МЕТОД ИЗВЛЕЧЕНИЯ ПРИЧИННО-СЛЕДСТВЕННЫХ СВЯЗЕЙ С ИСПОЛЬЗОВАНИЕМ ОПТМИЗИРОВАННЫХ БАЗ ФАКТОВ

Панов А.И., Швец А.В., Волкова Г.Д.
ИСА РАН, 117312, г. Москва, пр. 60-летия Октября, 9, pan@isa.ru

Аннотация. В работе предлагается и исследуется метод извлечения причинно-следственных (каузальных) бинарных отношений из множества баз фактов. Базы фактов строятся для целевых свойств каждого класса объектов. Описание классов формируются в результате обучения на данных из слабо формализованной предметной области. Обучение проводится с использованием коэволюционного генетического алгоритма, сокращающего начальное пространство признаков. По сформированным оптимизированным описаниям классов с помощью первого этапа ДСМ-метода осуществляется поиск причинно-следственных отношений для всех целевых свойств. Предложенный метод подходит как для анализа небольшого количества полных данных, так и для работы с массивами неполных данных большого размера. Проведен ряд модельных экспериментов с использованием базы медицинских данных MIMIC II.

Ключевые слова: машинное обучение, генетический алгоритм, причинно-следственные отношения, ДСМ-метод, AQ-обучение

Введение

Извлечение причинно-следственных (каузальных) отношений в массиве данных, особенно в случае неполных данных большого объема, имеет важное прикладное значение. Это связано с тем, что каузальное отношение является одним из небольшого количества базовых отношений, с помощью которых строится представление знаний в любых предметных областях, в том числе и в слабо формализованных [1,2]. Информация о причинно-следственных отношениях позволяет не только описывать и пополнять структуру знаний об объектах и явлениях в данной предметной области, но и проводить планирование и прогнозирование, что особенно ценно в тех направлениях, для которых не создано общих формальных моделей, таких как психология и медицина.

Существующие методы анализа каузальных отношений, как статистические, так и логические, сталкиваются с рядом существенных ограничений. Статистические методы анализа (корреляционный, факторный и др.) требовательны к качеству выборки, по которой

принимается решение о распространении найденных закономерностей на всю генеральную совокупность. Зачастую, особенно в практических задачах, имеющих дело с неполными и слабоструктурированными данными, проверить и удовлетворить статистические критерии не удается.

С другой стороны, логические методы (такие как ДСМ [3,4], анализ формальных понятий [5,6]), несмотря на большой прогресс в данной области за последнее время, до сих пор сталкиваются с проблемами масштабируемости и экспоненциальной сложности при работе с большими объемами данных. Однако существенные преимущества многих логических методов, заключающиеся в интуитивности, подобии алгоритма их работы процессу принятия решений экспертом, в слабых требованиях (либо полном их отсутствии) к качеству выборки, приводят к задаче предварительной обработки данных с целью снятия имеющихся ограничений для этих методов.

Один из наиболее часто используемых подходов предварительной обработки данных заключается в отборе наиболее информативных признаков, т.е. в сокращении пространства поиска причинно-следственных отношений. Существует достаточно большое количество методов выбора системы информативных признаков [7], однако задача поиска накладывает определенную специфику, что требует развития специальных подходов. Причины проявления того или иного свойства зачастую бывают скрытыми, что хорошо известно из факторного анализа в статистике. Например, на практике часто встречается ситуация, когда за проявление свойства отвечает целое множество других свойств, которые по отдельности не всегда являются информативными при разграничении одного класса объектов от другого.

Таким образом, возникает задача разработки метода, который включает предварительное обучение на данных и не позволяет исключать важные для описания причинно-следственных отношений свойства при редукции пространства признаков. Описанию данной задачи посвящен следующий раздел.

1. Постановка задачи

В разрабатываемом методе для определения значимых свойств предлагается модифицировать известный алгоритм индуктивного обучения AQ (quasi-minimal algorithm) [8,9] с помощью которого строятся правила, описывающие все объекты класса с учетом объектов других классов. Далее будем понимать под признаком некоторое имя (индекс) с соответствующим ему множеством допустимых значений, а под свойством – конкретное значение или дизъюнкцию значений признака. Каждое AQ правило для класса представляет

собой множество свойств, которые должны присутствовать у тестового объекта, чтобы он относился к этому классу. Иными словами с помощью классического AQ метода строится *отличительное описание* класса объектов. Кратко опишем алгоритм его работы:

1. Все объекты анализируемого класса объявить положительными примерами, остальные объекты – отрицательными примерами.

2. Произвольным образом выбрать опорный пример и построить из его свойств начальное правило.

3. Выбрать некоторое свойство и расширить начальное правило либо путем добавления к этому свойству дополнительного интервала значений соответствующего признака, либо путем удаления данного свойства из правила (обобщение по свойствам). При этом учитывается, что расширенное правило не должно покрывать ни одного отрицательного примера.

4. Повторять п.3 до тех пор, пока не будут получены максимально общие правила, которые будут являться листьями дерева правил.

5. Из наиболее общих правил выбрать наилучшее в том смысле, что оно покрывает максимальное количество положительных примеров и обладает минимальной длиной, вычисляемой по количеству участвующих в правиле свойств [10].

6. Удалить из рассматриваемых положительных примеров все, покрытые найденным правилом, и, если полученное множество не пусто, перейти к шагу 2.

В работе [10] было предложено в качестве наиболее информативных свойств для поиска причинно-следственных отношений использовать именно те свойства, которые входят в построенные правила для каждого класса. Однако, получаемый набор свойств обладает существенным недостатком связанным с тем, что при использовании AQ обучения, из-за случайного выбора начального объекта и расширяемых свойств, не гарантируется поиск именно таких правил, которые были бы минимальными по длине и максимальными по покрытию при рассмотрении всего множества возможных правил.

В [10] было предложено проводить несколько запусков AQ с определением статистики по встречающимся в правилах свойствам, однако такой подход окончательно не устраняет указанную проблему и страдает проблемами масштабируемости и производительности на больших массивах данных. Настоящая работа посвящена описанию метода анализа данных, использующего такую редукцию множества свойств, при которой сохраняются необходимые для поиска каузальных отношений свойства. Метод, как в части предварительного обучения, так и в части сформированных баз фактов, работает на большом массиве неполных данных без существенного ухудшения производительности.

2. Метод

2.1. Предварительная обработка

Пусть имеются три множества: $O=(o_1,o_2,\ldots,o_n)$ – множество объектов, $C=(c_1,c_2,\ldots,c_k)$ – совокупность непересекающихся множеств объектов, т.е. классов, $P=(p_1,p_2,\ldots,p_m)$ – множество признаков, каждый из которых обладает своим множеством допустимых значений W_j, $j=1,\ldots,m$. Составим матрицу значений признаков A (см. Табл. 1), в которой каждому признаку p_j поставим в соответствие столбец его значений (a_{1j}, a_{2j},…, a_{nj}), а каждому объекту o_i – его описание $<(p_1,a_{i1}), (p_2,a_{i2}),\ldots, (p_m,a_{im})>$, где пара (p_j,a_{ij}) называется свойством объекта.

Табл. 1. Общий вид матрицы A.

	p_1	p_2		p_m
o_1	a_{11}	a_{12}	…	a_{1m}
o_2	a_{21}	a_{22}	…	a_{2m}
	…	…	…	…
o_n	a_{n1}	a_{n2}	…	a_{nm}

В качестве допустимых признаков рассматриваются как номинальные, так и интервальные. Множество значений W_j интервального признака p_j подвергается дискретизации путем разделения на непересекающиеся подмножества. Каждое такое подмножество будем обозначать индексом w_i^j, где j – индекс признака, а i – индекс подмножества. В качестве алгоритма такого разбиения используется метод χ-слияния [11]:

1. Все объекты сортируются по значению интервального признака.

2. Запускается итеративный процесс попарного слияния тех соседних интервалов, для которых оказалось минимальным значение $\chi^2=\sum_{i=1}^{q}\sum_{j=1}^{k}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$, где q – количество сравниваемых интервалов, k – это количество классов, n_{ij} – количество примеров класса c_j в i-ом интервале, $e_{ij}=\frac{\sum_{j=1}^{k}n_{ij}\sum_{i=1}^{q}n_{ij}}{\sum_{j=1}^{k}\sum_{i=1}^{q}n_{ij}}$ – ожидаемая частота n_{ij}.

3. Слияние продолжается до тех пор, пока χ^2 не превысит некоторый порог, обычно равный 0.9.

Полученная дискретизованная матрица значений признаков A' поступает на этап обучения, в результате которого выбираются наиболее информативные признаки в смысле качества описания классов объектов.

2.2. Сокращение пространства признаков

Предлагаемый в работе метод для сокращения пространства признаков заключается в отборе наиболее информативных признаков путем построения специального набора AQ правил. Будем считать, что признак тем важнее, чем больше положительных примеров покрывает AQ правило с его участием. Для каждого положительного примера существует правило, покрывающее вместе с ним максимальное число других примеров. Набор таких правил, покрывающий все положительные примеры, и будет содержать наиболее значимые признаки. В обозначениях, введенных в п. 2.1, AQ правило представляет собой подмножество множества свойств *H*, где каждое свойство $h_j \in H$ имеет вид $(p_j, \bigcup_q w_q^j)$, где p_j – это имя (индекс) признака, а w_q^j – один из интервалов множества значений W_j признака p_j, которые строятся на этапе предварительной обработки данных.

Опишем общий алгоритм обнаружения наиболее информативных признаков:

1. Отметить все положительные примеры как непокрытые.
2. Построить правило *R*, которое покрывает наибольшее число положительных примеров, включая хотя бы один непокрытый пример, и не покрывает ни одного отрицательного примера, путем обобщения по свойствам, начиная с некоторого начального объекта o_i: $<(p_1, w_{i_1}^1),(p_2, w_{i_2}^2),\ldots,(p_m, w_{i_m}^m)>$.
3. Если таких правил несколько, выбрать правило, покрывающее наибольшее число непокрытых примеров. Если и таких правил несколько, выбрать правило наименьшей длины, т.е. правило, состоящее из минимального набора свойств, включающих минимальное число интервалов.
4. Отметить положительные примеры, покрытые найденным правилом.
5. Если не осталось непокрытых положительных примеров, перейти к шагу 6, иначе перейти к шагу 2.
6. Объединить свойства, входящие в найденные правила. Построенное объединение и содержит наиболее информативные признаки и будет использовано на этапе извлечения причинно-следственных отношений.

Представленный алгоритм включает в себя многократное решение оптимизационной задачи (шаги 2 и 3), которую можно свести к задаче оптимизации с бинарными переменными:

$$F(\bar{x}) \to \max,$$

где $\overline{x} = (x_{11},...,x_{1q_1},x_{21},...,x_{2q_2}...,x_{m1},...,x_{mq_m})$ – вектор, задающий правило, где $x_{ji} \in \{0,1\}$ – переменная, принимающая значение 1 при наличии *i*-го значения *j*-го признака (w_i^j) в правиле и 0 при его отсутствии.

Функционал $F(\overline{x})$ может быть задан следующим образом:

$$F(\overline{x}) = \begin{cases} aN_{\text{cov}} + \beta N_{un\text{cov}} - \gamma N_{ones}, N_{un\text{cov}} > 0; \\ \beta N_{\text{cov}} - \gamma N_{ones}, N_{un\text{cov}} = 0, \end{cases}$$

где N_{cov} – число всех положительных примеров, удовлетворяющих вектору $\overline{x}$, N_{uncov} – число еще непокрытых положительных примеров, удовлетворяющих $\overline{x}$, N_{ones} – число значений признаков, участвующих в правиле, α, β, γ – константы, заданные так, чтобы первое слагаемое в суммах функционала вносило наибольший вклад, а последнее слагаемое – наименьший вклад:

$$\alpha >> \beta >> \gamma > 0 .$$

Такой вид функционала гарантирует, что оптимальное правило будет удовлетворять условиям, заданным на шаге 2 и 3 алгоритма.

Решение этой задачи оптимизации затруднительно проводить с применением классических методов ввиду того, что при большом числе признаков и принимаемых ими значений задача имеет большую размерность, а функционал $F(\overline{x})$ задана алгоритмически. В связи с этим предлагается использовать генетический алгоритм (ГА), успешно применяющийся для решения нестандартных оптимизационных задач [12].

Опишем параметры генетического алгоритма. Индивид представляет собой правило, закодированное в виде бинарного вектора $\overline{x}$. Функция пригодности соответствует функционалу $F(\overline{x})$. Чтобы избежать перебора значений основных операторов генетического алгоритма, была применена модификация стандартного ГА, требующая настройки меньшего числа параметров, названная коэволюционным асимптотическим генетическим алгоритмом (КАГА) [13]. Эта модификация представляет собой несколько параллельно работающих вероятностных асимптотических ГА, которые конкурируют за общий ресурс – число индивидов в популяции, и обмениваются лучшими найденными решениями. Участвующие алгоритмы имеют адаптивный оператор мутации и различаются типом селекции. Такая комбинация алгоритмов позволяет избежать выбора значений операторов селекции, рекомбинации и мутации, которые являются индивидуальными для каждой решаемой задачи.

2.3. Извлечение причинно-следственных отношений

В результате этапа сокращения пространства признаков для каждого исследуемого класса объектов строится так называемое признаковое описание, в которое входят наиболее

информативные свойства в смысле, определенном в п. 2.2. Для класса c_k такое описание представляет собой множество свойств D_k, полученных на предыдущем этапе. Предварительно полученные описания классов проверяются на следующие случаи:

1) если два свойства $h_1 = (p_1, \bigcup_i w_i^1)$ и $h_2 = (p_2, \bigcup_j w_j^2)$ содержат один и тот же признак, т.е. $p_1=p_2$, и разные, но не вложенные, интервалы для этого признака, т.е. $\bigcup_i w_i^1 \neq \bigcup_j w_j^2$, $\bigcup_i w_i^1 \not\subset \bigcup_j w_j^2$ и $\bigcup_j w_j^2 \not\subset \bigcup_i w_i^1$, то такие свойства считаются конфликтными и оба исключаются из базы фактов;

2) если два свойства $h_1 = (p_1, \bigcup_i w_i^1)$ и $h_2 = (p_2, \bigcup_j w_j^2)$ содержат один и тот же признак, т.е. $p_1=p_2$, и разные вложенные интервалы для этого признака, т.е. $\bigcup_i w_i^1 \neq \bigcup_j w_j^2$ и либо $\bigcup_i w_i^1 \subset \bigcup_j w_j^2$, либо $\bigcup_j w_j^2 \subset \bigcup_i w_i^1$, то такие свойства считаются вложенными и исключается более общее свойство.

На основе описания классов строится множество баз фактов ДСМ-метода для каждого отдельного целевого свойства. В качестве целевых свойств могут быть выбраны все имеющиеся свойства, в том числе и классовое свойство, что приведет к поиску всех существующих в данном массиве данных причинно-следственных отношений. Объекты всех классов делятся на два непересекающихся множества $E^+(h_g)$ - положительные примеры для целевого свойства h_g и $E^-(h_g)$ – отрицательные примеры для целевого свойства h_g. При анализе такой базы фактов для некоторого признака класса c_l каждый объект класса c_k представляется в виде бинарного вектора $\overline{y}_i^k = (y_{i1}^k, y_{i2}^k, ..., y_{i|D_l|}^k)$, где $|D_l|$ - мощность описания класса c_l, а компонент y_{ij}^k принимает значение 1, когда у объекта присутствует j-ое свойство из D_l.

Таким образом, база фактов для некоторого целевого свойства h_g класса c_l представляет собой бинарную матрицу $M^l(h_g)$ размерности $|D_l| \times |O|$, где O – множество объектов всех классов. Задача поиска каузальных отношений для свойства h_g в таком случае сводится к поиску множества максимальных пересечений $I^l(h_g)$ по алгоритму Норриса [14,15]:

1. Берём i-ый объект из матрицы $M^l(h_g)$ - вектор $\overline{y}_i^k$.

2. Находим пересечение с имеющимися максимальными пересечениями: $T_i(h_g) = \{\overline{y}_j \wedge \overline{y}_i^k \mid \overline{y}_j \in I^l(h_g)\}$. Для каждого пересечения $\overline{y}_j$ запоминаем его множество образующих G_j, т.е. номера строчек матрицы $M^l(h_g)$, пересечением которых оно образовано.

3. Из множества $T_i(h_g)$ удаляем элементы, являющиеся подмножеством других элементов:

$$T_i'(h_g) = \{\overline{y}_j \mid \nexists \overline{y}_q \in T_i(h_g) : j \neq q, \overline{y}_q \wedge \neg\overline{y}_j = 0, G_j \subseteq G_q\}.$$

4. Для каждого имеющегося максимального пересечения $\overline{z}_j$ из $I^l(h_g)$

а) добавить индекс объекта в множество образующих G_i пересечения $\overline{z}_j$, если $\overline{y}_i^k \wedge \neg\overline{z}_j = 0$;

б) иначе добавить в $I^l(h_g)$ новое пересечение $\overline{w} = \overline{z}_j \wedge \overline{y}_i^k$, если оно не пусто и входит в множество $T_i'(h_g)$.

5. Добавить в $I^l(h_g)$ сам объект $\overline{y}_i^k$, если он не является подмножеством ни одного пересечения из $I^l(h_g)$.

6. Принять i=i+1 перейти к шагу 1, если не все объекты матрицы $M^l(h_g)$ просмотрены, иначе завершит алгоритм.

Полученное множество максимальных пересечений является множеством гипотез о причинах целевого свойства [16], которые генерируются на основе первого этапа ДСМ-метода. Элементы этого множества удовлетворяют следующим достаточным условиям принятия гипотез:

а) *условие сходства*: если все объекты, в которых встречается целевое свойство, имеют только один общий фактор (множество свойств), то этот фактор и является причиной рассматриваемого свойства;

б) *условие различия*: если оба объекта сходны по всем факторам (множествам свойств) кроме одного и этот фактор присутствует в том объекте, в котором присутствует целевое свойство, то он является причиной рассматриваемого свойства.

Необходимое и достаточное *условие абдукции*, о том, что полученные гипотезы о причинах объясняют начальное множество фактов, может быть удовлетворено при проверке таких условий, как запрет на контрпримеры, единственность причины и непротиворечивость гипотез и некоторые другие. Несмотря на то, что не все такие предикаты на данном этапе работы были проверены, будем в дальнейшем считать, что критерий абдукции выполнен, и все гипотезы из полученного множества объясняют начальное множество фактов.

Выдвигаемые в результате работы первого этапа ДСМ-метода гипотезы о причинах целевого свойства $h_g = (p_g, \bigcup_q w_q^g)$ для класса c_k являются множеством найденных алгоритмом Норриса свойств:

$$H(h_g, c_k) = < (p_{g_1}, \bigcup_q w_q^{g_1}), (p_{g_2}, \bigcup_q w_q^{g_2}), ..., (p_{g_t}, \bigcup_q w_q^{g_t}) >,$$

где t – сложность (длина) найденной причины. Построенное множество гипотез редуцируется путем удаления незначащих причин, длина которых превышает некоторый порог, и тех причин, которые являются подмножествами других.

3. Эксперимент

Проверка эффективности предложенного метода проводилась на базе медицинских данных MIMIC II [17], которая содержит большое количество результатов анализов разного вида для пациентов с различными заболеваниями с учетом истории (Рис. 1). При выборке объектов из базы рассматривались такие болезни, для которых длительность наблюдений и количество обращений были сбалансированы, т.е. для пациента имеется хотя бы одно непрерывное длительное обращение (40-70 дней) и таких пациентов не менее 50.

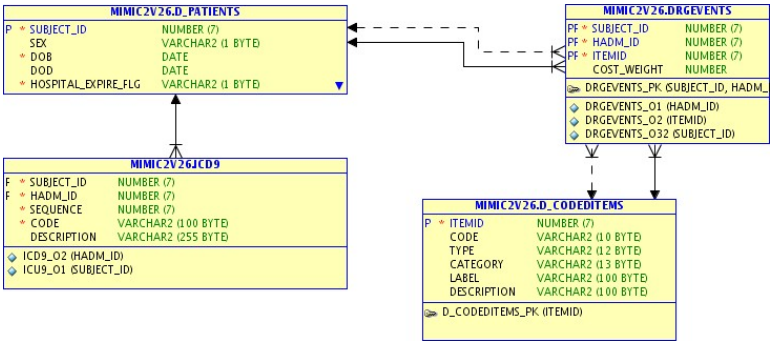

Рис. 1. Фрагмент базы данных MIMIC со связями основных таблиц: пациентов (D_PATIENTS) и диагнозов (DRGEVENTS).

Были выбраны две группы пациентов, которые имели два различных заболевания (65 и 60 пациентов соответственно). Применение коэволюционного алгоритма позволило построить правила, отделяющие пациентов разных классов по совокупности признаков. Участвующие в правилах признаки являются наиболее информативными и были использованы для построения гипотез, представленных ниже. Для полученных наборов правил было проведено сравнение с наборами правил, получаемых с помощью классического AQ-метода. Сравнение представлено в Табл. 2. Последняя колонка P показывает, какая часть правил покрывает большое число объектов (по крайней мере, одну десятую всех объектов).

Табл. 2. Сравнение методов построения покрытий.

	Число правил	Длина правил	Покрытые объекты (%)	P (%)
AQ-метод	21-23	9-11	100	10-15
GAAQ-метод	32-36	14-16	100	50-60

Оба метода построили правила, покрывающие все положительные объекты, при этом размер правил почти совпадает в обоих случаях. Необходимо отметить следующие различия. AQ-метод строит меньшее число правил, однако лишь небольшая часть из них покрывает большое число объектов, что является признаком того, что в правилах используются не самые значимые признаки. Разница в покрытиях продемонстрирована на Рис. 2, для наглядности приведены по два покрытия для каждого метода.

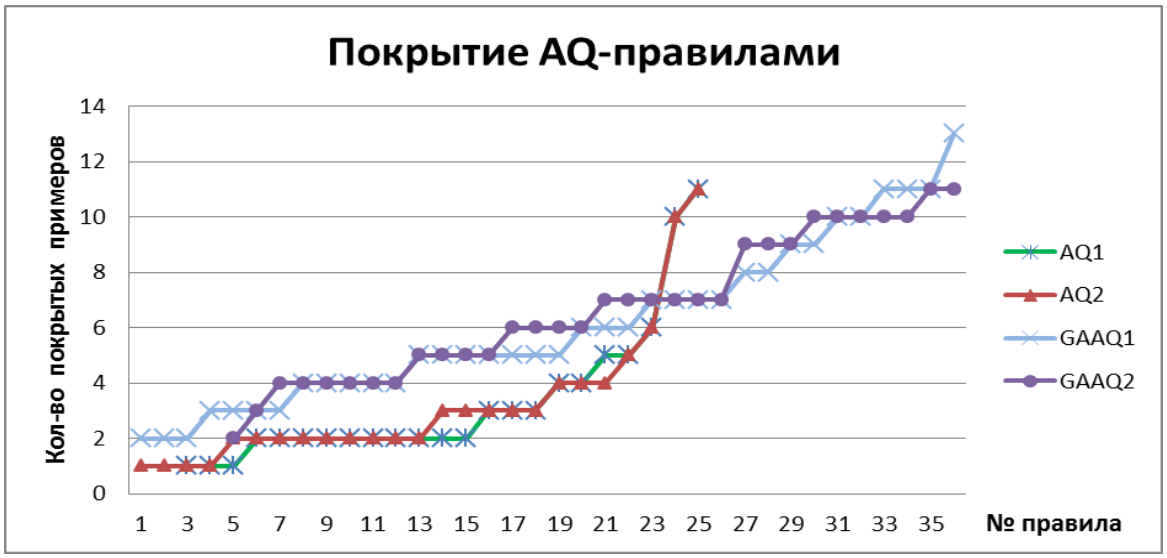

Рис. 2. Разница в покрытиях рассмотренными методами.

Из рисунка видно, что кривые, соответствующие предложенному методу, располагаются выше кривых AQ-метода. Интересующие правила, покрывающие одновременно более одной десятой объектов, располагаются в правой части графика и принадлежат в основном GAAQ-методу. AQ-метод генерирует в среднем три правила (лишь одна десятая всех правил), покрывающих более одной десятой всех объектов, в то время как GAAQ-метод строит более двадцати таких правил (более половины всех правил). Этим числам соответствует последняя колонка Табл. 2.

Другим отличием является то, что правила, которые строит GAAQ-метод, существенно отличаются друг от друга, тогда как правила AQ-метода похожи между собой и содержат не

все информативные признаки, требуемые для построения гипотез с помощью ДСМ-метода, что будет показано ниже.

Полученные разными методами наборы информативных признаков подавались на вход модулю анализа причинно-следственных отношений. База фактов во всех случаях формировалась из 30 свойств, встречающихся в самых лучших с точки зрения покрытых примеров правилах. На Рис. 3 представлены диаграммы количества найденных отношений для каждого класса. Как следует из рисунка, количество найденных отношений для набора, построенного с помощью GAAQ, больше, чем в случае AQ метода. Этот результат свидетельствует в пользу того, что описания классов, полученные разрабатываемым методом, обладают лучшим качеством с точки зрения поиска причинно-следственных отношений, не теряя при этом ни свойства полноты по покрываемым примерам, ни свойства минимальной длины описания.

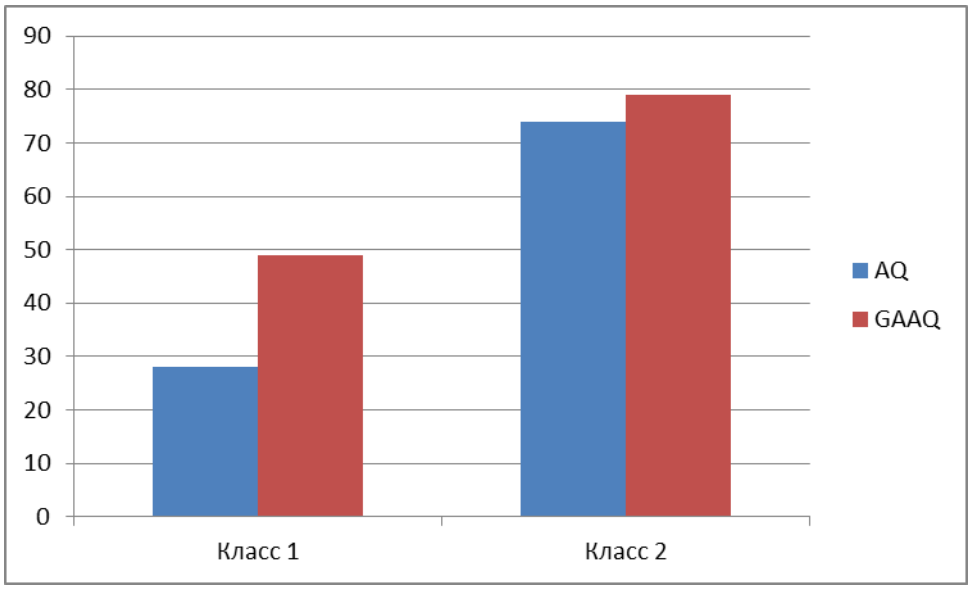

Рис. 3.Усредненное количество причинно-следственных отношений (высота диаграмм), найденных на базах фактов, построенных методами GAAQ и AQ.

Заключение

В работе построен и исследован метод извлечения причинно-следственных бинарных отношений с предварительным обучением, позволяющий анализировать неполные данные большого объема. В качестве метода сокращения пространства признаков был использован коэволюционный генетический алгоритм. Формируемое множество баз фактов анализировалось первым шагом ДСМ-метода с использованием алгоритма Норриса. Были проведены модельные эксперименты на медицинских данных из базы MIMIC II, результаты которых подтвердили эффективность предложенного метода. Следует отметить, что полностью

вопрос корректности применения метрических средств по сокращению пространства признаков для задачи извлечения причинно-следственных отношений остается открытым, не смотря на то, что экспериментальные результаты показывают оправданность их использования.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 13-07-12127 офи_м и № 14-07-31194 мол_а).

Литература

1. Осипов Г. С. Построение моделей предметных областей. Ч. I. Неоднородные семантические сети // Известия АН СССР. Техническая кибернетика. — 1990. — № 5. —С. 32–45.

2. Осипов Г. С. Приобретение знаний интеллектуальными системами: основы теории и технологии. — М. : Физматлит, 1997.

3. Финн В.К. Дистрибутивные решётки индуктивных процедур // Научно-техническая информация. — 2014. — № 11. — С. 1-31.

4. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ-метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. — 2010. — № 4. — С. 41–48.

5. B. Ganter, R. Wille Formal Concept Analysis: Mathematical Foundations. — Springer, 1999.

6. Кузнецов С.О., Объедков С.А. Алгоритмы построения множества всех понятий формального контекста и его диаграммы Хассе // Известия Академии Наук. Теория и системы управления. — 2001. — № 1. — С. 120-129.

7. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. — С. 270.

8. 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.

9. 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.

10. Панов А. И. Выявление причинно-следственных связей в данных психологического тестирования логическими методами // Искусственный интеллект и принятие решений. — 2013. — № 1. — С. 24–32.

11. 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.

12. Емельянов В. В., Курейчик В. В., Курейчик В. М. Теория и практика эволюционного моделирования. — М: Физматлит, 2003.

13. Заблоцкий С.Г., Семенкин Е.С., Швец А.В. Коэволюционный асимптотический генетический алгоритм для формирования предложений по слоговой модели в системе автоматического распознавания слитной речи // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнёва. — 2011. — № 3.

14. Norris E. M. Maximal rectangular relations // Fundamentals of Computation Theory / Ed. by Marek Karpinski. — Springer Berlin Heidelberg, 1977. — P. 476–481.

15. Kuznetsov S. O., Obiedkov S. A. Comparing performance of algorithms for generating concept lattices // Journal of Experimental and Theoretical Artificial Intelligence. — 2002. — Vol. 14. — P. 189–216.

16. Волкова А.Ю., Шестерникова О.П. О создании интеллектуальных систем, реализующих ДСМ-метод автоматического порождения гипотез, и результатах их применения для анализа медицинских данных // Научно-техническая информация. Сер. 2. – 2013. – № 5. – С. 10-15.

17. Massachusetts Institute of Technology. MIMIC II Database URL:https://mimic.physionet.org/database.html (дата обращения: 16.12.2014).

СВЕДЕНИЯ ОБ АВТОРАХ

Панов Александр Игоревич. Научный сотрудник ИСА РАН.
Окончил в 2009 г. Новосибирский государственный университет, в 2011 г. Московский физико-технический институт (государственный университет). Количество печатных работ: 23. Область научных интересов: методы машинного обучения, распознавание образов, когнитивное компьютерное моделирование, мультиагентные системы. E-mail: pan@isa.ru.

Швец Александр Валерьевич.
Младший научный сотрудник Института системного анализа Российской академии наук. Окончил Сибирский федеральный университет в 2011 году. Автор 16 научных работ. Область научных интересов: компьютерная лингвистика, математическое моделирование, методы оптимизации, искусственный интеллект. E-mail: shvets@isa.ru.

Волкова Галина Дмитриевна.
Профессор каф. Информационных технологий и вычислительных систем ФГБОУ ВПО МГТУ "СТАНКИН". Окончила Московский авиационный институт им.С.Орджоникидзе в 1976 году. Автор 161 научных и учебно-методических работ. Область научных интересов: автоматизация интеллектуального труда, когнитивные технологии проектирования, моделирование и представление знаний. E-mail: cog-par@yandex.ru.

Method of Extraction of Causal Relationships with Optimized Fact Bases

Abstract. In this paper a method of extraction of causal relationships on the set of fact bases obtained by learning on data from weakly formalized subject area is proposed and studied. Fact bases are built for the target properties of each object class. Learning is carried out with the use of co-evolutionary genetic algorithm, which reduces the initial feature space. The formed class descriptions obtained by means of the first phase of the JSM-method are used for search of causal relationships for all target properties. The proposed method is suitable both for the analysis of a small amount of data and for work with sets of incomplete data of the big size. A number of model experiments with use of base of medical data MIMIC II is carried out.

Keywords: machine learning, genetic algorithm, causal relationships, JSM-method, AQ-learning.

Information about authors

Panov Aleksandr. Researcher of Institute for Systems Analysis RAS. Graduated in 2009, Department. of Physics at Novosibirsk State University, in 2011, Department of Control/Management and Applied Mathematics at Moscow Institute of Physics and Technology (State University). Number of publications: 23. Research interests: machine learning, pattern recognition, cognitive computer modeling, multiagent systems. E-mail: pan@isa.ru.

Shvets Alexander. Junior researcher of Institute for Systems Analysis RAS. Graduated in 2011, Dept. of mathematics at Siberian Federal University. Number of publications: 16. Research interests: computer linguistics, mathematical modelling, optimization methods, artificial intelligence. E-mail: shvets@isa.ru.

Volkova Galina, professor of Moscow State University of Technology “STANKIN”, Department of Information Technology and Computer Systems. Graduated in 1976, Moscow Aviation Institute. Number of publication: 161. Research interests: automation of intellectual labor, cognitive design technologies, knowledge modeling and knowledge representation. E-mail: cog-par@yandex.ru.

References

1. Osipov G. S. Formulation of Subject Domain Models Part I. Heterogeneous Semantic Nets // Journal of Computer and Systems Sciences International. — 1992. —Vol. 30, No. 5. —P. 1–12.

2. Osipov G. S. Priobretenie znanij intellektual'nymi sistemami: osnovy teorii i texnologii. — Moscow: Fizmatlit, 1997.

3. Anshakov O. M., Skvortcov D. P., Finn V. K. On logical construction of JSM-method of automated hypotheses generation // DOKLADY AKADEMII NAUK SSSR, Vol. 320, No.6, pp. 1331-1336, 1991.

4. Finn V. K. On the definition of empirical regularities by the JSM method for the automatic generation of hypotheses // Scientific and Technical Information Processing. — 2014. — Vol. 39, no. 5. — P. 261–267.

5. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. — Springer, 1999.

6. Kuznetsov S.O., Obedkov S.A. Algorithms for Constructing a Set of AH Concepts of Formal Context and their Hasse Diagrams // Journal of Computer and Systems Sciences International. — 2001. —Vol. 40, No. 1. —P. 115–124.

7. Zagorujko N.G. Prikladnye metody analiza dannyx i znanij. - Novosibirsk: IM SO RAN, 1999, P. 270.

8. 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.

9. 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.

10. Panov A. I. Extraction of cause-effect relationships from psychological test data using logical methods // Scientific and Technical Information Processing. — 2014. — Vol. 41, no. 5. — P. 1–8.

11. 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.

12. Goldberg D. Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley Professional, 1989.

13. Zablotskiy S., Shvets A., Sidorov M., Semenkin E. and Minker W. Speech and Language Resources for LVCSR of Russian // Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). – Istanbul, 2012. – P. 3374-3377

14. Norris E. M. Maximal rectangular relations // Fundamentals of Computation Theory / Ed. by Marek Karpinski. — Springer Berlin Heidelberg, 1977. — P. 476–481.
15. Kuznetsov S. O., Obiedkov S. A. Comparing performance of algorithms for generating concept lattices // Journal of Experimental and Theoretical Artificial Intelligence. — 2002. — Vol. 14. — P. 189–216.
16. Volkova A. Yu. Analyzing the Data of Different Subject Fields Using the Procedures of the JSM Method for Automatic Hypothesis Generation // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, No. 3. – pp. 127–139.
17. Massachusetts Institute of Technology. MIMIC II Database. — 2014. — URL:https://mimic.physionet.org/database.html (access: 16.12.2014).