Экспериментальное исследование численных мер оценки ассоциативных и причинных связей в больших данных

О.Н. Тушканова

Аннотация. В работе приводится краткое описание сущности методов ассоциативного и причинного анализа данных и проблем, затрудняющие его применение в больших данных. Описывается схема ускоренного поиска множества причинных связей. Приводится список численных мер, предложенных к настоящему времени для оценки "силы" ассоциативной связи пары атрибутов в статистике, социологии, машинном обучении и интеллектуальном анализе данных. Приводятся результаты анализа их формальных свойств, в терминах которых формулируются необходимые условия, которым должны удовлетворять меры связи причинного характера. Описываются результаты экспериментального исследования выделенных численных мер, которые позволяют сформировать упорядоченный список наиболее перспективных мер, пригодных для оценки силы причинной связи.

Ключевые слова: ассоциативная мера, причинная мера, причинный анализ, большие данные.

Введение

Ключевые и наиболее сложные проблемы интеллектуального анализа данных в настоящее время относятся к его разделу, который называется *большие данные* (*англ. Big Data* [1]). Примерами больших данных являются потоки текстовых сообщений из социальных сетей, пространственно–временные данные результатов исследований окружающей среды, потоки данных о соединениях абонентов сотовой связи, данные интернет–торговли (содержащие информацию о покупателях, товарах, динамике и структуре покупок), данные о финансовых потоках банков с распределенными офисами, числовые и фотографические данные научно-технических исследований, разведывательные данные о различных объектах и другая информация военного назначения.

Такие данные чаще всего описываются очень большим количеством (сотни и тысячи) разнородных атрибутов, которые, в том числе, могут представлять собой изображения и тексты на естественном языке (т.е. быть неструктурированными), а также хранятся в распределенных базах данных.

Согласно [2], две основные цели анализа больших данных – это разработка эффективных методов прогнозирования и выявление существующих зависимостей между атрибутами данных.

Практика показала, что большинство методов анализа данных для принятия решений не работают в области больших данных либо в связи с их вычислительной неустойчивостью, либо из-за проблем вычислительной сложности. Следовательно, требуется разработка новых методов и алгоритмов анализа больших данных.

К числу перспективных в этой области относят большую группу методов, которые объединяются общим названием *ассоциативный* и *причинный* анализ данных. В этих методах акцент делается на выделение и анализ связей между атрибутами, задающими данные. На практике эти методы зарекомендовали себя с лучшей стороны в таких задачах, как анализ социальных сетей, например, для выявления опасных сообществ

или неформальных лидеров, персонификация рекламы в интернет–торговле, распознавание объектов в различных приложениях, диагностика сложной техники и т.д.

Важно отметить, что имеются теоретические и экспериментальные работы, в которых строго показано, что среди ассоциативных связей наиболее полезными с точки зрения информативности в задачах принятия решений являются связи причинного характера [3, 4].

Однако поиск причинных структур, традиционно используемый в причинном анализе, и основанный на построении и анализе байесовских сетей доверия [5], требует решения задач обучения экспоненциальной сложности относительно числа атрибутов [3]. Поэтому использование такой модели в задачах причинного анализа больших данных бесперспективно. Стоит также отметить, что авторы [6], считают, что в байесовской сети, построенной автоматически на некоторых данных, далеко не все выявленные между переменными связи являются причинными. Эти обстоятельства делают актуальной задачу разработки альтернативных методов поиска причинных связей в данных. Такие методы должны, прежде всего, обладать вычислительной эффективностью, что может быть достигнуто за счет простых алгоритмов фильтрации множества атрибутов, которые потенциально могут быть кандидатами в число причин для той или иной целевой переменной.

Известно, что причинная связь переменных является частным случаем ассоциативной связи [7, 8, 9], а для обнаружения последней предложено большое количество численных мер. Поэтому естественно простые численные меры оценки причинной связи искать среди мер, предложенных для оценки силы ассоциаций. Теоретический и экспериментальный анализ таких мер для определения их пригодности для решения задач причинного анализа больших данных является целью данной работы.

Дальнейшее содержание работы организовано следующим образом. Во втором разделе описаны необходимы свойства причинных мер связи и формальные требования к ним. В третьем разделе представлены меры ассоциативной связи атрибутов, предложенные к настоящему времени, среди которых следует искать меры причинной связи. Выполнен формальный анализ приведенных мер и среди них выделены те, которые соответствуют формальным требованиям к мерам причинной связи. В четвертом разделе описаны и обоснованы методы сравнительного экспериментального исследования выделенных мер. Численные результаты исследования позволяют сделать выводы относительно применимости тех или иных мер для в выявлении причинных связей между атрибутами. В заключения сформирована список наиболее перспективных мер с точки зрения способности выявлять причинные связи в больших данных.

1. Необходимые свойства причинных мер связи и формальные требования к ним

К числу наиболее известных работ в области ассоциативного и причинного анализа относятся работы [3, 4, 10-12]. В работах [3, 4] показывается, что *причинный анализ данных* фактически предлагает новый подход к решению классической задачи анализа данных, в частности, задачи поиска и отбора признаков.

Наиболее распространенная схема ускоренного поиска множества причинных связей состоит в том, что сначала для целевой переменной выделяются атрибуты данных, связанные с ней значимой ассоциативной связью. Для этого используется некоторая численная мера ассоциации. С ее помощью на первом шаге находится множество атрибутов, часть из которых может быть связана с целевой переменной причинной связью. На втором шаге для каждого такого атрибута проверяется гипотеза о том, что найденная для него ассоциативная связь с целевой переменной является причинной связью. Такая схема используется, например, в работе [9]. Объем вычислений при этом зависит, в частности, от того, насколько удачно выбрана мера ассоциативной связи, поскольку разные меры могут генерировать разные множества атрибутов и обладать разной вычислительной эффективностью.

При анализе известных мер ассоциативной связи с позиций причинности предпочтения на их множества должны формироваться с учетом трех их свойств:

1) насколько эффективна та или иная мера с вычислительной точки зрения (насколько быстро она может быть вычислена для всего множества пар атрибутов больших данных);

2) отражает ли эта мера причинный характер найденной связи между парой переменных, которые рассматриваются в качестве аргументов этой меры;

3) насколько полный перечень причин целевой переменной она позволяет находить.

Применительно к каждой мере ассоциативной связи, предложенной к настоящему времени, вопрос относительно способности этой меры выявлять причинные связи между атрибутами в данной работе решается в два этапа. Сначала анализируются формальные свойства меры для того, чтобы определить, может ли та или иная мера отобрать атрибуты, которые могут быть причинами целевой переменной. На втором этапе необходимо проводить экспериментальное исследование для окончательного вывода о целесообразности использования той или иной меры для поиска причин целевой переменной.

Отметим, что в качестве причин могут выступать не только отдельные атрибуты данных, но и паттерны данных. Под паттернами данных понимаются наборы значений нескольких атрибутов. Паттерны задаются на подмножестве атрибутов, в терминах которых представлены анализируемые данные. Примером паттернов являются *часто встречающиеся паттерны*, используемые при поиске ассоциативных правил [13]. Для паттернов, как и для одиночных атрибутов данных, численная мера их ассоциации с целевой переменной может характеризовать численную оценку перспективности паттерна при решении той или иной задачи. В данной работе внимание акцентируется на анализе ассоциаций в интересах поиска элементов причинных структур целевых переменных.

Для оценки свойств ассоциативной связи между переменными/паттернами (включенными, например, в ассоциативное правило) могут быть использованы объективные и/или субъективные меры связи. Субъективные меры принимают во внимание априорные знания пользователя о связях аргументов с целевой переменной и о семантике аргументов. Субъективные меры [14], такие как нетривиальность, способность предсказать действие, которое должно быть выполнено (*actionability*) и новизна, или неожиданность (способность отражать новое, не очевидное знание) обычно невозможно определить формально, поскольку они отражают знание, интерпретация которого дается экспертом. В данной работе субъективные меры не рассматриваются на этапах формального и экспериментального анализа. Они могут привлекаться для анализа семантических свойств причинных связей, полученных на основе формального анализа.

Объективные меры ориентированы на данные и принимают во внимание только их. Объективная мера есть некоторая функция $\mu = F(A,B)$, вычисляемая по выборке данных, содержащих атрибуты/паттерны A и B, которые интерпретируются как случайные события, а потому функция $F(A,B)$ является некоторой выборочной статистикой. В качестве аргументов функции $F(A,B)$ обычно рассматриваются выборочные оценки вероятностей событий A и B, так что формальное представление объективной меры связи атрибутов A и B имеет вид некоторой функции $\mu\left(n, p_A, p_B, p_{\bar{A}}, p_{\bar{B}}, p_{AB}, p_{\bar{A}B}, p_{A\bar{B}}, p_{\bar{A}\bar{B}}\right) \in R$. В этой формуле n – объем выборки, а символом p обозначены выборочные вероятности событий, указанных в его нижнем индексе. Важно также отметить, что значение меры связи может зависеть от порядка следования аргументов в ней, а в общем случае $F(A,B) \neq F(B,A)$. Объективные меры в общем случае могут принимать положительное и отрицательное значение, и задаются так, чтобы большее значение ее абсолютной величины соответствовало "*более сильному*" правилу.

Для проверки ассоциативной связи на принадлежность к кассу причинных связей предполагается использовать те меры, которые удовлетворяют некоторым формальным требованиям к мерам оценки причинности, которые, по сути, можно назвать *аксиомами* меры причинной связи. Эти требования перечислены ниже.

В силу того, что причинная связь является направленной [15], мера, которая сможет оценивать силу причинной связи должна быть некоммутативной, указывая направление связи. Обычно требуется, чтобы мера принимала значения в диапазоне [-1, 1] либо [0, 1]. И наконец, нулевое

значение меры должно указывать на отсутствие причинной связи между ее аргументами.

2. Формальный анализ и фильтрация мер ассоциативной связи

Список объективных мер связи пары атрибутов, предложенных к настоящему времени в статистике, социологии, машинном обучении и литературе по интеллектуальному анализу данных представлен на Рис. 1.

Анализ мер показывает, что некоммутативными являются следующие меры: *уверенность*, *мера Лапласа*, *J-мера*, *убеждение*, *добавочное значение*, *индекс Джини*, *мера Клозгена*, *мера Сибега и Шонауера*, *фактор определенности* и *коэффициент регрессии*. Остальные меры

№	Мера	Формула для вычисления	Область значений	Источник
1	ϕ-коэффициент	$\frac{p_{AB}-p_A p_B}{\sqrt{p_A p_B (1-p_A)(1-p_B)}}$	[-1; 1]	[16]
2	Соотношение шансов (α)	$\frac{p_{AB}p_{\overline{A}\overline{B}}}{p_{A\overline{B}}p_{\overline{A}B}}$	[0; ∞]	[17]
3	Q-коэффициент ассоциации Юла	$\frac{p_{AB}p_{\overline{A}\overline{B}}-p_{A\overline{B}}p_{\overline{A}B}}{p_{AB}p_{\overline{A}\overline{B}}+p_{A\overline{B}}p_{\overline{A}B}}$	[-1; 1]	[18]
4	Y-коэффициент ассоциации Юла	$\frac{\sqrt{p_{AB}p_{\overline{A}\overline{B}}}-\sqrt{p_{A\overline{B}}p_{\overline{A}B}}}{\sqrt{p_{AB}p_{\overline{A}\overline{B}}}+\sqrt{p_{A\overline{B}}p_{\overline{A}B}}}$	[-1; 1]	[19]
5	κ-коэффициент	$\frac{p_{AB}+p_{\overline{A}\overline{B}}-p_A p_B-p_{\overline{A}}p_{\overline{B}}}{1-p_A p_B-p_{\overline{A}}p_{\overline{B}}}$	[-1; 1]	[19]
6	J-мера (J)	$p_{AB}\log\frac{p_{B\|A}}{p_B}+p_{A\overline{B}}\log\frac{p_{\overline{B}\|A}}{p_{\overline{B}}}$	[0; 1]	[20]
7	Индекс Джини (G)	$p_A\left(p^2_{B\|A}+p^2_{\overline{B}\|A}\right)+p_{\overline{A}}\left(p^2_{B\|\overline{A}}+p^2_{\overline{B}\|\overline{A}}\right)-p^2_B-p^2_{\overline{B}}$	[0; 1]	[19,20]
8	Поддержка (sup)	p_{AB}	[0; 1]	[13]
9	Уверенность (conf) для ассоциативного пр-ла	$p_{B\|A}$	[0; 1]	[13]
10	Мера Лапласа (L)	$(np_{AB}+1)/(np_B+2)$	[0; 1]	[7]
11	Убеждение (V)	$p_A p_{\overline{B}}/p_{A\overline{B}}$	[0.5; ∞]	[8]
12	Фактор интереса (I)	$p_{AB}/p_A p_B$	[0; ∞]	[21]
13	Косинус (IS)	$p_{AB}/\sqrt{p_A p_B}$	[0; 1]	[22]
14	Мера Пятецкого-Шапиро (PS)	$p_{AB}-p_A p_B$	[-0.25; 0.25]	[21]
15	Фактор определенности (F)	$(p_{B\|A}-p_B)/(1-p_B)$	[-1; 1]	[19]
16	Добавочное значение (AV)	$p_{B\|A}-p_B$	[-0.5; 1]	[19]
17	Коллективная сила (S)	$\frac{p_{AB}+p_{\overline{A}\overline{B}}}{p_A p_B+p_{\overline{A}}\cdot p_{\overline{B}}}\times\frac{1-p_A p_B-p_{\overline{A}}p_{\overline{B}}}{1-p_{AB}-p_{\overline{A}\overline{B}}}$	[0; ∞]	[23]
18	Мера Жаккара (ζ)	$p_{AB}/(p_A+p_B-p_A p_B)$	[0; 1]	[24]
19	Мера Клозгена (K)	$\sqrt{p_{AB}}\cdot p_{B\|A}-p_B$	$[\sqrt{(2/\sqrt{3}-1)}\cdot 2-\sqrt{3}-1/\sqrt{3}, 2/3\sqrt{3}]$	[16]
20	Информационная выгода (IG)	$\log(p_{AB}/(p_A\cdot p_B))$	$[-\infty; \log(1/p_A)]$	[17]
21	Мера Сибега и Шонауера (SEB)	$p_{AB}/p_{A\overline{B}}$	[0; ∞]	[25]
22	Коэффициент регрессии	$(p_{AB}-p_A p_B)/(p_A\cdot(1-p_A))$	[-1, 1]	[26, 27]

Рис. 1. Численные меры ассоциативной связи

являются коммутативными, т.е. для них $F(A \rightarrow B) = F(B \rightarrow A)$.

Следующие меры имеют нулевое значение при отсутствии связи, но их диапазон значений отличен от [–1, 1] ([0, 1]): *убеждение*, *добавочное значение*, *мера Клозгена*, *мера Сибега и Шонауера*.

Итак, можно сделать вывод, что всем трем критериям удовлетворяют только шесть мер, а именно *уверенность*, *мера Лапласа*, *J-мера*, *индекс Джини*, *фактор определенности* и *коэффициент регрессии*. Однако, в ходе экспериментального исследования, помимо перечисленных мер, для выяснения дополнительных свойств будут проанализированы также *убеждение*, *добавочное значение*, *мера Клозгена*, *мера Сибега и Шонауера*.

3. Экспериментальное исследование мер связи

3.1. Методика экспериментального исследования

Можно использовать несколько вариантов экспериментальной оценки способности ассоциативной меры выявлять причинные зависимости между переменными. Первый из них – это сравнить множество причин, выявленных некоторой мерой, со списком причин, которые заранее известны для некоторого набора данных. Второй вариант – это сравнить причинные связи, выявленные с помощью анализируемой меры, со связями, полученными с помощью байесовской сети доверия [5] для одного и того же набора данных, если следовать общепринятому мнению о том, что причинными связями являются те и только те, которые выявляются этой сетью. Последнее утверждение, хотя оно и не бесспорно, может быть принято, если опираться на результаты работ [3, 4], в которых показано, что байесовская сеть позволяет выделить связи, весьма полезные для восстановления модели целевой переменной. Существуют, однако, работы, которые предостерегают от такой интерпретации байесовских сетей, например, [6]. Добавим, что экспоненциальная сложность построения байесовской сети дополнительно снижает ценность ее практического использования как эталона для построения причинных связей. Это метод исследования не рассматривается в данной работе.

Так как конечной целью поиска причинных связей в данных зачастую бывает их использование в процессе предсказания значений целевой переменной, в качестве еще одного способа оценки меры связи с точки зрения причинности может выступить оценка эффективности выявленных с ее помощью правил в предсказании значений целевой переменной.

Четвертый способ сравнения мер в рассматриваемом здесь смысле – это участие в соревновании по выявлению причинных связей и сравнение своих результатов с другими.

Методы один, три и четыре использованы в данном исследовании. Далее описываются результаты экспериментальных исследований мер, выделенных в предыдущем разделе, для различных наборов данных.

3.2. Причинные правила для набора данных Adult (метод 1)

В ходе исследования удалось найти только одну работу [28], в которой приводится список (точнее его фрагмент) правил, отражающих направленную причинную связь переменных и целевой переменной (назовем их для краткости причинными правилами) для модифицированного открытого набора данных *Adult* [29].

Этот набор данных представляет собой фрагмент базы данных переписи населения, проведенной бюро переписи населения США в 1994 году. На основании данных о некотором индивиде необходимо предсказать, получает ли он более $50,000 в год. Модифицированный набор данных Adult, использованный в работе [28], содержит 30160 примеров без пропущенных значений, каждый из которых описан 99 бинарными признаками.

Для выявления причинных правил в работе [28] используется разработанная авторами методика *CAR*, основанная на ретроспективном *когортном* (англ. *cohort*) исследовании, широко применяемом в медицине и социологии. С деталями методики и алгоритмами извлечения причинных правил можно ознакомиться в работе [28].

Найденные причинные связи представляются в виде однолитерных правил с меткой класса в заключении:

Если <атрибут>=true, то класс = <метка класса>,

причем метка класса принимает значение 1, если индивид зарабатывает более $50 000, и значение 0 – в противном случае. Полученный список правил авторы сравнивают с правилами, выявленными с помощью двух методик построения локальных причинных структур – *CCC* [30] и *CCU* [9].

Проведенный эксперимент включает формирование наборов причинных правил с помощью каждой из анализируемых мер двумя способами и их сравнение с результатами работы [28]. Опишем кратко способ извлечения причинных правил из данных с помощью мер причинной связи. Сначала для каждого бинарного атрибута в данных были сформированы два правила следующего вида: (1)*"если <атрибут>=true, то класс = 0"* и (2)*"если <атрибут>=true, то класс = 1"*. Затем для каждого из этих правил была вычислена одна из исследуемых мер μ. После этого правило (1 или 2), которое имеет меньшее значение исследуемой меры μ, отбрасывалось. Списки правил, полученные для каждой исследуемой меры в отдельности, являются итоговыми списками первого варианта данного метода исследования. Затем списки правил, полученные для каждой исследуемой меры, были сокращены с использованием некоторого порога θ для значения этой исследуемой меры[1]. Эти сокращенные списки являются итоговыми списками второго варианта данного метода исследования. Детали эксперимента и все полученные численные результаты размещены по адресу [31].

При формировании правил первым способом полное совпадение с *CAR* в эксперименте продемонстрировали *коэффициент регрессии*, *убеждение*, *добавочное значение*, *фактор определенности* и *мера Клозгена*. При сравнении со списками правил, полученных с помощью методик *CCC* [30] и *CCU* [9], полное совпадение по части наличия правил в списке также продемонстрировали *коэффициент регрессии*, *убеждение*, *добавочное значение*, *фактор определенности* и *мера Клозгена*.

Приведем результаты эксперимента при формировании правил вторым способом. Наилучшее совпадение со списком правил из работы [28] в этом случае демонстрирует *мера Клозгена* (18 правил из 30), далее следуют *коэффициент регрессии* и *добавочное значение* (12 правил из 30). Сравнение со списками правил, построенных методиками *CCC* [30] и *CCU* [9] показывает, что наилучшее совпадение вновь показывает *мера Клозгена*, далее следуют *коэффициент регрессии*, *добавочное значение, фактор определенности* и *убеждение*.

Следует отметить, что перечисленные меры, помимо правил, генерируемых методами [9, 28, 30], генерируют и другие правила. Их можно отнести как к лишним правилам, так и к правилам, которые являются причинными, но пропущены методами [9, 28, 30]. Заметим также, что в работе [28] представлены только 30 правил из 50 правил, найденных авторами.

Однако, простого сравнения количества общих правил в списках, полученных разными методами, недостаточно для того, чтобы сделать окончательные выводы, тем более, если сравнение выполнено только на одном наборе данных. Более убедительные результаты можно получить с помощью метода 3, который рассматривается в следующем подразделе.

3.3. Сравнение мер на основании их классификационных возможностей (метод 3)

На основе наборов правил, полученных с помощью мер, и набора правил из работы [28] были построены простые классификаторы для модифицированного набора данных *Adult* следующим образом:

- каждое правило из сформированного для каждой меры списка рассматривалось как отдельный классификатор, который может голосовать только за тот класс, который представлен в заключении соответствующего правила;
- объединение решений этих отдельных классификаторов выполняется по схеме простого голосования.

Для оценки качества исследуемых классификаторов использованы следующие данные: матрица неточностей (англ. confusion matrix) [32]; среднеквадратическое отклонение [33]; чувствительность (англ. true positive rate) [34], коэффициент ложной тревоги (англ. false positive rate – FPR) [34], точность (англ. precision) [35], полнота (англ. recall) [35], F-мера (англ. F-measure) [35] классификатора для каждого

[1] Правильный выбор значения порога θ зависит от решаемой задачи и конкретного набора данных. Методика выбора этого значения выходит за рамки данного исследования

класса; площадь под ROC-кривой для классификатора (англ. AUC) [34]. Тестирование классификаторов выполнялось с помощью процедуры скользящего контроля с 10 блоками (англ. 10-fold cross-validation) [36].

Детальное представление результатов может быть найдено по адресу [31], где приведены оценки построенных классификаторов по всем перечисленным метрикам. Дополнительно для сравнения использовался алгоритм классификации *BayesNet* [6], реализованный в системе *Weka* [37]. Напомним, что речь идет о классификации модифицированного набора данных *Adult*. Фрагмент результатов эксперимента представлен в Табл. 1.

По результатам анализа данных, полностью приведенных по адресу [31], можно сделать следующие выводы:

1. Классификаторы, использующие набор правил, полученных с помощью мер *уверенность*, *мера Лапласа* и *мера Сибега и Шонауэра*, являются неэффективными и непригодны для классификации, так как не позволяют классифицировать экземпляры одного из классов набора данных, поэтому далее они не рассматриваются.

2. Наихудшими по всем параметрам являются меры *индекс Джини* и *J-мера*.

3. Наименьшее *среднеквадратичное отклонение* показал классификатор на основе *меры Клозгена*. Эта мера также оказалась лучшей среди оставшихся по показателям *чувствительность, полнота* и *F-мера.*

4. Классификаторы на основе мер коэффициент регрессии, добавочное значение, фактор определенности и убеждение незначительно хуже классификатора на основе меры Клозгена по показателям среднеквадратичное отклонение, чувствительность, полнота и F-мера, однако, лучше его по показателям коэффициент ложной тревоги и точность, что представляется очень важным.

5. Классификаторы на основе мер *коэффициент регрессии*, *добавочное значение, убеждение*, *фактор определенности* и *меры Клозгена*, превосходят классификатор, построенный на основе правил, представленных в [28], *по всем характеристикам*.

6. Наилучшее значение площади под ROC-кривой продемонстрировали классификаторы, построенные на основе коэффициента регрессии, добавочного значения и убеждения, фактора определенности и меры Клозгена.

Отметим, что классификатор *BayesNet*, построенный в инструментальной среде *Weka*, смог классифицировать все экземпляры и оказался наилучшим по показателям *среднеквадратичное отклонение, чувствительность, полнота* и *F-мера*. Однако, его построение заняло несоизмеримое большее время, и он уступил классификаторам на основе *добавочного значения*, *убеждения*, *коэффициента регрессии*, *фактора определенности* и *меры Клозгена* по показателям *коэффициент ложной тревоги, точность* и *площадь под ROC-кривой*.

Полученные результаты позволяют сделать вывод о том, что классификатор *BayesNet* не подходит *для работы с большими данными*, в отличие от алгоритмов классификации на основе правил, полученных с помощью мер *коэффициент регрессии*, *добавочное значение*, *убеждение*, *фактор определенности* и *меры Клозгена*.

3.4. Соревнование “Causality Challenge №1: Causation and Prediction” (метод 4)

Соревнование “*Causality Challenge №1: Causation and Prediction*” [38] проходило с 15 декабря 2007 г. по 30 апреля 2008 г. Участникам были предложены четыре задачи в различных прикладных областях. Целью соревнования было как можно точнее предсказать значение бинарной целевой переменной для тестовых выборок. Кроме того, участникам было предложено пред-

Табл. 1. Результаты анализа классификаторов (фрагмент)

	J	G	V	AV	K	F	R	CAR	Bayes-Net
RMSE	0.785	0.722	0.479	0.482	0.474	0.479	0.479	0.507	0.3659
AUC	0.527	0.5	0.799	0.799	0.792	0.797	0.798	0.6	0.745
Precision	0.658	0.629	0.835	0.836	0.83	0.835	0.835	0.761	0.81
FPR	0.331	0.476	0.177	0.169	0.193	0.176	0.175	0.366	0.326

ставить список переменных (признаков), которые они использовали в предсказании. Задачи были разработаны таким образом, что знание причинно-следственных связей между переменными и целевой переменной должно было повышать точность предсказания (классификации). Эффективность поиска причинных связей соответствовала оценке, которая показывала, насколько хорошо признаки, отобранные участниками, совпадают с Марковским покрытием для целевой переменной в тестовой выборке. Частью анализа результатов соревнования было исследование, насколько эта оценка коррелирует с точностью прогнозирования. Несмотря на то, что конкурс закончился в апреле 2008 года, платформа открыта для загрузки внеконкурсных результатов и просмотра всех результатов соревнования.

Для экспериментов с мерами связи, рассмотренными в данной работе, использован набор данных *CINA*, содержащий 16033 обучающих и 10000 тестовых примеров, каждый из которых описывается 132 бинарными признаками. Он был сформирован на основе набора данных *Adult*, имена атрибутов в котором были скрыты. К исходным 44 атрибутам было добавлено 88 искусственно созданных переменных, которые не являются причинами целевой переменной. В обучающих данных некоторые из этих искусственных переменных являются следствиями целевой переменной и/или других исходных переменных. Другие переменные не имеют ассоциативных связей с ними. Таким образом, некоторые из искусственных переменных могут коррелировать с целевой переменной в обучающих данных, но они не являются ее причинами.

Внеконкурсные результаты соревнования для набора данных *CINA0* были получены для *меры Клозгена*, *коэффициента регрессии*, *добавочного значения*, *убеждения* и *фактора определенности*. Для каждой из перечисленных мер были сформированы списки выявленных ими причин, а также построены классификаторы, как описано в разделе 3.3. Списки причин и результаты классификации загружены на сайт соревнования.

Для сравнения результатов участников соревнования использовались следующие оценки:

1. *Fscore* – оценка соответствия представленных участником причинно-следственных связей для набора данных реальным причинно-следственным связям. Список всех истинных причинно-следственных связей между признаками (переменными) и целевой переменной известен только организаторам. *Fscore* представляет собой площадь под *ROC*-кривой для представленной участником классификации “причинных” и “не причинных” признаков набора данных. При этом “не причинными” организаторы считают признаки, которые не относятся к Марковскому покрытию целевой переменной.

2. *Dscore* – оценка, соответствующая площади под *ROC*-кривой для классификации примеров обучающего набора.

3. *Tscore* – оценка, соответствующая площади под *ROC*-кривой для классификации примеров тестового набора.

Отметим, что для ранжирования участников в турнирной таблице соревнования использовалась только оценка *Tscore*. Но так как целью данной работы является выявление причинных связей в данных, то в рамках данного исследования более важной является оценка *Fscore.*

Внеконкурсные результаты соревнования для набора данных *CINA0* и перечисленных выше мер представлены в Табл. 2.

Табл. 2. Внеконкурсные результаты соревнования [39]

Мера	Поиск причин	Точность классификации	
	Fscore	Dscore	Tscore
Мера Клозгена	0.9276	0.9015	0.9045
Коэффициент регрессии	0.8893	0.9009	0.9035
Добавочное значение	0.8813	0.8903	0.8952
Убеждение	0.8665	0.9136	0.9136
Фактор определенности	0.8665	0.9118	0.9132

Отметим, что значения *Fscore* для всех пяти мер, анализируемых в работе, попали в список 5% лучших результатов. Результаты классификации (*Dscore* и *Tscore*) в списки лучших не попали, но целью было не построение наилучших классификаторов. Напомним, что в данной работе для оценки качества мер причинности использованы классификаторы, построенные по схеме простого голосования.

Заключение

В работе представлен краткий обзор численных мер оценки ассоциативных правил,

разработанных в статистике, социологии, машинном обучении и интеллектуальном анализе данных, и проведен сравнительный анализ нескольких их ключевых свойств, важных с позиций причинного анализа данных.

Приведены аксиомы для мер причинной связи [27], и из всего списка мер ассоциативной связи выделены те, которые потенциально могут быть использованы в качестве мер для оценки "силы" причинных связей. Эти меры детально исследованы экспериментально, в частности, для них

- проведено сравнение наборов правил, полученных с помощью выделенных мер, с наборами правил, которые получены другими исследователями, для некоторого общего набора данных;
- выполнено сравнение результатов ассоциативной классификации с помощью классификаторов, построенного на основе правил, полученных с помощью выделенных мер;
- получены внеконкурсные результаты соревнования "*Causality Challenge №1: Causation and Prediction*" для простых классификаторов, построенных с помощью наиболее перспективных мер.

На основании анализа экспериментальных результатов сделан вывод о том, что наиболее перспективными с точки зрения способности выявлять правила, представляющие причинные связи в данных, являются следующие меры:

1. *коэффициент регрессии;*
2. *мера Клозгена;*
3. *убеждение;*
4. *фактор уверенности.*

Напомним, что при анализе больших данных *критически важной* является *сложность* используемых алгоритмов. Отметим, что вычисление значений перечисленных мер можно выполнить за *один проход* по данным. Соответственно, подобный подход к оценке правил, которые потенциально могут отражать причинные связи между переменными и целевой переменной, очень хорошо подходит для задач анализа больших данных.

Литература

1. Fan J., Han F., Liu H. Challenges of Big Data Analysis // National Science Review. 2014. No. 1. spp. 293-314.
2. Bickel P. Discussion on the paper "Sure independence screening for ultrahigh dimensional feature space" by Fan and Lv // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008. No. 70(5). pp. 883–884.
3. Aliferis C.F., Statnikov A., et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation // Journal of Machine Learning Research. 2010. No. 11. pp. 171-234
4. Aliferis C.F., Statnikov A., et al. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions // Journal of Machine Learning Research. 2010. No. 11. pp. 235 – 299
5. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Representation and Reasoning Series (2nd printing ed.). San Francisco. California: Morgan Kaufmann. 1988
6. Witten I.H., Frank E., Hall M.A. Data Mining: Practical machine learning tools and techniques (3rd Edition). San Francisco. California: Morgan Kaufmann. 2011
7. Clark P., Boswell R. Rule induction with cn2: some recent improvements // Proceedings of the European Working Session on Learning EWSL-91, Porto, Portugal. 1991. pp.151-163.
8. Clark P, Brin S., Motwani R., Ullman J., Tsur S. Dynamic itemset counting and implication rules for market basket data // Proceedings of ACM-SIGMOD International Conference on Management of Data, Montreal, Canada. 1997. pp. 255-264
9. Silverstein C., Brin S., Motwani R., Ullman J., Scalable techniques for mining causal structures // Journal of Data Mining and Knowledge Discovery.2000. No. 4. pp. 163–192
10. Adamo J.-M. Data Mining for Association Rules and Sequential Patterns // Berlin: Springer. 2000
11. Han J., Kamber M. Data Mining: Concepts and Techniques, 2nd ed. (J. Gray, Ed.). San Francisco, California: Morgan Kaufmann. 2006
12. Agrawal R., Sricant R. Fast Algorithm for Mining Association rules // Proc. of the 20th Intern. Conference on Very Large Databases. Santiago, Chile. 1994. pp. 487-499
13. Agrawal R., Imielinski T., Swami A. Mining association rules between sets of items in large databases // Proceedings of ACM SIGMOD International Conf. on Management of Data. In P. Buneman, & S. Jajodia (eds.). 1993. pp. 207-216
14. Yafi E., Alam M.A., Biswas R. Development of subjective measures of interestingness: From unexpectedness to shocking // Proceedings of World Academy of Science, Engineering and Tech. No. 26. 2007. pp. 368-370
15. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1981. 302 с.
16. Mosteller F. Association and estimation in contingency tables // Journal of American Statistical Association. 1968. No. 63 (321). pp. 1-26
17. Lenca P., Vaillant B., Meyer P., Lallich S. Association Rule Interestingness Measures // Experimental and Theoretical Studies. Quality Measures in Data Mining. 2007. Vol. 43. pp. 51-76
18. Yule G.U. On the methods of measuring association between two attributes // J. R. Stat. Soc. 75. 1912. pp. 579-642

19. Tan P.N., Kumar V., Srivastava J. Selecting the right objective measure for association analysis // Joirnal of Information Systems - KDD. 2004. No. 4. pp. 293-313
20. Wikipedia.org: the free encyclopedia. Gini coefficient. URL: http://en.wikipedia.org/: http://en.wikipedia.org/wiki/Gini_coefficient (дата обращения 01.06.2015 г.)
21. Piatetsky-Shapiro G. Discovery, analysis and presentation of strong rules // G. Piatetsky-Shapiro, & W. Frawley (Eds.), Knowledge Discovery in Databases. Cambridge, MA: MIT Press, 1991. pp. 229-248
22. Tan P., Kumar V. Interestingness measures for association patterns: A perspective. Technical Report TR00-036 // Proceeedings of Workshop on Postprocessing in Machine Learning and Data, Mining University of Minnesota, Department of Computer Science. 2000
23. Sahar S., Mansour Y. An empirical evaluation of objective interestingness criteria // Proceeedings of SPIE Conference on Data Mining and Knowledge Discovery, Orlando, FL. 1999. pp. 63-74
24. Wikipedia.org: the free encyclopedia. Jaccard index. URL: http://en.wikipedia.org/wiki/: http://en.wikipedia.org/wiki/Jaccard_index (дата обращения 01.06.2015 г.)
25. Sebag M., Schoenauer M. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases // Proc. of the European Knowledge Acquisition Workshop EKAW'88. 1988. pp 28.1-28.20
26. Иоффе А.Я., Марков В.И., Петухов Г.Б. и др. Вероятностные методы в прикладной кибернетике: Учебное пособие. Под ред. Р.М. Юсупова. Л. 1976. 424 с.
27. Городецкий В.И., Самойлов В.В. Ассоциативный и причинный анализ и ассоциативные байесовские сети // Труды СПИИРАН. 2009. № 9. С. 13-65.
28. Li J., Le T.D, Liu L., Liu J., Jin Z., Sun B. Mining causal association rules // Proc. of The First IEEE ICDM Workshop on Causal Discovery (CD 2013). 2013. pp. 114-123.
29. UCI Machine Learning Repository. URL: http://archive.ics.uci.edu/ml/ (дата обращения 01.06.2015 г.)
30. Cooper G.F. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships // Journal of Data Mining and Knowledge Discovery. 1997. No. 1. pp. 203–224
31. Результаты экспериментов. URL: https://drive.google.com/folderview?id=0ByiklOTai_zZUUlIRjVJWVdwZTg&usp=sharing (дата обращения 01.06.2015 г.)
32. Wikipedia.org: the free encyclopedia. Confusion matrix. URL: http://en.wikipedia.org/wiki/Confusion_matrix (дата обращения 01.06.2015 г.)
33. Wikipedia.org: свободная энциклопедия. Среднеквадратичсекое отклонение. URL: https://ru.wikipedia.org/wiki/Среднеквадратичсекое_отклонение (дата обращения 01.06.2015 г.)
34. Wikipedia.org: свободная энциклопедия. ROC-кривая. URL: https://ru.wikipedia.org/wiki/ROC-кривая (дата обращения 01.06.2015 г.)
35. Wikipedia.org: the free encyclopedia. Precision and recall. URL: http://en.wikipedia.org/wiki/Precision_and_recall (дата обращения 01.06.2015 г.)
36. Wikipedia.org: свободная энциклопедия. Скользящий контроль. URL: http://machinelearning.ru/wiki/index.php?title=Скользящий_контроль (дата обращения 01.06.2015 г.)
37. Weka 3: Data Mining Software. URL: http://www.cs.waikato.ac.nz/ml/weka (дата обращения 01.06.2015 г.)
38. Causality Challenge №1: Causation and Prediction. URL: http://www.causality.inf.ethz.ch/ challenge.php (дата обращения 01.06.2015 г.)
39. Causality Challenge №1: Causation and Prediction. Результаты соревнования. URL: http://www.causality.inf.ethz.ch/challenge.php?page=results&ds=cina0 (дата обращения 01.06.2015 г.)

Тушканова Ольга Николаевна. Аспирант лаборатории интеллектуальных систем СПИИРАН. Окончила Южный федеральный университет (г. Ростов-на-Дону) в 2011 году. Автор 16 печатных работ. Область научных интересов: интеллектуальный анализ данных и извлечение знаний, многоагентные системы, рекомендующие системы, облачные технологии, онтологии. E-mail: tushkanova.on@gmail.com