{j=1}^k \sum_{i=1}^m A_{ij}} - \text{ожидаемая частота } A_{ij}.$$

Затем начинается построение так называемых АQ-правил, которые представляют собой конъюнкцию свойств объектов. Свойством  $h_j$

будем называть дизъюнкцию интервалов значений признака:  $p_j = \bigcup_q w_q$ .

Кратко опишем алгоритм построения AQ-правил, следуя [10]. На первом шаге AQ-обучения все объекты анализируемого класса  $c_k$  объявляются положительными примерами, а объекты других классов – отрицательными. Затем произвольным образом выбирается опорный пример, для которого строится начальное правило из свойств объекта:  $R_k^1 = \{p_1 = w_{i_1}, p_2 = w_{i_2}, \dots, p_m = w_{i_m}\}$ . Далее правило расширяется путем добавления интервалов к какому-либо признаку, либо удалением из правила одного из признаков таким образом, чтобы результирующее правило не покрывало ни одного отрицательного примера. В итоге получается дерево правил, листьями которого являются наиболее общие правила, описывающие определенное количество положительных примеров. Из наиболее общих правил выбирается наилучшее, т.е. то, которое покрывает наибольшее количество положительных примеров и обладает минимальной длиной. Мерой длины правила  $R$  является его сложность  $Sx(R)$ , вычисляемая как сумма весов составляющих правила. Веса составляющих, в соответствии с [4], представлены в Табл. 1.

Табл. 1. Веса операторов, встречающихся в AQ-правилах

| Оператор                   | Вес |
|----------------------------|-----|
| Конъюнкция свойств         | 4   |
| Дизъюнкция свойств         | 10  |
| Дизъюнкция интервалов      | 2   |
| Интервал                   | 2   |
| Оператор больше или меньше | 1   |
| Оператор равенства         | 1   |
| Оператор «не равно»        | 2   |

В результате работы AQ метода, каждому классу  $c_k$  ставится в соответствие набор правил, представляющих собой конъюнкцию свойств:

$$\mathfrak{R}_k = \{R_{ki} \mid R_{ki} = \bigcap_j (p_j = \bigcup_q w_q)\}.$$

Так как начальный объект, с которого начинается построение правил, может быть выбран произвольным образом и, в зависимости от этого, результаты будут различаться, то процедура запуска AQ-метода осуществляется несколько раз. Процесс расширения объектных свойств до классовых, зависящий от порядка встречающихся объектов, похож на процесс образования естественных понятий в модели Брунера [13].

Из всех получаемых наборов правил  $\mathfrak{R}_k$  для класса  $c_k$  выбирается минимальный по сложности набор:

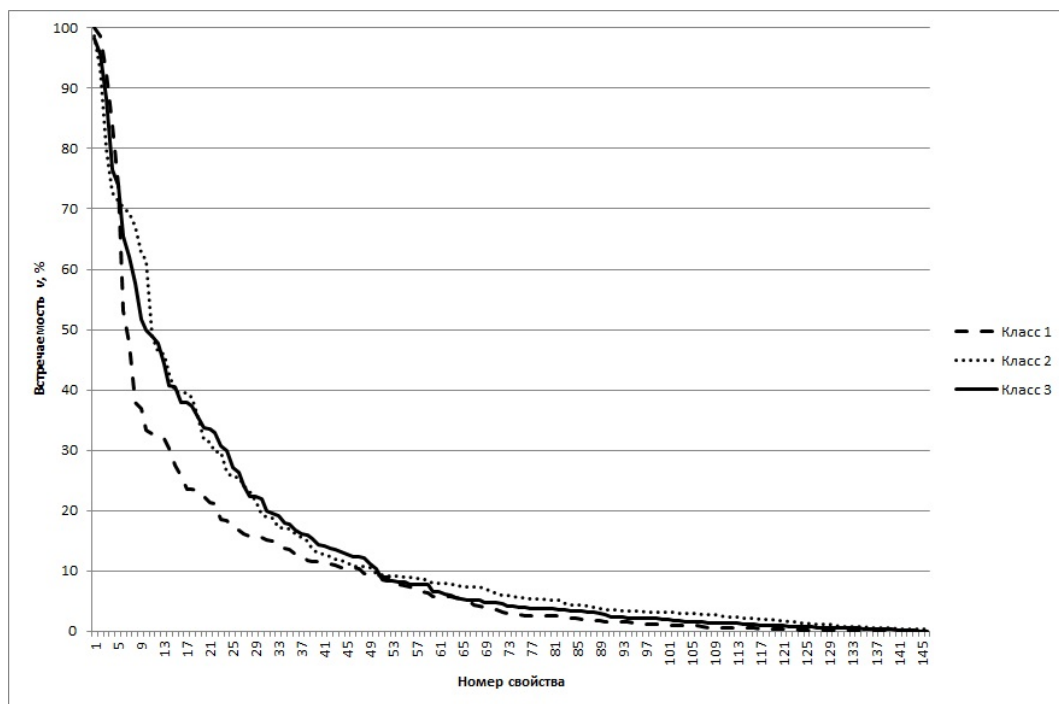


Рис. 1. Распределение свойств классов по встречаемости

$$\mathfrak{R}_k^* = \min_{R_{ik} \in \mathfrak{R}_k} \mathfrak{R}_k.$$

На основе построенного набора правил  $\mathfrak{R}_k^*$  производится анализ структуры класса. Объекты класса, которые описываются одним правилом, образуют подкласс. Подклассы образуют структуру исследуемого класса объектов. Если в структуре класса выделяется подкласс, состоящий из одного объекта, то этому объекту необходимо уделить особое внимание (возможно, при тестировании соответствующего испытуемого были допущены ошибки, либо испытуемый выпадает из общей выборки).

В процессе многократных запусков процедуры построения AQ-правил производится подсчет встречаемости  $v(h_j)$  свойства  $h_j$  в правилах для данного класса. На Рис.1 представлены характерные графики распределения свойств классов с различным уровнем враждебности для тестовых данных.

Описание класса  $\mathfrak{D}_k$  составляют значимые свойства, т.е. те свойства, встречаемость которых выше  $1/3$ :  $\mathfrak{D}_k = \{h_j \mid v(h_j) > 1/3\}$ .

Для определения того, насколько описание класса пригодно для выявления причинно-следственных связей предложены два критерия.

#### 1. Критерий сложности класса

$$\theta_1(c_k) = \frac{|\mathfrak{D}_k|}{5m},$$

где  $|\mathfrak{D}_k|$  означает мощность описания класса  $c_k$ ,

$5m = m \sum_2^q i$  – максимальное количество возможных свойств в том случае, когда шкала разбивается на  $q$  (в нашем случае на три интервала). Минимальное значение критерия равно  $1/5m$  в том случае, когда все объекты класса описываются одним отличительным свойством. Максимальное значение равно 1 в том случае, когда для описания класса задействованы все возможные свойства. Чем меньше значение критерия  $\theta_1$ , тем ниже размерность пространства поиска причинно-следственных связей, поэтому оно не должно быть близко ни к 0, ни к 1. Оптимальное эмпирическое значение равно 0,1. Критический порог с большей стороны зависит от производительных мощностей вычислительной машины, но редко превышает значения  $5/m$ .

#### 2. Критерий однородности класса

$$\theta_2(c_k) = \frac{\sum_{\mathfrak{R}_k^*} Cx(R_{ik})}{\sum_{\mathfrak{R}_k^*} Cov(R_{ik})},$$

где  $Cov(R_{ik})$  означает количество положительных примеров, описываемых правилом  $R_{ik}$ . Критерий принимает минимальное значение, равное  $1/N$ , где  $N$  – количество объектов в классе, в том случае, когда класс описывается одним правилом, покрывающим все примеры, и максимальное значение, равное 1, в том случае, когда каждый объект класса описывается своим правилом. Критерий  $\theta_2$  определяет, насколько объекты внутри класса похожи друг на друга: чем меньше его значение, тем более объекты сходны между собой. Эмпирическое критическое значение равно 0,5.

В качестве примера описаний классов и значений критериев, приведем таблицу свойств классов, образованных из тестовых данных (Табл. 2).

Уровень «низкий» соответствует первому интервалу значений признака, «средний» – второму, «высокий» – третьему, «ниже среднего» – объединению первого и второго интервалов, уровень «выше среднего» – объединению второго и третьего интервалов.

Из Табл. 2 видно, что множества свойств классов могут пересекаться друг с другом, т.е. одни и те же свойства являются важными как для одного, так и для другого класса. В описание одного класса могут входить свойства, соответствующие разным интервалам значений одного признака (например, «Phis. Agr. – ниже среднего» и «Phis. Agr. – высокий» для класса 2, такие свойства в Табл.2 подчеркнуты). Наличие большого количества таких свойств служит свидетельством в пользу высокой сложности данного класса в силу критерия  $\theta_1$ .

#### 1.3. Построение каузальных гипотез

В том случае, когда значения критериев  $\theta_1$  и  $\theta_2$  свидетельствуют о возможности выявления причинно-следственных связей, по описанию класса объектов составляется база фактов ДСМ-метода. Предварительно в целях сокращения пространства поиска для каждого класса проводится сокращение множества свойств описания двумя способами:

Табл. 2. Описания классов с различным уровнем враждебности

| Класс 1 (низкая враждебность)<br>$N=25$ ,<br>$\theta_1=0,094$ , $\theta_2=0,12$                                                                                                                                                                                                            | Класс 2 (средняя враждебность)<br>$N=20$ ,<br>$\theta_1=0,125$ , $\theta_2=0,24$                                                                                                                                                                                                                                                                                                 | Класс 3 (высокая враждебность)<br>$N=25$ ,<br>$\theta_1=0,125$ , $\theta_2=0,17$                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Конфл. – ниже среднего<br>Агресс. – низкий<br>Anger – ниже среднего<br>ШЭС – выше среднего<br>ШЛСМ – ниже среднего<br>Контрол – ниже среднего<br>СКС – высокий<br>Отчужд. – низкий<br>Политика – выше среднего<br>ОШКМ – выше среднего<br>ШВ – ниже сред./выше сред.<br>ШЛ – выше среднего | Агресс. – ниже среднего<br>Anger – высокий<br>ШЭС – выше сред./низ.<br>ШЛСМ – ниже сред./выс.<br>Контрол – ниже среднего<br>СКС – выше среднего<br>Отчужд. – выше среднего<br>Политика – ниже среднего<br>ОШКМ – выше среднего<br>ШВ – выше среднего<br><br>Phis. Agr. – ниже сред./выс.<br>ШЭМ – ниже среднего<br>ШПС – ниже среднего<br>ШКМ – выше среднего<br>Удача – средний | ШЭС – выше среднего<br><br>Контрол – выше среднего<br>СКС – ниже среднего<br><br>Политика – выше среднего<br>ОШКМ – выше сред./низ.<br><br>ШЛ – выше среднего<br>Phis. Agr. – выше среднего<br><br>ШПС – выше среднего<br>ШКМ – выше среднего<br>Удача – ниже среднего<br>Напряж. – выше среднего<br>Добр. – ниже сред./выше сред.<br>Законом. – выше среднего<br>ШНО – выше среднего<br>Личный опыт – выше среднего<br>Воспитание – средний<br>Спорт – ниже среднего |

1) если два свойства  $h_1 : (p_1 = \bigcup_i w_i)$  и  $h_2 : (p_2 = \bigcup_j w_j)$  содержат один и тот же признак, т.е.  $p_1 = p_2$ , и разные, но не вложенные, интервалы для этого признака, т.е.  $\bigcup_i w_i \neq \bigcup_j w_j$ ,

$\bigcup_i w_i \not\subset \bigcup_j w_j$  и  $\bigcup_j w_j \not\subset \bigcup_i w_i$ , то такие свойства считаются конфликтными и оба исключаются из базы фактов;

2) если два свойства  $h_1 : (p_1 = \bigcup_i w_i)$  и  $h_2 : (p_2 = \bigcup_j w_j)$  содержат один и тот же признак, т.е.  $p_1 = p_2$ , и разные вложенные интервалы для этого признака, т.е.  $\bigcup_i w_i \neq \bigcup_j w_j$  и либо

$\bigcup_i w_i \subset \bigcup_j w_j$ , либо  $\bigcup_j w_j \subset \bigcup_i w_i$ , то такие свойства считаются вложенными и исключается более общее свойство.

Из всех свойств описания класса поочередно выбирается целевое, для которого на основе оставшихся свойств строятся гипотезы о причинно-следственных связях. Объектами базы фактов являются как объекты исследуемого класса, так и других классов.

Причинно-следственные связи выявляются в ДСМ-методе на основе следующих допущений [5]:

- *гипотеза схождения*: если все случаи наблюдаемого явления имеют только один общий фактор, то он и является причиной рассматриваемого явления;

- *гипотеза различия*: если оба случая сходны по всем факторам, кроме одного, и этот фактор присутствует в случае, когда явление наступает, то он является причиной рассматриваемого явления;

- *гипотеза абдукции*: если множество фактов объясняет множество выдвинутых гипотез, то эти гипотезы правдоподобны.

В качестве алгоритмической реализации метода автоматического порождения гипотез использовались алгоритмы Аншакова и Объедкова [7, 9].

Выдвигаемые в результате ДСМ-метода гипотезы о наличии причины целевого свойства  $h_g : (p_g = \bigcup_q w_q)$  для класса  $c_k$  имеют вид конъюнкции свойств:

$$H(h_g, c_k) = \bigcap_{j \neq g} (p_j = \bigcup_q w_q).$$

Если причина состоит из одного свойства, она называется простой, иначе – комплексной. Гипотеза ДСМ-метода, в отличие от статисти-

ческой, говорит о детерминации свойств, а не о зависимости признаков (шкал). Свойство представляет собой конкретное подмножество значений признака (высокое, низкое значение), поэтому связь признаков не подразумевает связь свойств и наоборот.

Множество гипотез, построенных для свойства  $h_g$ , редуцируется по длине и по вложенности. Если комплексная причина содержит более трех свойств, она считается незначимой. Причины, которые являются расширением других причин за счет конъюнкции свойств, также исключаются из множества гипотез.

#### 1.4. Общая схема алгоритма

Приведем общий алгоритм, лежащий в основе предлагаемого метода выявления причинно-следственных связей.

##### Алгоритм AQ+JSM

*Вход:* множество объектов  $O=\{o_{ij}\}$ , разбитых на множество классов  $C=\{c_k\}$ , множество признаков  $P=\{p_j\}$ , матрица значений  $A_{ij}=\{a_{ij}\}$ .

*Шаг 1:* из множества  $C$  выбирается исследуемый класс  $c_k$  и из него – начальный объект для AQ-метода.

*Шаг 2:* с помощью AQ-обучения для класса  $c_k$  строится набор правил  $\mathfrak{R}_k$ , начиная с начального объекта.

*Шаг 3:* для каждого свойства  $h_j$  из набора правил  $\mathfrak{R}_k$  подсчитывается встречаемость (с накоплением); выбирается новый начальный объект, если таковой имеется, и переходим к шагу 2, иначе выбирается новый класс  $c_k$  для исследования и также переходим к шагу 2; если все классы были рассмотрены переходим к шагу 4.

*Шаг 4:* для каждого класса составляется его описание  $\mathfrak{D}_k$  и выбирается наилучшее правило  $\mathfrak{R}_k^*$ , вычисляются критерии  $\theta_1(c_k)$  и  $\theta_2(c_k)$ ; если значения какого-либо из них выше (ниже) критических порогов, то алгоритм заканчивает работу с пустым множеством причин для каждого класса, иначе переходим к шагу 5.

*Шаг 5:* для каждого класса  $c_k$  на основе  $\mathfrak{D}_k$  формируется база фактов, в которой проводится сокращение множества свойств по вложенности и конфликтности.

*Шаг 6:* выбирается исследуемый класс  $c_k$  и с помощью первого шага ДСМ-метода для него строятся гипотезы о наличии причинно-

следственных связей между свойствами класса  $\mathfrak{H}_k=\{H(h_g, c_k)\}$ .

*Шаг 7:* множество гипотез  $\mathfrak{H}_k$  сокращается путем исключения причин, длина которых больше критической, и причин, которые включены в другие причины, если остаются неисследованные классы, выбирается один из них и переходим к шагу 6.

*Выход:*  $\{c_k, \mathfrak{H}_k\}$  – множество пар «класс – множество гипотез».

## 2. Эксперимент

Разрабатываемый метод выявления причинно-следственных связей был реализован в виде программной системы анализа данных психологического тестирования. Проведенные эксперименты с данными, описанными в первой части статьи, позволили построить и проинтерпретировать около десяти каузальных связей. Далее представлен сравнительный анализ полученных результатов с результатами статистической обработки тех же данных.

С помощью алгоритма AQ+JSM была построена следующая гипотеза:

$H(\text{«Конфликтность – ниже среднего», «Класс 1»}) = (\text{«Агрессивность – ниже среднего»}) \wedge (\text{«ШВ – выше среднего»})$ .

Данная запись интерпретируется так, что у испытуемых с низкой враждебностью причиной того, что их конфликтность ниже среднего, является одновременно низкая агрессивность и уровень ШВ выше среднего.

Корреляционный анализ для той же группы испытуемых выделил достаточно много связей признака «Конфликтность» с другими признаками. В Табл. 3 показаны коэффициенты корреляции с уровнем значимости  $p < 0,05$ .

В построенной прямой пошаговым методом модели множественной регрессии ( $MultipleR=0,91$ ;  $F=34$ ) содержится три предиктора с высоко значимыми коэффициентами: «Пол», «Напряженность» и «Воспитание».

Таким образом, в данном случае результаты, выделенные статистическими и индуктивными методами, оказались не пересекающимися, что свидетельствует о разных принципах, лежащих в их основе.

Для свойства «ШЭС – выше среднего» AQ+JSM построил четыре гипотезы:

Табл. 3. Корреляции признака «Конфликтность» ( $p < 0,05$ )

| Гр. по возр. | Пол     | Напряж. | Отчужд.   | Агресс. | Phis. Agr. | Anger | Политика |
|--------------|---------|---------|-----------|---------|------------|-------|----------|
| -0,499       | -0,495  | 0,823   | 0,658     | 0,794   | 0,597      | 0,522 | 0,451    |
| Личный опыт  | Воспит. | Спорт   | Интеграц. |         |            |       |          |
| 0,523        | 0,454   | 0,479   | 0,475     |         |            |       |          |

1.  $H(\text{«ШЭС – выше среднего»}, \text{«Класс 1»}) = (\text{«СКС – высокий»}) \wedge (\text{«ОШКМ – выше среднего»}) \wedge (\text{«ШВ – выше среднего»})$ ;

2.  $H(\text{«ШЭС – выше среднего»}, \text{«Класс 1»}) = (\text{«СКС – высокий»}) \wedge (\text{«ШЛСМ – ниже среднего»}) \wedge (\text{«ШВ – выше среднего»})$ ;

3.  $H(\text{«ШЭС – выше среднего»}, \text{«Класс 1»}) = (\text{«СКС – высокий»}) \wedge (\text{«Политика – выше среднего»}) \wedge (\text{«ШВ – выше среднего»})$ ;

4.  $H(\text{«ШЭС – выше среднего»}, \text{«Класс 1»}) = (\text{«СКС – высокий»}) \wedge (\text{«Отчужд. – низкий»}) \wedge (\text{«ШВ – выше среднего»})$ .

Из анализа гипотез можно сделать вывод, что комплексное свойство («СКС – высокий»)  $\wedge$  («ШВ – выше среднего») входит во все гипотезы, что свидетельствует о его большом влиянии на рассматриваемое свойство, но не служит его причиной.

Значимые корреляционные связи между признаком «ШЭС» и другими признаками приведены в Табл. 4.

Пошаговый метод построения модели множественной регрессии ( $MultipleR=0,95$ ;  $F=43$ ) указал на четыре предиктора: ОШКМ, ШЭМ, ШПС и ШЛ.

В данном случае и индуктивный и статистический методы указали на большую роль признака ОШКМ в определении ШЭС. В ДСМ-гипотезах участвуют также такие признаки как СКС, ШЛМС, в то время как ни регрессионный, ни корреляционный анализы на них не указывают.

Для свойства «ШВ – выше среднего» ДСМ-методом было построено четыре основных гипотезы о наличии простых причин:

1.  $H(\text{«ШВ – выше среднего»}, \text{«Класс 1»}) = (\text{«ШЭС – выше среднего»})$ ,

2.  $H(\text{«ШВ – выше среднего»}, \text{«Класс 1»}) = (\text{«ОШКМ – выше среднего»})$ ,

3.  $H(\text{«ШВ – выше среднего»}, \text{«Класс 1»}) = (\text{«СКС – высокий»})$ ,

4.  $H(\text{«ШВ – выше среднего»}, \text{«Класс 1»}) = (\text{«Отчужд. – низкий»})$ .

Табл. 4. Корреляции признака «ШЭС» ( $p < 0,05$ )

| СКС    | Воспитание | СМИ    | ОШКМ  | ШЭМ    | ШВ    |
|--------|------------|--------|-------|--------|-------|
| -0,405 | -0,518     | -0,598 | 0,806 | -0,426 | 0,507 |

Табл. 5. Корреляции признака «Конфликтность» ( $p < 0,05$ )

| Агресс. | СМИ    | ОШКМ  | ШЭС   | ШКМ    |
|---------|--------|-------|-------|--------|
| 0,407   | -0,530 | 0,472 | 0,507 | -0,409 |

Все четыре причины являются независимыми, т.е. подтверждение любой гипотезы означает проявление целевого свойства.

Значимые корреляционные связи между признаками приведены в Табл. 5.

Результат пошагового метода построения модели множественной регрессии с низкой значимостью ( $MultipleR=0,66$ ;  $F=8,4$ ) указал на два предиктора: «Пол» и «СМИ».

В данном случае корреляционный и индуктивный методы указали на два общих признака: ШЭС и ОШКМ. Причем симметрия в корреляциях между ШЭС и ШВ очевидна, а в ДСМ-методе взаимная обусловленность не является обязательным результатом.

Таким образом, предлагаемый логический метод выявления причинно-следственных связей позволил выдвинуть интерпретируемые гипотезы о наличии каузальных связей между свойствами класса объектов, представляющих данные психологического тестирования. Важно отметить, что такие гипотезы невозможно построить на основе только статистических результатов.

## Заключение

В работе представлен метод выявления причинно-следственных связей в данных психологического тестирования, основными составляющими которого являются построение описания классов на основе правил AQ-обучения, и анализ гипотез о причинно-следственных связях, которые строятся на первом этапе ДСМ-метода. Про-

веденное сравнение результатов, получаемых на основе разработанного индуктивного метода и статистического подходов, показало, что данный метод представляет собой альтернативный статистическим пакетам инструмент работы с экспериментальными данными. В ряде случаев получаемые гипотезы, указывающие на зависимости между свойствами, невозможно построить на основе статистических результатов. Предлагаемый метод может быть рекомендован для анализа слабоструктурированных данных в различных областях знаний.

## Литература

1. Kerber R. ChiMerge: Discretization of Numeric Attributes // In Proc. AAAI-92, Ninth National Conference Artificial Intelligence. AAAI Press/The MIT Press. 1992. P. 123-128.
2. Michalski R.S. AQVAL/1-Computer Implementation of Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition // Proc. Of the First Int. Joint Conf. on Pattern Recognition. Washington, DS, 1973. P. 3-17.
3. Smirnov I. V. Producing Rules for Ascertaining Meanings of Minimal Semantic-Syntactic Text Units. //Automatic Documentation and Mathematical Linguistics. Vol. 41, No. 3, Allerton Press, 2007. P. 108-113.
4. The aq21 natural induction program for pattern discovery: Initial version and its novel features / Janusz Wojtusiak, Ryszard S. Michalski, Kenneth A. Kaufman, Jaroslaw Pietrzykowski // ICTAI. 2006. P. 523-526.
5. Аншаков О.М. Об одной интерпретации ДСМ-метода автоматического порождения гипотез // Научно-техническая информация, Сер. 2. Информационные процессы и системы. 1999. №1. С. 45-53.
6. Григорьев П.А. Sword-системы или ДСМ-системы для цепочек, использующих статистические соображения // Научно-техническая информация, Сер. 2. Информационные процессы и системы. 1996. №5-6. С. 45-51.
7. Забежайло М.И., Ивашко В.Г., Кузнецов С.О., Михеевкова М.А., Хазановский К.П., Аншаков О.М. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез // Научно-техническая информация, Сер. 2. Информационные процессы и системы. 1987. №10. С. 1-14.
8. Климова С.Г., Михеевкова М.А., Панкратов Д.В. ДСМ-метод как метод выявления детерминант социального поведения // Научно-техническая информация, Сер. 2. Информационные процессы и системы. 1999. №12. С.3-14.
9. Объедков С.А. Алгоритмические аспекты ДСМ-метода автоматического порождения гипотез. / НТИ, Серия 2. Вып. 1-2, 1999. С. 64-74.
10. Осипов Г.С. Методы искусственного интеллекта. М.: ФИЗМАТЛИТ, 2011. 296с.
11. Панов А.И. Методика интеллектуального анализа результатов психологического тестирования // Труды I Всероссийской научной конференции молодых ученых. 2010. Т. I. С. 39-45.
12. Панов А.И. Применение методов искусственного интеллекта в обработке психологических данных // Экспериментальная психология в России: традиции и перспективы / Под ред. В.А. Барабанщикова. М.: Изд-во «Институт психологии РАН», 2010. С. 153-159.
13. Панов А.И., Чудова Н.В. Моделирование процесса образования естественных понятий методами искусственного интеллекта // Четвертая международная конференция по когнитивной науке: Тезисы докладов: В 2 т. Томск, 22-26 июня 2010г. 2010. Т. 2. С. 455.
14. Смирнов И.В. Метод автоматического установления значений минимальных синтаксических единиц текста // Информационные технологии и вычислительные системы. 2008. №3. С. 30-45.
15. Тихомиров И.А., Смирнов И.В. Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине Exactus. //Труды международной конференции Диалог'2009. С. 483-487.
16. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. 1991. Т. 15. С. 54-101.
17. Финн В.К. Каузальный анализ данных в интеллектуальных системах // Научно-техническая информация, Сер. 2. Информационные процессы и системы. 2000. №11. С. 1-5.

**Панов Александр Игоревич.** Инженер-исследователь ИСА РАН. Окончил Новосибирский государственный университет в 2009 году и Московский физико-технический институт (Государственный университет) в 2011 году. Автор 11 печатных работ. Область научных интересов: методы машинного обучения, когнитивное компьютерное моделирование. E-mail: pan@isa.ru.