ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 3** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 004.89 : 004.056

А.А. Грушо, М.И. Забежайло, А.А. Зацаринный, В.О. Писковский

О некоторых методах и технологиях искусственного интеллекта, используемых при защите облачных вычислений*

Представлен обзор применения технологий интеллектуального анализа данных в прикладных системах обеспечения информационной безопасности. Основное внимание уделено новому активно развивающемуся направлению – облачным вычислительным средам (в том числе – так называемым туманным вычислениям). Обсуждаются как текущее состояние, так и перспективные возможности использования моделей и методов искусственного интеллекта при решении задач информационной безопасности.

Ключевые слова: *искусственный интеллект, интеллектуальный анализ данных, облачные вычисления, информационная безопасность, математические модели и методы*

ВВЕДЕНИЕ

Со времени своего зарождения как самостоятельного направления исследований и разработок в 40-50-х гг. двадцатого века *искусственный интеллект* (ИИ) всегда ассоциировался с решением с помощью компьютеров таких трудных задач, которые требовали привлечения интеллекта человека (в том числе – знаний, умения рассуждать, выдвигать и верифицировать гипотезы, строить доказательства и опровержения и т.п.). Вряд ли кто-либо сегодня будет оспаривать наличие в области обеспечения защиты

*Работа выполнена при поддержке РФФИ (проект № 15-29-07981 офи-м).

данных и информационной безопасности целого ряда подобных трудных задач. А необходимость решать их, анализируя все более существенные объемы данных, причем делая это в условиях все более жестких ограничений по времени на оценку ситуации и принятие рациональных (обоснованных, корректных, а если возможно – оптимальных) решений, определяет все более привлекательные перспективы использования и в этой области новейших достижений – компьютерных систем интеллектуального анализа данных. Цель настоящей работы – дать беглый обзор текущего состояния и ближайших перспектив использования методов и технологий искусственного интеллекта в частном, однако критически значимом секторе ИТ-систем и решений – в задачах обеспечения информационной безопасности облачных вычислительных сред (ИБ ОВС).

ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ ОБЛАЧНЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕД: ПОПЫТКА ВЗГЛЯНУТЬ НА ЭТУ ПРОБЛЕМАТИКУ С РАБОЧЕГО МЕСТА ИБ-ЭКСПЕРТА

Облачные вычисления (cloud computing) – относительно новый, быстро развивающийся комплекс технологий, привлекательный открывающимися заманчивыми преимуществами (например, [1]) и, вместе с тем, обремененный новыми проблемами, в частности – в области обеспечения информационной безопасности (ИБ). Особая и все более критичная роль ИБ-проблематики зафиксирована на самом высоком государственном уровне [2].Сегодня для облачных вычислений характерны устойчиво высокие темпы развития. Блестящий пример взрывного развития этой проблематики – так называемые «туманные вычисления» (fog computing). Лидеры глобального ИТ-рынка объединили свои усилия в развитии этого класса информационных технологий [3], создав проблемно-ориентированное профессиональное сообщество [4-5] и сфокусировав его усилия в том числе и вокруг актуальных прикладных проблем (см., например, развитие так называемого интернета вещей – [6]).

Критически важными характеристиками обсуждаемой проблематики (cloud и в частности fog computing) являются быстро растущие объемы, а также актуальные скорости информационного обмена. В части ИБ для этого сектора информационных технологий определяющими стали все более жесткие требования к режиму процессно-реального времени в мониторинге (в том числе – анализе содержимого) трафика облачных сред и противодействии угрозам информационной безопасности.

Актуальные требования к объемам и скорости анализа данных при решении задач ИБ облачных вычислительных сред не оставляют каких-либо надежд на успех систем мониторинга и защиты, функционирующих в так называемом «ручном» режиме изучения данных и поддержки принятия соответствующих решений. Всеобщая «автоматизация» и применение «искусственного интеллекта» – популярная «мантра», активно продвигаемая сегодня и в обсуждаемой нами области – обеспечении ИБ облачных вычислительных сред [7-11]. Практически значимые прикладные результаты здесь связаны, в первую очередь, с использованием математических моделей, методов и прикладных программно-технических комплексов *интеллектуального анализа данных* (ИАД) – компьютерного анализа данных средствами интеллектуальных систем.

В обширном множестве (как полуформальных, так и формальных) вариантов уточнения понятий *интеллект*, *искусственный интеллект*, *компьютерная система искусственного интеллекта* и т.п. представляется полезным опереться на подробный обзор и детальный анализ имеющихся нескольких десятков точек зрения на обсуждаемую предметную область, которые были предприняты С. Леггом и М. Хаттером (Legg S., Hutter M – [12,13]). Суммируя анализируемые ими толкования, будем связывать обсуждаемую проблематику со *способностями вычислять* и *рассуждать*, *воспринимать* и *осознавать отношения* и *аналогии*, *обучаться* (на примерах), *хранить* и *искать информацию*, *использовать естественный язык*, *классифицировать* и *обобщать*, *адаптироваться к новым ситуациям* в процессе *решения трудных проблем* [12].

В отличие от обзора [14], где нас в первую очередь интересовали «академические» основания и возможности применения методов ИАД при обеспечении информационной безопасности облачных вычислительных сред, в настоящей работе мы попробуем взглянуть на рассматриваемую предметную область с точки зрения критически важного персонажа – *эксперта по информационной безопасности*. Весомых аргументов в пользу такого выбора можно предъявить немало, однако мы ограничимся лишь следующими двумя: – *особой ролью* такого эксперта (в том числе – необходимостью вести себя *рационально* и *ответственно* в быстро меняющейся среде) и необходимостью «вооружить» эксперта эффективными компьютерными инструментами ИАД, позволяющими *расширить* его *профессиональные возможности*.

Рассматривая особую роль эксперта, занятого мониторингом и противодействием угрозам ИБ в режиме «боевого дежурства», представляется целесообразным уточнить: Что такое *рациональное* поведение эксперта в рассматриваемой ситуации? Что ему следовало бы знать и чем (какими «инструментами» – знаниями, методами, …) ему следовало бы владеть, чтобы оказаться успешным в своей профессиональной деятельности? (С одной стороны, как уже было отмечено выше, объемы требующих мониторинга данных и скорость информационного обмена не оставляют надежд на успех «ручного» анализа текущих данных, т.е. без «автоматизации» не обойтись, с другой стороны – «*дознавательный*»[1] характер деятельности такого эксперта вряд ли может быть полностью формализован, чтобы успешно«перенести»этот тип деятельности в компьютерную систему в форме

[1] Ориентированный на *дознание*, *расследование*, *анализ фактов*, *выдвижение* и *опровержение (либо доказательство) версий*, на *обоснованность* (неоспоримость) принимаемых *решений* и т.п..

надежных программных «инструментов», способных эффективно работать как в диалоговом, так и автономном режиме). Подход, в рамках которого реализуется партнерское взаимодействие *ИБ-эксперт&компьютер,* не исключает полную автоматизацию принятия решении (т.е. работу компьютерной интеллектуальной системы в *автоматическом* режиме). Наоборот, по нашему мнению, полная автоматизация анализа и принятия решений машиной должна быть естественным развитием отношений *ИБ-эксперт&компьютер*.

В свою очередь, «партнерский» режим взаимодействия эксперта и используемых им систем искусственного интеллекта (систем *расширения его познавательных возможностей* и *поддержки рационального поведения*[2]) требует определенных усилий с обеих сторон – и от эксперта (например, в части формализации его действий), и от компьютерной системы (в части формирования ее архитектуры и организации функционирования). Как обеспечить возможности эксперту и интеллектуальной компьютерной системе «говорить» на одном языке? Как обеспечить «*бесшовные*» переходы проводимого в конкретной ситуации интеллектуального анализа данных «из головы и рук» ИБ-эксперта в «ассистирующую» ему интеллектуальную компьютерную систему и обратно? Проблем здесь не мало. Однако в первую очередь, следует обратить внимание на три важнейшие задачи, эффективное решение которых определяет успешность рассматриваемых интеллектуальных систем:

1) выбор адекватных *средств представления знаний* (дающих возможность и компьютерной интеллектуальной системе оперировать необходимыми эксперту знаниями);

2) выбор процедурных *средств порождения зависимостей* из данных (позволяющих использовать и компьютерную систему для извлечения скрытых зависимостей из постоянно накапливаемых эмпирических данных, причем делать это абсолютно «прозрачным» для ИБ-эксперта образом);

3) возможность вести *каузальный анализ* (выделять *причины* анализируемых событий – инцидентов, атак и т.п., что позволяет организовать надежное *противодействие* регистрируемым вредоносным активностям).

Успешное решение этих трех задач критически важно, так как фокусировка только на тех или иных «инструментальных» аспектах, например: новых возможностях так называемого глубокого машинного обучения, нейронных сетей нового поколения и т.п. – как правило, оставляет в «тени» интегральный конечный результат – *оценку степени безопасности защищаемого объекта*, за которую собственно и несет *ответственность* обсуждаемый нами ИБ-эксперт.

Таким образом, если посмотреть на работус ИБ-эксперта, так сказать, с «высоты птичьего полета», то несложно увидеть, что в целостной цепочке его действий существенная роль отводится не только тем или иным инструментам интеллектуального анализа конкретного набора данных, но и таким процессам:

- поиск (как правило, неформальный, управляемый лишь искусностью и опытом самого ИБ-эксперта) подходящих способов формализованного представления имеющихся эмпирических данных и знаний для их последующего компьютерного изучения (порождения зависимостей из накапливаемых данных),
- анализ «поведения» порождаемых зависимостей при расширении (пополнении новыми элементами) накапливаемых обучающих выборок описаний прецедентов.

Несложно убедиться, что это – уже больше чем традиционное машинное обучение (понимаемое как специализированная «аппроксимация» данных фиксированной выборки прецедентов соответствующим набором эмпирических зависимостей), это – своего рода *проблемно-ориентированное исследование*, которое требует адекватных средств компьютерной поддержки на каждом этапе – от выбора средств представления данных и знаний до анализа устойчивости («наследуемости» при расширении обучающих выборок описаний инцидентов), формируемых в результате его причинных зависимостей. Особо отметим, что *ответственность* за результаты такого исследования при всем разнообразии и могуществе используемых компьютерных систем (так называемого «искусственного интеллекта») остается на ИБ-эксперте.

Итак, (по возможности) следуя выбранному нами позиционированию (интеллектуальные системы и безопасность облачных вычислительных сред (ОВС) глазами ИБ-эксперта), обсудим более подробно:

- чем располагает ИБ-эксперт в рамках своей штатной деятельности, какие типы процедур исполняет и как ему в этой работе способны содействовать компьютерные системы интеллектуального анализа данных?
- в чем именно проявляются существующие угрозы ИБ ОВС?
- что собственно в них особенного (принципиально нового)?
- «канонические» представления регуляторов о проблематике ИБ в ОВС;
- актуальные представления о классификации сетевых атак;
- текущее состояние нормативной базы по проблематике ИБ ОВС;
- основные типы угроз ИБ в облачных вычислениях;
- что лидеры рынка считают приоритетами при формировании типовых подходов в противодействии угрозам ИБ ОВС;
- несколько примеров успешных проблемно-ориентированных технологий зашиты ОВС;
- некоторые *критически важные особенности* рассматриваемой предметной области – обеспечения ИБ ОВС.

О ТИПОВЫХ ДЕЙСТВИЯХ ИБ-ЭКСПЕРТА В РАМКАХ ЕГО ШТАТНОЙ ДЕЯТЕЛЬНОСТИ

Наряду со стандартным перечнем ИБ-активностей (например, [15, 16]), таких как обеспечение сохранности данных и шифрование, защита данных при передаче, аутентификация\защита паролем, а также разделение и изоляция пользователей (вместе с используемыми ими виртуальными машинами и

[2] Формулировка, предложенная проф. В.К. Финном.

сетями), штатный набор действий ИБ-эксперта должен быть пополнен следующим перечнем базовых активностей:

- поддержание в актуальном состоянии реестра (базы фактов и знаний – БФ&З) ранее идентифицированных ИБ-событий (вредоносных воздействий, атак, попыток проникновения, …);
- мониторинг (и документирование) поведения объекта защиты;
- фиксация аномалий (сбоев, флюктуаций, отклонений от заданных параметров SLA[3], а также заданных требований актуальной политики безопасности) и подробное документирование зафиксированных инцидентов;
- идентификация в зафиксированной аномалии целенаправленных вредоносных воздействий. Поиск и идентификация (в БФ&З) описаний аналогичных инцидентов (на базе ранее накопленного опыта как успешных, так и неуспешных действий адекватного противодействия идентифицированному вредоносному воздействию), а в случае отсутствия таковых – подробное изучение характеристик идентифицированного вредоносного воздействия и выбор адекватных средств противодействия.

Базовая роль и место интеллектуальной компьютерной системы – «ассистента» ИБ-эксперта в таких действиях:

- вести (поддерживать) проблемно-ориентированную БФ&З по идентифицированным ИБ-событиям, а также релевантным им сведениям;
- обеспечивать быстрый поиск в БФ&З инцидентов-аналогов, в том числе – «распознавание»(на сколько это возможно) развивающегося вредоносного воздействия на наиболее ранних стадиях, а также подбор рекомендаций (поддержка принятия решений) по организации в текущей ситуации эффективного противодействия угрозам информационной безопасности.

Фактически, здесь можно вести речь о компьютерных возможностях быстрого опознания (в процессе обучения на прецедентах) ранее изученных вредоносных воздействий и организации (с учетом накапливаемого опыта) эффективного противодействия их влиянию.

При выявлении нового (не описанного в БФ&З) вредоносного воздействия, основная роль компьютерной системы-ассистента – поддержать (прежде всего – *скоростью* и *объемами* анализа данных) «*дознавательные*» действия ИБ-эксперта, в том числе его возможности *вычислять* и *рассуждать* (например – использовать как достоверный логический вывод, так и те или иные эвристики, адекватные анализируемым эмпирическим данным), *воспринимать* и *осознавать отношения* и *аналогии* между совокупностями фактов, *обучаться* (на ранее изученных и описанных в БФ&З прецедентах), формировать абдуктивные объяснения выявляемым зависимостям и закономерностям, … *классифицировать* и *обобщать*, *адаптироваться к новым ситуациям* в процессе исчерпывающего анализа вновь идентифицированного инцидента. Именно здесь оказываются востребованными уже упоминавшиеся нами возможности «*бесшовной*» интеграции проблемно-ориентированных рассуждений ИБ-эксперта и поддерживающей его действия интеллектуальной компьютерной системы. Именно здесь оказываются критически важными и востребованными соответствующие функциональные возможности обсуждаемых компьютерных систем искусственного интеллекта. И чем более «тонкими» и «гибкими» окажутся характеристики закладываемых в такие системы математических моделей, методов и алгоритмов интеллектуального анализа данных, тем с большей уверенностью можно будет рассчитывать на успех исполнения ИБ-экспертом штатных обязанностей.

Рассуждая неформально, при поиске аналогов наблюдаемому аномальному поведению объекта защиты ИБ-эксперт сталкивается с тремя типами задач (порождающими затем соответствующие варианты «ответных действий»): идентификация объекта, идентификация действия\конфигурации («та или не та», уже известная или же новая) и идентификация целостности объекта\конфигурации. В свою очередь, процедурная «ниша» для использования интеллектуальных систем при решении этих трех типов задач может выглядеть следующим образом:

а) идентифицируем\классифицируем фиксируемые инциденты,

б) подбираем (с учетом постоянно накапливаемого опыта) адекватные средства противодействия вредоносным воздействиям каждого идентифицируемого типа,

в) выстраиваем систему быстрого обнаружения и классификации вновь возникающих инцидентов для того, чтобы

г) оперативно противодействовать таковым.

Инструментальным средством, реализующим представленную функциональность может быть проблемно-ориентированная (учитывающая специфику анализируемых данных – описаний прецедентов, характер скрытых в данных каузальных зависимостей и т.п.) интеллектуальная система обучения на прецедентах, распознавания и классификации вновь идентифицируемых инцидентов. Ее основная роль – расширение соответствующих «*дознавательных*» возможностей и поддержка принятия *рациональных* решений ИБ-экспертом в его штатной деятельности.

Следует отметить, что при всей очевидности предложенной нами схемы интеллектуального анализа данных и поддержки принятия решений успешная реализация ее на практике требует существенных усилий. Причем не только при выборе конкретных *математических моделей* и *методов* или же наделенных подходящими *выразительными возможностями* инструментальных программных систем. В целом ряде конкретных ситуаций трудности, с которыми столкнулись исполнители соответствующих прикладных проектов, в основе своей связаны с непониманием важности представленных нами проблем «*бесшовной*» интеграции активностей ИБ-экспертов и поддерживающих их штатную деятельность интеллектуальных компьютерных систем. (Пример ситуации такого типа дает проект разработки «Единой

[3] Service Level Agreement – соглашение об уровне сервиса

кросс-канальной системы выявления мошенничества», развернутый в ПАО Сбербанк [17], когда при полном понимании важности решаемых в Проекте задач, поддержке со стороны руководства Банка (см. в частности [9-11]) и несмотря на привлечение к организации Проекта компетентных консультантов, сроки его реализации уже неоднократно сдвигались, а к содержанию планируемых результатов у профессиональных экспертов имеется ряд существенных вопросов).

ОСНОВНЫЕ УГРОЗЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ОВС И ИХ ПРИНЦИПИАЛЬНЫЕ ОСОБЕННОСТИ

По данным [18-21] (см. также [1,Гл. 5]) наиболее существенными угрозами безопасности облачных вычислительных сред на текущий момент являются:

- злоупотребления и некорректное использование облачных сервисов\вычислений;
- архитектурные ограничения на доступ к инфраструктуре облака и динамический характер облачной вычислительной среды (контролируемость всех развернутых в ОВС виртуальных машин, удаленный доступ к вычислительным ресурсам, динамические изменения в текущем ландшафте задействованных виртуальных машин, уязвимость бездействующих виртуальных машин и т.п.);
- небезопасные интерфейсы (несанкционированное проникновение\взлом интерфейсов и API[4]);
- вредоносные инсайдеры (злоумышленники-инсайдеры);
- проблемы с распределенным использованием технологий (в том числе – с синхронизацией использования ресурсов облачной инфраструктуры различными пользователями);
- утечки или потери данных;
- взломы (уязвимости) аккаунтов или сервисов;
- неизвестные профили рисков (недостаточная осведомленность пользователей об особенностях облачных вычислений и возникающих при переходе в облачную вычислительную среду рисков и угроз);
- компрометация учетных записей и обход штатных процедур аутентификации;
- несанкционированное использование (кража) учетных записей;
- целенаправленные кибератаки на элементы инфраструктуры облака и вредоносные воздействия на архитектуру и механизмы управления облачными сервисами;
- DDoS-атаки[5];
- проникновение неавторизованных агентов – «чужаков» (Rogue Devices, Rogues)– в закрытую для них инфраструктуру облака;
- уязвимости беспроводного доступа (в том числе – специфичные для беспроводного доступа угрозы и атаки, включая разведку конфигурации сети и т.п.);
- уязвимости сетей и отдельных сетевых устройств (в том числе – угрозы утечки информации из проводного сегмента интегрированной сети);
- наличие недекларированных функциональных возможностей, а также вредоносных «закладок» в используемом программном и аппаратном обеспечении.

Как и ранее (см., например, [14]), особого внимания требует проблема идентификации и защиты периметра облачной вычислительной среды. Здесь в крупных центрах обработки данных (ЦОД) ситуацию в значительной мере позволяют контролировать подходы на основе программно-конфигурируемых сетей (Software Defined Networks – SDN [22-25]), однако в быстро развивающейся области «туманных вычислений» даже SDN-подход (разделение уровней собственно исполнения сетевого трафика и управления им через систему SDN-контроллеров) пока еще не дает удовлетворительных по надежности результатов.

Не менее деликатная (и критически значимая) проблематика информационной безопасности все более активно проявляется в области fog computing в связи с архитектурными особенностями используемых здесь вычислительных устройств. На сегодняшний день подавляющее большинство встроенных вычислителей, задействованных в области Internet of Things («интернета вещей»), не ориентированы на использование продвинутых (а в ряде случаев – вообще каких-либо) средств защиты и обеспечения информационной безопасности. Принимая во внимание возможности объединить в организованно действующее сообщество большое количество таких устройств (например, с целью организации массированной DDoS-атаки), злоумышленнику, вообще говоря, совсем не сложно организовать целенаправленное вредоносное воздействие на вычислители подобного типа, перехватив управление ими ради использования сообществ подобных устройств в своих целях. Так, после зафиксированной 21 октября 2016 г. DDoS-атаки на нескольких наиболее популярных в США сайтов, включая Twitter, Amazon, Netflix и Spotify (в которой оказались задействованными холодильники, термостаты, камеры наблюдения, видеорегистраторы и т.п.), профессиональные эксперты призвали правительство США разработать новые стандарты безопасности, которыми могли бы руководствоваться производители техники и оборудования, подключаемых к «интернету вещей» [26, 27]).

Если «взглянуть» на представленный перечень угроз, как мы и условились, с рабочего места ИБ-эксперта, то придется констатировать, что в ряде существенных случаев соответствующие угрозы формулируются как бы с позиции «*кто виноват?*», а реагировать на них необходимо, отвечая уже на другой вопрос – «*что делать*?». (Ситуация хорошо известная, например, из исследований по компьютерному моделированию игры в шахматы, где стартовая и финальная позиции сформулированы на вообще-то, разных языках: стартовая – позиционно, расстановкой фигур на конкретных полях игрового поля, а финальная – функционально, демонстрацией отсутствия допустимых правилами игры ходов в случае победы одной из сторон, или же демонстрацией отсутствия возможности поставить какую-либо из сторон в такое положение в случае ничьей). В рассматриваемой нами ситуации (при применении интеллектуальных

[4]Application Programming Interface – программный интерфейс приложений

[5] Distributed Denial of Service – DDoS.

компьютерных систем в задачах обеспечения информационной безопасности ОВС) в ряде случаев оказывается необходимым строить детальные соответствия – пары вида *тип угрозы =>варианты противодействия*, формируемые по результатам выполняемых ИБ-экспертом (во взаимодействии с проблемно-ориентированной компьютерной интеллектуальной системой) необходимых «*дознавательных*» процедур (в том числе – проблемно-ориентированного интеллектуального анализа накапливаемых эмпирических данных).

«КАНОНИЧЕСКИЕ» ПРЕДСТАВЛЕНИЯ МЕТОДОЛОГОВ И РЕГУЛЯТОРОВ О ПРОБЛЕМАТИКЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИВ ОВС

Наиболее полное представление о структуре, моделях и особенностях облачных вычислений, в том числе – комплексный взгляд на проблемы обеспечения информационной безопасности ОВС, по-видимому, можно найти в документах NIST (National Institute of Standards and Technology) – Национального института стандартов и технологий США (например, [28-30]).

В информационно-технологической концепции, предложенной экспертами NIST, облачные вычисления определяются пятью базовыми характеристиками [28]:

- самообслуживание по требованию (*on-demand self-service*);
- широкий выбор устройств доступа к сети (*broad network access*);
- объединение ресурсов (*resource pooling*);
- быстрое масштабирование («эластичность») облака (*rapid elasticity*);
- измеримость и учет сервисов, контроль ресурсов (*measured service*)

и имеют три модели обслуживания (сервис-модели):

- программное обеспечение как услуга (*Software as a Service – SaaS*);
- платформа как услуга (*Platform as a Service – PaaS*);
- инфраструктура как услуга (*Infrastructure as a Service – IaaS*);

а также четыре модели развертывания (deployment models):

- частное облако (*Private cloud*);
- облако сообщества пользователей (*Community cloud*);
- публичное облако (*Public cloud*);
- гибридное (интегрирующее те или иные из перечисленных выше компонентов и обеспечивающее балансировку нагрузки между ними) облако (*Hybrid cloud*).

С точки зрения аналитиков NIST, ключевыми аспектами безопасности облачных вычислений являются [29]:

- управление облачной вычислительной средой (*Governance*);
- способность оперировать в соответствии с заданными законами, регуляторными правилами, стандартами и спецификациями (*Compliance*);
- обеспечение доверенного режима использования облачной вычислительной среды (*Trust*);
- целостность архитектуры облачной среды (*Architecture*);
- управление идентификацией и доступом (*Identity and Access Management*);
- изоляция используемого в облачной среде программного обеспечения (*Software Isolation*);
- защита данных (*Data Protection*);
- доступность ресурсов и сервисов пользователям облачной вычислительной среды (*Availability*);
- адекватная реакция на возникающие\регистрируемые инциденты с безопасностью облачной вычислительной среды (*Incident Response*).

В качестве референс-модели систем обнаружения и предотвращения вторжений (Intrusion Detection and Prevention Systems – IDPS) аналитики NIST предлагают [31] рассматривать «инструментальные» программные системы, которые позволяют идентифицировать инциденты, накапливать детальную информацию о них, предпринимать усилия по их пресечению, формировать отчеты о них для администраторов информационной безопасности. IDPS используются как для выявления проблем с задействованными политиками безопасности, так и для того, чтобы документировать существующие угрозы и удерживать участников информационного обмена от нарушения политик безопасности. Типология IDPS-систем определяется особенностями событий, которые им приходится отслеживать, а также путей («каналов») реализации этих событий. Базовые типы таких систем – решения, ориентированные на контроль корректности конфигурирования компьютерных сетей, на беспроводные технологии, анализирующие поведение компьютерных сетей и, наконец, ориентированные на анализ функционирования хост-компьютеров.

Обширный и полезный обзор стандартов в области использования облачных вычислений предлагает Н.А. Храмцовская [32]. Рекомендации по управлению рисками информационной безопасности (в том числе – в соответствии с требованиями Системы менеджмента информационной безопасностью согласно ISO/IEC 27001) предлагает Международный стандарт ISO/IEC 27005:2011 [33]. Перспективы развития и совершенствования требований по защите информации при использовании облачных вычислительных сред представляет проект соответствующего стандарта [34].

ПОЗИЦИЯ ЛИДЕРОВ ИТ-РЫНКА ПО ОТНОШЕНИЮ К ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ОБЛАЧНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Среди лидеров глобального ИТ-рынка, по-видимому, бесспорным законодателем мод в обсуждаемой нами проблематике следует считать компанию Cisco. Предлагаемые ею комплексные программно-аппаратные решения в области обеспечения безопасности в ряде случаев оказываются стандартом *de facto* в соответствующем секторе ИТ-индустрии (достаточно яркий пример – актуальное решение по защите от DDoS-атак [35], которое имеет смысл рассматривать в контексте предлагаемых компанией классификации се-

тевых атак а также методов и средств противодействия их влиянию – см., например, [36, 37]).

Не менее уважаемый участник этого сегмента рынка – Intel. Результатом интеграции новейших специализированных системно-технических решений Intel с проблемно-ориентированными программными решениями McAfee стали новый подход и новая платформа для обеспечения безопасности облачных технологий – Cloud Security Platform [38]. Ключевые функциональные возможности этого решения, позволяющие обеспечивать безопасную циркуляцию информации между корпоративной сетью и облачной вычислительной средой,– это:

- двунаправленная защита http-трафика между корпоративной сетью и облаком (*Web Security*);
- защита мобильных устройств с помощью специализированного программного обеспечения (ПО), противодействующего вредоносному ПО (так называемого anti-mal ware), и проблемно-ориентированных корпоративных политик фильтрации web-трафика (*Mobile Security*);
- защита почты, интегрированная с системой защиты от утечек информации (*Data Leak Prevention system* – DLP) в исходящем трафике (*Email Security*);
- балансировщик нагрузки на уровне шлюза (*Web Services Security – Intel Expressway Service Gateway*).

При этом каждый из компонентов предлагаемого решения может быть реализован как физическое или виртуальное устройство, а также как облачный сервис (SaaS), причем независимо от способа реализации любого из других компонентов Cloud Security Platform.

Особый (как по динамике его развития, так и в плане критичности проблем, связанных с его функционированием) сектор глобального рынка информационной безопасности ОВС – это рынок уязвимостей (дефектов безопасности) широко используемых ИТ-решений (так называемых *vulnerabilities*).Здесь, наряду с источниками, фактически возведенными в статус классических (см., например, справочник [39]), все более весомую роль приобретают информационные ресурсы (см., например, [40-43]), основная задача которых – предоставлять актуальную информацию об уязвимостях популярных ИТ-решений (вплоть до так называемых «уязвимостях нулевого дня[6]»). Этот сегмент рынка ИБ-решений (в том числе – ИБ ОВС) стимулирует ключевых игроков (как разработчиков, так и «экспертов») постоянно держать на данном направлении и серьезные профессиональные компетенции, и значительные материальные, а также технологические ресурсы.

НЕСКОЛЬКО ПРИМЕРОВ УСПЕШНЫХ ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ ТЕХНОЛОГИЙ ЗАЩИТЫ ОБЛАЧНЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕД

Обсуждаемая проблематика, где одна из сторон атакует, а другая стремится надежно противодействовать таким атакам (и при этом обе стороны, вообще говоря, не ограничены в выборе путей и средств в своем целенаправленном поведении), – пример так называемой *открытой* (динамически меняющейся за счет пополнения новыми фактами, действиями, …) предметной области. Именно это ее свойство – *открытость*, по-видимому, не оставляет надежд на появление своего рода «*универсальной таблетки*» – универсального средства (подхода, инструмента, …), обеспечивающего надежное решение любых возникающих здесь задач.

Более естественным представляется подход, в рамках которого «однородные» по своей природе атакующие действия (вместе с эффективными средствами противодействия их влиянию) группируются в классы, «инвентаризацию» и сопровождение которых можно организовать средствами соответствующих интеллектуальных компьютерных систем. Причем интеллектуальность последних может проявляться в способности не только быстро диагностировать принадлежность того или иного наблюдаемого вредоносного воздействия к какому-либо из уже известных и в той или иной мере изученных классов (т.е. в способности «сэкономить» ИБ-эксперту время и технические ресурсы на анализ данных и принятие рациональных решений), но и поддержать (проводимые в рамках соответствующих «*дознавательных*» действий) рассуждения ИБ-эксперта автоматическим формированием на базе имеющихся эмпирических данных тех или иных гипотез (аналогий, обобщений, абдуктивных объяснений и т.п.) о характеристиках анализируемых вредоносных активностей. Таким образом, классическая тактика «разделяй и властвуй» вполне эффективна и в рассматриваемой нами ситуации. Некоторые примеры подходов и решений этого типа мы и обсудим далее.

Принимая во внимание de-facto существующую сегодня глубокую дифференциацию уязвимостей (см., например, уже упоминавшееся выше руководство [39], а также ресурсы ([44-46]) представляется естественным, что в практически значимых приложениях инструментарий противодействия вредоносным внешним воздействиям на ОВС также оказывается диверсифицированным по «однородным» направлениям (функциональным характеристикам угроз и вредоносных активностей, требующих привлечения такого инструментария для обеспечения штатного режима функционирования ОВС). Приведем несколько примеров.

В работах [47, 48] представлен подход к обнаружению сетевых атак, основанный на моделях классификации сетевого трафика. Показано, что на текущем этапе развития облачных вычислений, актуален ряд уязвимостей, связанных не только с классическими угрозами для распределенных систем, но и с принципиально новыми угрозами, порожденными *спецификой виртуализации*, а также наличием дополнительных уязвимых компонентов, реализующих предоставление облачных услуг. Например, компонент администрирования, компоненты организации коммутации сетевого трафика внутри ОВС, компоненты организации предоставления ресурсов для виртуальных машин и др. Авторами этих исследований разработана методика сбора данных и анализа трафика, проходящего в виртуализированной компь-

[6] Уязвимостях еще только планируемых к выводу на рынок решений.

ютерной сети облачной вычислительной среды (КС ОВС), а также метод обнаружения и блокировки сетевых атак на них, позволяющий работать в автоматическом режиме (без обращения к администратору). Предложенный метод использует проблемно-ориентированные средства интеллектуального анализа данных – переобучение моделей классификации трафика – и позволяет учитывать динамические изменения как структуры, так и объема легитимного трафика виртуализированной компьютерной сети ОВС. На базе предложенного метода реализована система защиты для облачной вычислительной платформы OpenStack.

Успешное применение развиваемых в работах [47, 48] методов интеллектуального анализа данных (в частности, машинного обучения, классификации и др.) требует эффективного использования аппаратных и системных возможностей облачных вычислений. В работах [49, 50] предлагаются подход, позволяющий реализовать подобный интеллектуальный анализ данных в режиме параллельных вычислений в облачной среде, формальная модель алгоритмов интеллектуального анализа данных на основе модели акторов, позволяющая описывать распараллеливание процесса вычислений как по данным, так и по задачам и не накладывающая ограничения на функции обработки данных, а также разработан метод конфигурирования узлов системы распределенного ИАД, позволяющий учитывать свойства обрабатываемых данных, свойства среды выполнения и алгоритма интеллектуального анализа данных. Работоспособность предлагаемых конструкций демонстрируется на примерах вычислительных ИАД-экспериментов в облаке. (Для их организации в репозитории виртуальных машин облака сохраняются заранее настроенные группы виртуальных машин, с предустановленным на них программным обеспечением, объединяемые в виртуальную сеть. Эти виртуальные машины запускаются из заранее сохраненных образов при выполнении интеллектуального анализа данных в соответствии с запросами пользователя).

Другой пример проблемно-ориентированного класса специализированных методов и систем противодействия угрозам информационной безопасности ОВС – технологии, ориентированные на уязвимости взаимодействия с памятью. Так, в работах [51, 52] рассмотрены модели и методы обнаружения и фильтрации шелкодов (shellcode) – вредоносного исполнимого кода (ВИК), эксплуатирующего *уязвимости работы с памятью* (т.е. уязвимости, которые возникают тогда, когда некоторый код в программе записывает в память больше данных, чем было предусмотрено разработчиком соответствующего приложения). Типичными примерами таких уязвимостей являются *переполнение стека*, *переполнение кучи*, а также некоторых других служебных структур. Авторы обсуждаемых работ создали модель распознавания объектов (ВИК), позволяющую обнаруживать в них набор специфических признаков и относить такие объекты к некоторым из заданных классов. Для решения этой задачи идентификации объекта создана проблемно-ориентированная программная система – инструментальная среда Demorpheus, работоспособность которой демонстрируется на примерах выявления шелкодов различных классов в процессе сканирования высокоскоростных каналов передачи данных, где таким путем существенно снижается вычислительная нагрузка на штатные системы обнаружения и противодействия вторжениям (IDS\IPS).

Практика показывает, что нередко пользователи облака могут сами устанавливать уязвимое программное обеспечение на своих виртуальных машинах (VM – virtual machine), что существенно снижает безопасность задействованной ОВС. По мнению авторов работы [53], критически важная задача состоит в том, чтобы эффективно обнаруживать уязвимости и атаки, создав систему реагирования при точном определении атак и сведя к минимуму последствия нарушения безопасности в облаке. К сожалению, в облачной среде атаки могут оказаться более эффективными, чем в классических компьютерных сетях, так как пользователи обычно разделяют общие вычислительные ресурсы. В частности, будучи подключенными через некие коммутирующие устройства, они могут делить оперативную память и файловую систему с потенциальным злоумышленником. Такая угроза реализуема и для методов виртуализации, и для виртуальных операционных систем, и для программного обеспечения, и для виртуальной сети. Все это делает привлекательным для злоумышленников атаки, цель которых – компрометировать множество виртуальных машин, используемых в облачных виртуальных средах.

В связи с массовым переходом на облачные вычисления, безопасность в облаке является одним из наиболее важных вопросов, которые стали предметом многих научных исследований и разработок в последние несколько лет. В частности, злоумышленники могут изучить уязвимые места облачной системы и скомпрометировать виртуальные машины для развертывания дальнейших крупномасштабных «Распределенных атак отказа в обслуживании» (Distributed Denial of Service – DDoS), которые обычно включают действия на ранней стадии, такие как многоступенчатая эксплуатация уязвимостей, низкочастотное сканирование, использование VM в качестве зомби, и, наконец, DDoS-атаки через зомби VM. В облачных системах, особенно, таких как «Инфраструктура как услуга» (Infrastructure as a Service– IaaS), обнаружить предварительные действия для зомби-атаки крайне сложно. Это связано именно с тем, что пользователи облака могут сами устанавливать уязвимые приложения. Для предотвращения компрометации VM, авторы обсуждаемого исследования предлагают многофазные распределенные детекторы уязвимостей, систему мер, а также механизм для выбора контрмер, который строит граф атак на основе аналитических моделей и реконфигурируемой виртуальной сети. Предлагаемая структура использует OF-протокол управления процессом обработки данных (протокол OpenFlow) для программного конфигурирования сети. С его помощью для улучшения обнаружения атак и смягчения их последствий реализуется мониторинг и контролируемая область поверх распределенных программируемых виртуальных коммутаторов. Реконфигурирование виртуальной сети используется также с целью обнаруживать и противодействовать использованию зом-

би VM. Предлагается также набор контрмер, которые могут быть применены к облачной виртуальной сети в зависимости от доступных методов и средств защиты. Здесь и перенаправление трафика, а также его изоляция и глубокий анализ, формирование правил фильтрации, изменение MAC-[7] и IP-адресов, блокировки портов, карантин, реконфигурация и изменение топологии сети.

Стратегия изменения конфигурации сети в основном связана с действиями на втором и третьем уровнях сетевой модели OSI[8]. На втором уровне отслеживаются виртуальные мосты и виртуальные локальные сети, являющиеся основным компонентом для соединения двух VM в облачной виртуальной сети. Виртуальный мост является объектом, который соединяет VIFs (виртуальные интерфейсы). VM на разных мостах изолированы на втором уровне. VIFs – на одном виртуальном мосту, но с разными тегами VLAN[9], не могут общаться друг с другом напрямую. Основываясь на изоляции второго уровня, можно организовать реконфигурацию сети для изоляции подозрительных VM. В результате этой контрмеры исключаются пути атак в графе, заставляя атакующего исследовать альтернативные пути. Третий уровень реконфигурации – это еще один способ исключить путь атаки. Через контроллер сети на каждом OFS (Open Flow Switch) или OVS (Open Virtual Switch) может быть модифицирована таблица потоков для изменения топологии сети.

Отмечается также, что подход реконфигурирования виртуальной сети на нижнем уровне имеет преимущество в том, что на верхних уровнях приложения будут испытывать минимальное воздействие. В частности, такой подход возможен только при использовании программного коммутатора для автоматической реконфигурации в динамической среде. Такая контрмера, как изоляция трафика может быть реализована на OVS и OFS, ограничивая и перенаправляя подозрительные потоки. Особо подчеркивается, что, когда в виртуальной сети обнаруживается подозрительная активность или сканирование портов, важно определить, являются ли эти действия злонамеренными. В такой ситуации изменение конфигурации сети заставит злоумышленника выполнить больше исследований, что в свою очередь сделает его поведение более заметным.

Одной из критически уязвимых «точек» нормального функционирования облачных вычислительных сред являются проблемы, которые могут возникнуть при конкурентной обработке данных облачными сервисами. Для иллюстрации этого положения обратимся к паре примеров.

Отлаженный годами автоматический процесс установки сложного программного обеспечения вдруг начал давать сбои («ошибаться») при использовании компьютеров следующего поколения. Наиболее вероятная причина – нарушение последовательности установки программных компонентов. На новой технике процесс установки компонента, для развёртывания которого обычно требовалось значительное время, в ряде случаев получил возможность обогнать другой процесс, который обычно, в силу конструктивных особенностей аппаратуры, к определённому моменту уже завершался. В результате – неминуемый сбой всякий раз при использовании более современной (быстрой) вычислительной техники. Очевидная проблема – рассинхронизация процессов. Предложенный временный способ обхода возникающих трудностей и решения проблемы – перенести некоторые файлы установки на диск. Объяснение этому вполне очевидное – разработчики не могли ни предусмотреть, ни эмулировать описанную ситуацию на аппаратуре предыдущего поколения. В результате компания переписала полностью установщик, а для временного обхода выпустила patch 1507768 [54].

Другой пример – выполнение системами управления базами данных операций, объединённых в группы транзакций, с целью сохранения определённых свойств. Например, в случае обработки финансовой информации, транзакции должны сохранять суммарное количество денег в обрабатываемых записях, т.е. набор объединённых в группы операций над данными не должен приводить к появлению или исчезновению денежных средств ни при каких обстоятельствах. В противном случае, это приведёт к использованию подобных обстоятельств с неблаговидными целями со всеми вытекающими последствиями, включая отказ от использования таким образом построенных информационных систем. Очевидный безопасный способ выполнения рассматриваемых действий – выстроить все операции, касающиеся хотя бы частично пересекающихся данных, в непротиворечивую последовательность. При этом конкурентный доступ к данным исключается, все операции, касающиеся данных, становятся в одну очередь, что приводит к неоправданному снижению производительности обработки или ситуациям, когда обработка становится невозможной (deadlock). Для выхода из сложившейся ситуации используются различные механизмы управления целостностью, отличаемые уровнем изоляции транзакций. В зависимости от установленного уровня изоляции допускаются условия, которые потенциально могут приводить к аномалиям при работе с данными, но тем не менее гарантируют их целостность в соответствии с выбранным уровнем ([55-60]). Один из способов обеспечить целостность транзакций при конкурентном использовании одного и тоже набора данных – это многоверсионность, использование копии данных, выполненных до начала конкурентной транзакции.

В случае автоматических изменений конфигурации сети рассинхронизация таких процессов приводит к нарушению политики безопасности в конкретный момент и увеличению поверхности атаки на распределённые системы. Различным способам защиты в ситуациях такого класса посвящено уже много работ. Одним из вариантов защиты здесь является контролируемое использование уровней изоляции для сетевых конфигураций и поддержки многоверсионности конфигураций [61-65].

[7]Media Access Control – управление доступом к среде.
[8]Open System Interconnection – 7-уровневая каноническая модель сетевого взаимодействия открытых систем.
[9]Virtual Local Area Network – виртуальная локальная сеть.

В работе [63] построен протокол для безопасной реконфигурации SDN-сетей на базе протокола OpeFlow. Найден полиномиальный алгоритм для поиска безопасных реконфигураций. В предложенном здесь алгоритме реконфигурация происходит в процессе последовательного применения правил обновления на отдельном коммутаторе.

Задача реконфигурации состоит в сохранении условий корректности при переходе от одной конфигурации к другой. Эти условия сформулированы в [61]. Однако следует учитывать, что изменения конфигурации не должны останавливать (или же существенным образом замедлять) в ОВС работу других штатных процессов.

Условия, обеспечивающие корректность такого реконфигурирования, сформулированы в работе [65], где также рассмотрен свой вариант подхода к решению задачи безопасной реконфигурации облачных вычислительных сред.

Контроль и управление информационными потоками в облачной среде – важные механизмы информационной безопасности [66]. Ключевая проблема в реализации таких механизмов – скорость коммутации пакетов в соответствии с заданными правилами. Критичная проблема в данном случае – идентификация текущего (анализируемого) объекта – заголовка пакета – в заданном множестве аналогичных объектов, а критичная «техническая» характеристика – это скорость поиска в соответствующей коммутационной таблице (КТ) заголовка пакета, поступающего на вход коммутатора. С формальной точки зрения – это скорость поиска предъявляемого КТ заголовка пакета (булевского вектора фиксированной длины) в множестве строк этой таблицы (в реальных приложениях содержащем десятки, а иногда – и сотни тысяч заголовков). Массовый приход пакетов и необходимость обрабатывать их в режиме процессно-реального времени (в том числе – в условиях внесения динамических изменений в текущее состояние КТ) делают эту область весьма чувствительной к любым возможностям сократить трудоемкость соответствующих процедур проверки.

В работе [67] предлагается алгоритм (использующий специальную процедурную технику формирования замыканий Галуа), позволяющий ускорить проверку встречаемости входных заголовках в КТ (опирающийся на быстрые процедуры проверки замкнутости множеств), а также приведены примеры использования предлагаемой техники в ряде задач управления потоками сообщений в компьютерных сетях и контроля содержимого пересылаемых пакетов.

Рассмотренные особенности предложенной технологии позволяют говорить о ней как о еще одном инструментальном основании для разработки активно развиваемых в настоящее время специализированных компьютерных решений-«акселераторов» [14], которые расширяют спектр возможностей для все более востребованного увеличения производительности проблемно-ориентированных прикладных систем интеллектуального анализа данных.

А.В.Волков в работе [68] приводит весьма убедительные аргументы, утверждая, что информация, находящаяся в облачных вычислительных средах, может подвергнуться атаке посредством уязвимостей как собственно облачной среды, так и системы управления облачными сервисами. Автор подчеркивает актуальность задачи *идентификации угроз* информационной безопасности ОВС и предлагает свой вариант анализа таких угроз, дополненный анализом существующих методов защиты и обеспечения информационной безопасности облачных вычислений. Проведенный им обзор и полученные результаты изучения наиболее распространенных угроз информационной безопасности ОВС позволяют аргументировать предлагаемую на их основе классификацию атак и методов защиты, а также дают убедительные аргументы для вывода о том, что наиболее опасными для нормального функционирования облачных вычислительных сред являются атаки на их виртуальную структуру.

В работе [69] предложена укрупненная архитектура экспериментального образца интеллектуальной системы защиты от компьютерных атак, охватывающая 5 основных блоков – подсистем:

1) сбора информации о циркулирующих в сети пакетах (например, по мнению авторов, это может быть известная система обнаружения и предупреждения вторжений Snort [70] или же подобная ей);

2) анализа ситуаций и идентификации компьютерных атак или вторжений, использующая информацию от программных сетевых сенсоров и базы знаний, в которую, по мнению авторов, должны входить модули корреляционного анализа и поддержки принятия решения на основе методов искусственного интеллекта;

3) адекватной реакции на идентифицированные атаки и попытки вторжения, обеспечивающая оперативную реакцию на возникшие угрозы (сетевые атаки, вредоносное программное обеспечение и т.п.) в соответствии с принятой политикой безопасности;

4) управления компонентами обнаружения и противодействия идентифицируемым атакам (в том числе – сетевым) и попыткам вторжения;

5) хранения накапливаемой информации об ИБ-событиях (инцидентах, сетевых атаках, …), эффективных средствах противодействия таковым (в том числе – истории как успешных, так и неуспешных действий по отражению компьютерных атак, по задействованным настройкам и политикам безопасности и т.п.), а также информационной поддержки принятия ИБ-экспертом рациональных решений.

Авторы [71] считают целесообразным использование в рамках подобной системы проблемно-ориентированных средств нейросетевого мониторинга сетевых атак на системы облачных вычислений с включением конвейерно-параллельной обработки. В качестве основных алгоритмов классификации они предлагают использовать многослойные сети прямого распространения, сети Кохонена, полиномиальное расстояние Евклида–Махаланобиса и метод опорных векторов. База данных сетевых атак KDD-99 [72] предложена в качестве обучающей выборки для машинного обучения интеллектуальной системы нейросетевого мониторинга и других механизмов выявления аномального поведения сети.

НЕСКОЛЬКО ДОПОЛНИТЕЛЬНЫХ ЗАМЕЧАНИЙ О КРИТИЧЕСКИ ВАЖНЫХ ОСОБЕННОСТЯХ РАССМАТРИВАЕМОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Переход к использованию облачных (в том числе – fog computing) вычислительных моделей и сред еще более расширяет и диверсифицирует перечень актуальных на текущий момент угроз информационной безопасности. Уже наблюдаемое сегодня разнообразие потенциально реализуемых злоумышленниками вредоносных действий не оставляет надежд на существование своего рода «*универсальной таблетки*» – универсального средства для отражения таких атак. (В обозначенном контексте представляется полезным задуматься как об ожиданиях и надеждах на возможности активно продвигаемых в настоящее время решений искусственного интеллекта «нового поколения» – так называемого глубокого обучения, нейросетевых технологий и др., так и, разумеется, о границах успешной применимости этих подходов).

С точки зрения ИБ-эксперта успех в рассматриваемой нами предметной области – проблематике обеспечения информационной безопасности ОВС – в значительной мере определяется возможностями вести анализ наблюдаемых фактов и порождать на их основе «*устойчивые*» (к *варьированию* как средств представления *данных*, так и средств *экспертизы* результатов) эмпирические зависимости и закономерности, которые способны характеризовать *причинность* в возникновении и развитии анализируемых вредоносных воздействий. При этом используемое здесь понятие устойчивости должно естественным образом учитывать и классическое толкование этого термина в смысле Ж.Адамара [73] и А.Н.Тихонова [74, 75]. Именно устойчивость (в обозначенном выше понимании) получаемых причинных зависимостей позволяет рассчитывать на эффективность основанного на их использовании противодействия идентифицируемым вредоносным активностям.

В стремлении сформировать подобные закономерности из накапливаемых эмпирических данных было бы крайне желательно при выборе и использовании соответствующих методов и технологий искусственного интеллекта не пасть жертвой своего рода «*подмены понятий*»: критически важно осознавать, что проблема вовсе не в том, чтобы построить оптимальную *интерполяцию* конкретной обучающей выборки описаний прецедентов (классическая постановка задачи машинного обучения – см., например, весьма популярные сегодня «заклинания» о глубинном машинном обучении, нейронных сетях и т.п.), а в том, чтобы *выявить устойчивую* (см. выше) *закономерность*, понимание и учет которой позволит эффективно противодействовать *варьируемым* попыткам реализовать целенаправленные вредоносные воздействия, идентифицируемым в *варьируемой* среде.

Особую роль в поиске таких закономерностей играет эффект BigData: данные доступны во все больших и больших объемах. Однако, на наш взгляд, ключевая проблема здесь вовсе не в том, что этих данных *очень много*. Критично, что такие постоянно растущие коллекции данных носят *открытый* характер (т.е. проблема на в *Big,* а в *Open*). Именно в пополняемых новыми фактами (а также другой релевантной информацией) обучающих выборках необходимо искать устойчивые закономерности. В свою очередь, устойчивость (наследуемость при расширении исходных обучающих выборок прецедентов) таких закономерностей должна сохраняться как при варьировании (например, за счет степеней детальности) выразительных возможностей используемых средств описания анализируемых прецедентов, так и при варьировании (по крайней мере в классах «*однородных*» – реализующих аналогичную «логику» обработки данных и принятия решений) средств экспертизы. (Сегодня понимание важности учета специфики обработки BigData в контексте динамически меняющихся условий анализа данных и принятия решений характерно уже не только для экспертов, так сказать, «*на передовой*», но и для ряда топ-менеджеров высшего звена [11]).

При этом полезно иметь ввиду, что решение проблемы *Big* можно искать достаточно хорошо известными, причем – постоянно развиваемыми и совершенствуемыми «техническими» подходами: здесь кое-что (а в ряде случаев – довольно многое) можно не только порождать в режиме параллельного счета, но и вычислять все быстрее и быстрее с помощью проблемно-ориентированных систем-«ускорителей» (специализированных программно-аппаратных решений-акселераторов) [14, 67].

Характеризуя критически важные элементы специфики обсуждаемой предметной области – обеспечения информационной безопасности ОВС, следует особо выделить проблему создания адекватных средств работы с малыми (статистически незначимыми) выборками. Полезно помнить, что в ряде случаев представляющие собой попытки взлома компьютерные атаки характеризуются весьма малым числом идентифицированных инцидентов. (И как же быть с нейронными сетями и глубоким обучением в такой ситуации? Ведь на практике ИБ-эксперт сталкивается не только с проблемой поиска ответа на «классический» вопрос «*кто виноват*?», но и вынужден определяться и по другому – не менее злободневному – вопросу «*что делать*?»).

Наконец, важно понимать, что поиск обсуждаемых эмпирических зависимостей и закономерностей – это уже не обычное машинное обучение. Это – больше, чем классическое машинное обучение. Это – проблемно-ориентированные исследования соответствующей (сегодня, как правило, характеризуемой как BigData и быстро растущей) предметной области, требующие разработки адекватных подходов в части «автоматизации», т.е. создания соответствующих математических моделей, методов и алгоритмов интеллектуального анализа данных вместе с реализующими их прикладными интеллектуальными компьютерными системами. При этом *интерактивный* режим классификации вновь идентифицируемых инцидентов ранее неизвестного типа требует особого внимания к проблеме «*бесшовной*» *интеграции* используемых ИБ-экспертом математических методов, моделей и алгоритмических процедур с соответствующим «когнитивным арсеналом» взаимодействующей с экспертом интеллектуальной

системы анализа данных и поддержки принятия рациональных решений. (По-видимому, здесь также вполне уместен вопрос: как же быть с нейронными сетями и «глубоким обучением» в этих ситуациях *интерактивного* режима взаимодействия эксперта и компьютерной системы?).

ЗАКЛЮЧЕНИЕ

Подведем итоги предпринятого обсуждения. Итак, проблематика информационной безопасности облачных вычислительных сред – актуальная и критически значимая область современных информационных технологий. Для этой области характерен ряд трудных (в том числе – плохо формализованных) задач, решение которых человеком-экспертом требует привлечения продвинутых *интеллектуальных* возможностей и навыков. Актуальная специфика задач обеспечения информационной безопасности ОВС (в том числе – постоянно растущие объемы требующей изучения информации, жесткие ограничения на режим процессно-реального времени анализа данных и принятия управленческих решений и др.) не оставляют возможности ИБ-эксперту исполнять свои штатные функции исключительно в «ручном режиме»: ему необходимы проблемно-ориентированные компьютерные инструменты интеллектуального анализа данных (ИАД), способные расширить его «*дознавательные*» возможности, а также обеспечить рациональные основания как для принимаемых им решений, так и выполняемых действий (по обеспечению информационной безопасности). Таким образом, актуальной оказывается задача «*бесшовной*» интеграции профессиональных действий ИБ-эксперта и поддерживающей его компьютерной системы: и эксперт, и взаимодействующая с ним интеллектуальная система должны «оперировать» в *однородной понятийной*, *алгоритмической* и *инструментальной среде* (использовать однородные средства представления данных и знаний, однородные формальные модели рассуждений и т.п.). От того, насколько согласованными будут эти «инструменты» интеллектуального анализа данных, используемые ИБ-экспертом и его «партнером» – компьютерной системой, критичным образом зависит надежность и эффективность соответствующих решений и средств обеспечения информационной безопасности.

Дополнительно заметим, что предложенный нами взгляд на использование интеллектуальных компьютерных систем – представление о них как о «партнерах» ИБ-эксперта (или же, следуя идее У.Р. Эшби [76], – *усилителях интеллектуальных возможностей* специалиста) – при оценке качества получаемых ими результатов естественным образом требует выхода за пределы уже ставших традиционными понятийных схем в стиле знаменитого *теста Тьюринга* [77]. В обсуждаемом нами случае интеллектуальная компьютерная система (даже если она работает в автоматическом, а не интерактивном режиме) вовсе не рассматривается как «*конкурент*» или «*заменитель*» ИБ-эксперта. Ее ключевая роль – в максимально возможной мере взять на себя «рутинную работу» по анализу данных, формированию и оцениванию характеристик альтернативных вариантов принятия необходимых решений по противодействию выявляемым ИБ-угрозам, предоставив ИБ-эксперту дополнительные ресурсы и свободу действий в оценке полноты, эффективности и надежности реализуемых мер по обеспечению безопасного функционирования объекта защиты.

СПИСОК ЛИТЕРАТУРЫ

1. Риз Дж. Облачные вычисления. – СПб.: «БХВ-Петербург», 2011. – 288 с.
2. Доктрина информационной безопасности РФ. – URL: http://www.scrf.gov.ru/documents/6/5.html
3. Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are. – White paper. – CISCO. – 2015. – 6 p. – URL: http://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf
4. Open Fog Consortium. – URL: http://www.openfogconsortium.org/
5. Open Fog Consortium. White Papers. – URL: https://www.openfogconsortium.org/resources/#white-papers
6. McKendrick J. Fog Computing: a New IoT Architecture? – RT Insight. – URL: https://www.rtinsights.com/what-is-fog-computing-open-consortium/
7. Буланов К. СМИ узнали о потере российским банком 100 млн руб. после кибератаки. – РБК, 01.12.2016. – URL: http://www.rbc.ru/finances/01/12/2016/583f7b749a79477669f0f916?from=main
8. Кошкаров А. Хакеры похитили у банков более 1 млрд руб. за полгода. - РБК, 19.07.2016. – URL: http://www.rbc.ru/finances/19/07/2016/578e222d9a794700144a8d80?from=main
9. Сбербанк: хакеры крадут больше всех. – Вести-экономика, 10.06.2016. – URL: http://www.vestifinance.ru/articles/71772
10. Горячева В. Сбербанк защитится on-line. – Коммерсант, 29.11.2016 года. – URL: http://www.kommersant.ru/doc/3155975?utm_source=kommersant&utm_medium=finance&utm_campaign=four
11. Греф Г.О. Эволюционируй или вымрешь. (Выступлением на 3 ежегодном конвенте выпускников бизнес-школы «Сколково», 24.05.2016 г.). – URL: http://trends.skolkovo.ru/2016/05/evolyutsioniruy-ili-vyimresh-german-gref-na-konvente-vyipusknikov-skolkovo/
12. Legg S., Hutter M. A Collection of Definitions of Intelligence. - Technical Report. - IDSIA-07-07. – June 15, 2007. – 12 p.– URL: https://arxiv.org/pdf/0706.3639.pdf
13. Legg S., Hutter M. Universal Intelligence: A Definition of Machine Intelligence // Minds & Machines. - Dec. 2007. – Vol.17 (4). – P. 391–444. – URL: http://www.vetta.org/documents/legg-hutter-2007-universal-intelligence.pdf
14. Грушо А.А., Забежайло М.И., Зацаринный А.А., Писковский В.О., Борохов С.В. О возможностях приложений интеллектуального анализа данных в задачах обеспечения информационной безопасности облачных сред // Научно-

техническая информация. Сер. 2. –2015. – № 11. – С. 1–11; Grusho A., Zabezhailo M., Zatsarinnyi A., Piskovskii V. On the Potential Applications of Data Mining for Information Security Provision of Cloud-Based Environments // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, № 6. – P.193-201.
15. Security Guidance Working Group. – URL: https://cloudsecurityalliance.org/group/security-guidance//
16. Решения по защите от угроз безопасности от компании CSA. – URL: https://habrahabr.ru/post/183168/
17. Выбор организаций на выполнение работ и оказанию услуг в рамках разработки «Единой кросс-канальной системы выявления мошенничества» для нужд ПАО Сбербанк. - ЗАКУПКА №31503176509. – URL: http://zakupki.gov.ru/223/purchase/public/purchase/info/common-info.html?noticeInfoId=5302472
18. Top Threats to Cloud Computing V1.0. - Cloud Security Alliance. – 2010. – 14 p. – URL: http://www.cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf
19. The Treacherous Twelve' Cloud Computing Top Threats in 2016. – Cloud Security Alliance. - Feb. 29, 2016. – URL: https://cloudsecurityalliance.org/group/top-threats/#_downloads
20. Top Threats to Cloud Computing. Update 2016-2017. - Cloud Security Alliance. – Dec. 4, 2016. – URL: https://cloudsecurityalliance.org/group/top-threats/#_downloads
21. Угрозы облачных вычислений и методы их защиты. – «PVSM.RU» - Новости ИТ. – URL: http://www.pvsm.ru/virtualizatsiya/36437/print/
22. Software Defined Networks. – URL: http://habrahabr.ru/company/muk/blog/251959/
23. ETSI: Network Function Virtualization. – URL: http://www.etsi.org/technologies-clusters/technologies/nfv
24. The Internet2 Community: Enabling the Future. – URL: http://www.internet2.edu/
25. GENI: Exploring Networks of the Future. – URL: https://www.geni.net
26. Миклашевская А. «Интернет вещей» может быть смертельно опасным. – Коммерсант. – 2016, 11 декабря. – URL: http://www.kommersant.ru/doc/3168591?utm_source=kommersant&utm_medium=tech&utm_campaign=four
27. SubCommTech and SubCMT Examine Recent Cyber Attacks. The Energy and Commerce Committee (114th Congress). – Press Release. - 2016, Nov.16. – URL: https://energycommerce.house.gov/news-center/press-releases/subcommtech-and-subcmt-examine-recent-cyber-attacks
28. Mell P., Grance T. The NIST Definition of Cloud Computing. – URL: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf
29. Jansen W., Grance T. Guidelines on Security and Privacy in Public Cloud Computing. – NIST SP 800–144. – URL: http://nvlpubs.nist.gov/ nistpubs/Legacy/SP/nistspecialpublication800-144.pdf
30. NIST Cloud Computing Security Reference Architecture. – URL: https://www.nist.gov/sites/default/files/documents/itl/cloud/NIST_SP-500-291_Version-2_2013_June18_FINAL.pdf
31. Scarfone K., Mell P. Guide to Intrusion Detection and Prevention Systems (IDPS). – NIST Special Publication 800-94. Revision 1. – Jul. 2012. – 111 p. – URL: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.259.5143&rep=rep1&type=pdf
32. Храмцовская Н.А. Стандарты и руководства по использованию облачных вычислений. – URL: http://www.eos.ru/upload/Information%20Management_13-03_12-21.pdf
33. ISO/IEC 27040:2015(en). Information technology – Security techniques – Storage security. – URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:27040:ed-1:v1:en
34. Проект ГОСТ РХХХХХ–20ХХ Защита информации. Требования по защите информации, обрабатываемой с использованием технологий «облачных вычислений». Общие положения. – ТЕХЭКСПЕРТ. – URL: http://docs.cntd.ru/ document/ 1200102839
35. Cisco ASR 9000 DDoS Protection Solution. – URL: http://www.cisco.com/c/en/us/solutions/service-provider/service-provider-security-solutions/asr-9000-ddos-protection-solution.html
36. Кадер М. Сетевые атаки. – URL: http://lagman-join.narod.ru/spy/CNEWS/cisco_attacks.html
37. Классификация сетевых атак. – CISCO. – 20.06.2005. – URL: http://www.internet-technologies.ru/articles/article_237.html
38. Intel Cloud Security Platform. – URL: http://itnews.com.ua/60064.html
39. Anley C., Heasman J., Linder F., Richarte G. The Shellcoder's Handbook. Discovering and Exploiting Security Holes. – Indianapolis: Wiley Publ.. – 2007. – xxiv+718 p.
40. Vulnerabilities in remote desktop could allow remote code execution. – Microsoft Security TechCenter. Microsoft security bulletin ms12- 020 – critical. – URL: http://technet.microsoft.com/enus/security/bulletin/ms12-020.
41. Vulnerability in oracle java runtime environment could allow remote code execution. - Multi-State Information Sharing and Analysis Center. – URL: https://msisac.cisecurity.org/advisories/2013/2013-041.cfm
42. CVE details. Security vulnerabilities published in 2013. – URL: http://www.cvedetails.com/ vulnerability-list.php
43. Rashid F.Y. The dirty dozen: 12 cloud security threats. – InfoWorld (IDG). – 2016, Mar. 11. – URL: http://www.infoworld.com/article/3041078/security/ the-dirty-dozen-12-cloud-security-threats.html
44. World's most used penetration testing software. – Rapid7's Metasploit. – URL: www.metasploit.com
45. Rapid7: Accelerating IT Insight. – URL: http://www.rapid7.com/company/index.php
46. Sun E., Haley P. 2015 Incident Detection and Response Survey Results. – Boston (MA): Rapid7. – URL: https://www.rapid7.com/company/resources/ incident-detection-response-survey.php

47. Борисенко К.А. Методы и модель организации защиты виртуализированных компьютерных сетей распределенных облачных вычислительных сред от сетевых атак. – Дисс. на соиск. уч. ст. канд.тех.наук. – СПб.: ЛЭТИ. – 2016. – 173 с. – URL: http://www.eltech.ru/assets/files/nauka/dissertacii/2016/Borisenko/Dissertaciya-Borisenko.pdf
48. Borisenko K., Rukavitsyn A., Gurtov A., Shorov A. Detecting the Origin of DDoS Attacks in OpenStack Cloud Platform Using Data Mining Techniques // Internet of Things, Smart Spaces, and Next generation Networks and Systems. – 2016. – P. 303-315.
49. Kholod I., Petuhov I., Kapustin N. Creation of Data Mining Cloud Service on the Actor Mode // Internet of Things, Smart Spaces, and Next Generation Networks and Systems. LNCS. – 2015. – №9247. – P. 585-599.
50. Петухов И.В.Представление алгоритмов интеллектуального анализа данных и их реализация в распределенных средах на основе модели акторов. – Дисс. на соиск. уч. ст. канд. тех. наук. – СПб.: ЛЭТИ. – 2016. – 116 с. – URL: http://www.eltech.ru/assets/files/nauka/dissertacii/2016/Petuhov/Dissertaciya_Petuhov.pdf
51. Гайворонская С.А. Исследование методов обнаружения шеллкодов в высокоскоростных каналах передачи данных. – Дисс. на соиск. уч. ст. канд.тех.наук. – М.: МГУ. – 2014. – 133 с. – URL: https://cs.msu.ru/sites/cmc/files/theses/gaivoronski-diss.pdf
52. Gaivoronski S., Gamayunov D. Hide and seek: Worms digging at the internet backbones and edges // Proceedings of the 7th Spring/Summer Young Researchers' Colloquium on Software Engineering (SYRCoSE 2013). – Kazan (Russia): National Research Technical University, 2013. – P. 94–107.
53. Крамаров Л.С., Бабенко Л.К. Обнаружение сетевых атак и выбор контрмер в облачных системах // Известия ЮФУ. Технические науки. – 2013. – №12 (149). – С. 94-101.
54. Varghese G. Resolving Oracle 8i problems with changes in hardware // Papers from the Proceedings of the 17th NACCQ 2004 (Edited by Samuel Mann and Tony Clear). – P. 452-454.
55. Revilak S., O'Neil P., O'NeilE.Precisely Serializable Snapshot Isolation // Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. – IEEE, 2011. – P. 482–493.
56. Eswaran P., Gray J., Lorie R., Traiger I. The Notions of Consistency and Predicate Locks in a Database System // Communications of the ACM. – 1976. – Vol. 19,№11. – P. 624–633.
57. Bernstein P.A., Hadzilacos V., Goodman N. Concurrency Control and Recovery in Database Systems. – Reading, MA, US: Addison-Wesley, 1987. – 264 p.
58. O'Neil P. Database: Principles, Programming, Performance. – San-Francisco, CA, US: Morgan Kaufmann Publ., 1994. – 860 p.
59. Конолли Т., Бегг К., Страчан А. Базы данных: проектирование, реализация и сопровождение. Теория и практика / пер. с англ. – М.: Вильямс, 2001. – 1120 с.; Conolly T. M., Begg C.E. Data base Systems. A Practical Approach to Design, Implementation, and Management. 2nd ed. – Reading, MA, US: Addison-Wesley, 1999. – 1093 p.
60. Berenson H., Bernstein P.., Gray J., Melton J., O'Nei E., O'Neil P. A Critique of ANSI SQL Isolation Levels. - Proceedings of the 1995 ACM SIGMOD international conference on Management of data. – New York: ACM, 1995. – P. 1–10.
61. Reitblatt R., Foster N., Rexford J., Walker D. Consistent Updates for Software-Defined Networks: Change You Can Believe In! - HotNets '11. - 2011, Nov. 14-15. - №7. - P. 1-6.
62. Reitblatt M., Foster N., Rexford J., Schlesinger C., Walker D. Abstractions for network update // Proc. Assoc. for Compt. Mach. SIGCOMM Conf. – 2012. – P. 323-334.
63. McGeer R. A correct, zero-overhead protocol for network updates // Proc. of the 2-nd ACM SIGCOMM workshop on Hot topics in software defined networking (2013). – 2013. – P. 161-162. – URL: http://conferences.sigcomm.org/sigcomm/2013/papers/hotsdn/p161.pdf
64. McGeer R. Declarative Verifiable SDI Specifications // IEEE Security and Privacy Workshops. – 2016. – P. 198-203. – URL: http://spw16.langsec.org/papers/mcgeer-verifiable-sdi-specs.pdf.
65. Грушо А.А., Забежайло М.И., Зацаринный А.А., Писковский В.О. Безопасная автоматическая реконфигурация облачных вычислительных сред // Системы и средства информатики.– 2016. – Т. 26, № 3. – С. 83–92.
66. Грушо А.А., Забежайло М.И., Зацаринный А.А. Контроль и управление информационными потоками в облачной среде // Информатика и ее применения. – 2015. – Т. 9, № 4. – С. 95–101.
67. Грушо А.А., Забежайло М.И., Зацаринный А.А. Об одном способе сокращения вычислений при формировании замыканий Галуа // Информатика и ее применения.– 2016. – Т. 10, № 4. – С. 97-106.
68. Волков В.А. Анализ угроз и методов защиты облачных сервисов // Молодий вчений. – 2015. – № 12(27). – С.38-43. – URL: http://molodyvcheny.in.ua/files/journal/2015/12/05.pdf
69. Емельянова Ю.Г., Фраленко В.П. Анализ проблем и перспективы создания интеллектуальной системы обнаружения и предотвращения сетевых атак на облачные вычисления // Программные системы: теория и приложения. – 2011. – № 4(8). – С. 17–31.
70. Snort. – URL: https://www.snort.org
71. Талалаев А.А., Тищенко И.П., Хачумов В.М., Фраленко В.П. Разработка нейросетевого модуля мониторинга аномальной сетевой

активности // Нейрокомпьютеры: разработка и применение. – 2011. – № 7. – С. 32–38.
72. Fifth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. – URL: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
73. Hadamard J. Sur les problèmes aux dérivéespartielles et leur signification physique. – 1902. – P. 49–52.
74. Тихонов А. Н. Об устойчивости обратных задач // Доклад АН СССР. – 1943. – Т. 39, № 5. – С. 195-198.
75. Тихонов А.Н. Корректные и некорректные задачи // Большая советская энциклопедия / гл. ред. А. М. Прохоров. 3-е изд. – М.: Советская энциклопедия, 1969–1978.
76. Эшби У.Р. Введение в кибернетику. – М.: Иностранная литература, 1959. – 432 с. – URL: http://pcp.vub.ac.be/ASHBBOOK.html
77. Turing A.M. Computing Machinery and Intelligence. - *Mind*. - 59 (Oct. 1950). – P. 433-460.

Материал поступил в редакцию 12.01.17.

Сведения об авторах

ГРУШО Александр Александрович – доктор физико-математических наук, профессор, заведующий лабораторией Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), Москва
e-mail: grusho@yandex.ru

ЗАБЕЖАЙЛО Михаил Иванович – доктор физико-математических наук, заведующий лабораторией Института проблем информатики ФИЦ ИУ РАН
e-mail: m.zabezhailo@yandex.ru

ЗАЦАРИННЫЙ Александр Алексеевич – доктор технических наук, профессор, заместитель директора Института проблем информатики ФИЦ ИУ РАН
e-mail: alex250451@mail.ru

ПИСКОВСКИЙ Виктор Олегович – кандидат физико-математических наук, ведущий специалист Центра прикладных исследований компьютерных сетей, Москва
e-mail: vpvp80@yandex.ru

УДК [004.89 : 510.64] : 616.37

М.А. Агафонов, О.П. Шестерникова, Л.В. Винокурова, Е.С. Панкратова, В.К.Финн

О принципах и логических средствах, реализуемых в интеллектуальной системе для гастроэнтерологии

Представлены принципы и логические средства компьютерной интеллектуальной системы, которая решает важную практическую задачу: обнаруживает причины развития сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом, а также эта система способна порождать причины отсутствия развития сахарного диабета у упомянутых больных.

Факты, исследуемые в этой системе, представлены посредством историй болезней, описанных на специально созданном формальном языке (его словарь содержится в конце статьи).

Отметим, что применение процедур выявления причин изучаемых эффектов и обнаружения их сохранения при расширениях массивов фактов требует разделения фактов на позитивные ((+)-факты) и негативные ((–)-факты).

В нашем исследовании позитивным фактом является наличие развития сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом, а его отсутствие у таких больных – отрицательным фактом.

Ключевые слова: *ДСМ-метод, эмпирические закономерности, компьютерная интеллектуальная система, доказательная медицина, панкреатит, диабет*

§1. ДСМ-МЕТОД АВТОМАТИЗИРОВАННОЙ ПОДДЕРЖКИ НАУЧНЫХ ИССЛЕДОВАНИЙ И ДСМ-РАССУЖДЕНИЯ

Компьютерные интеллектуальные системы (ИС) являются современным средством анализа данных для выявления причин изучаемых эффектов, которые используются также для объяснения полученных результатов. Применение интеллектуальных систем в медицине для анализа клинических данных и решения диагностических задач является примером использования информационных технологий (IT) для обнаружения патогенетических механизмов заболеваний и формирования групп риска.

Возможность получения новых знаний из расширяющихся массивов клинических данных (баз фактов) является полезным средством применения интеллектуальных систем для обнаружения эмпирических закономерностей, содержащихся в данных.

В настоящей работе использовалась компьютерная интеллектуальная система, реализующая ДСМ-метод автоматизированной поддержки научных исследований (ИС-ДСМ, ДСМ-метод АПНИ).

ДСМ-метод АПНИ имеет следующие компоненты: условия применимости, ДСМ-рассуждения, представление знаний в виде открытых теорий (множеств утверждений, существенным образом характеризующих изучаемое явление), организация процедур индуктивных выводов посредством их упорядочения по «логической силе» (выводимости одной из другой), средства обнаружения эмпирических закономерностей (ЭЗК), интеллектуальные системы типа ДСМ (ИС-ДСМ).

Условиями применимости ДСМ-метода АПНИ являются: 1) возможность установления сходства фактов (примеров); 2) наличие в базе фактов как **позитивных**, обладающих изучаемым эффектом, так и **негативных** фактов, им не обладающих, а также существование в базе фактов **позитивных и негативных** причин этих эффектов. Позитивные причины эффектов являются условиями, вынуждающими их проявления, а негативные причины, соответственно, являются условиями, вынуждающими отсутствие изучаемых эффектов. Позитивными причинами в клинических исследованиях могут быть как патогенетические механизмы заболеваний, так и факторы риска (поведенческие активности и влияния внешней среды).

Существенно отметить, что ДСМ-метод содержит возможности распознавания различных **структурно выраженных** сходств, могущими быть комбинациями различных факторов, представляющих причины эффектов (патогенетических механизмов и факторов

риска) [1]. Следовательно, ДСМ-метод обладает возможностями обнаружения **множественности** причин изучаемых эффектов.

Извлечение позитивных причин ((+)-причин) и негативных причин ((–)-причин) из массивов клинических данных посредством ДСМ-метода АПНИ возможно в силу специфического строения ДСМ-рассуждений, реализующих **качественный** анализ данных и предсказание изучаемых эффектов на основе аргументированных обобщений расширяющихся (пополняемых) баз фактов (экспериментальных массивов накопленных клинических данных).

ДСМ-рассуждение образовано тремя процедурами – индукцией, аналогией и абдукцией. Посредством процедуры индукции распознаются сходства (+)-фактов и (–)- фактов, соответственно. Обнаруженные сходства используются для порождения причин наличия эффекта ((+)-причин) и причин отсутствия эффекта ((–)-причин).

Обнаруженные гипотезы о (±)-причинах изучаемого эффекта образуют фрагмент порождаемой в автоматическом режиме базы знаний интеллектуальной системы (ИС-ДСМ). Процедура индукции формализуется посредством специальных правил правдоподобного (индуктивного вывода). Результатами применения правил индуктивного вывода являются (±)-гипотезы о причинах изучаемых эффектов. Эти гипотезы используются для предсказания высказываний, предложенных на прогноз. В настоящем исследовании на прогноз предлагаются истории болезней с диагнозом «хронический панкреатит», относительно которых должно быть получено предсказание о развитии сахарного диабета или о невозможности такого развития.

После последовательного применения правил индуктивного вывода (индукции) и правил вывода по аналогии порожденные гипотезы принимаются посредством объяснения начального состояния базы фактов (БФ) полученными гипотезами о (±)-причинах. Эта процедура называется **абдукцией**.

Схема принятия гипотез посредством абдуктивного объяснения базы фактов

(1) БФ – множество (±)-фактов

(2) Н – множество гипотез $\left(H = H_1 \cup H_2\right)$,

где H_2 – гипотезы о (±)-причинах, а H_1 – гипотезы о предсказаниях

(3) $E\left(H_2, БФ\right)$ – H_2 объясняют БФ

Все гипотезы из Н – правдоподобны (принимаются)

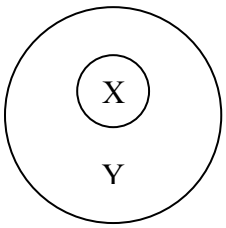
Пусть h^+ принадлежит H_2, а «C обладает Q^+» (+) – факт из $БФ^+$ (все (+)-примеры из БФ), тогда h^+ объясняет «C обладает Q^+», если h^+ содержится в С, где С – история болезни, а h^+ есть подмножество характеристик из С таких, что они обнаружены в сходных с С историях болезней, содержащих эффект Q^+ – развитие сахарного диабета у больных хроническим панкреатитом. Аналогичное имеет место для H_2, где h^- – гипотеза о (–)-причине, характеризующей условия отсутствия развития сахарного диабета панкреатогенного генеза.

Сформулированные условия определяют предикат $E\left(H_2, БФ\right)$ – «гипотезы о (±)-причинах» объясняют базу фактов (историй болезней из БФ).

База фактов содержит истории болезней С с двумя диагнозами, соответственно: «хронический панкреатит и последующее развитие сахарного диабета (Q^+)» (это (+)-пример), «хронический панкреатит и отсутствие развития сахарного диабета или сахарный диабет непанкреатогенного генеза (Q^-)» (это (–)-пример). Кроме того, БФ содержит множество историй болезней без установленного диагноза (Q^τ) (это (τ)-пример, где τ – обозначение неопределенности).

Вводятся обозначения для оценок историй болезней (они называются истинностными значениями): 1 (фактическая истина), -1 (фактическая ложь), τ (неопределенность), 0 (фактическое противоречие). Символы "∪" и "∩" обозначают, соответственно, объединение и пересечение множеств, а символ "ø" есть пустое множество, не имеющее элементов, тогда база фактов, представляющая истории болезней, есть объединение трех составляющих её множеств (подмножеств): $БФ = БФ^+ \cup БФ^- \cup БФ^\tau$ таких, что они не пересекаются (не имеют общих элементов, $БФ^+ \cap БФ^- = ø$, $БФ^+ \cap БФ^\tau = ø$ и т.д.).

Посредством $\subseteq$ обозначают отношения включения множества X в множество $Y\,(X \subseteq Y)$, что представимо кругами Эйлера.



Очевидно, что $БФ^\sigma \subseteq БФ$, где $\sigma = +, -, \tau$; а если $X \subseteq Y$ и $Y \subseteq X$, то $X = Y$. Если $X \subseteq Y$, то говорят, что X – подмножество Y. Если $X \neq Y$ и $X \subseteq Y$, то это обозначают $X \subset Y$ (строгое включение подмножества X в Y).

Обозначим также посредством |X| – число элементов множества X, тогда, $|БФ^\sigma|$, где σ = +, –, τ, есть число элементов $БФ^\sigma$ (т.е. историй болезней, являющихся (σ)-примерами). Пусть $|БФ^\tau| = m_0$, где элементы $БФ^\tau$ те истории болезней, относительно которых надо получить ответ на вопрос: будет ли развитие сахарного диабета панкреатогенного генеза у данного больного или нет?

Пусть $\tilde{БФ}^\sigma$, где $\sigma = +, -$, есть часть базы фактов БФ такая, что каждый её элемент (история болезни) имеет объяснение, т. е. имеет место $E(\tilde{БФ}^\sigma, H_2)$. Это означает, что каждая история болезни из $\tilde{БФ}^\sigma$ С

имеет h^σ такое, что $h^\sigma \subset C$, а h^σ объясняет факт «С имеет Q^σ», где $\sigma = +,-$.

Очевидно, что $\tilde{Б}\tilde{Ф}^\sigma \subset БФ$. Определяется функция $\rho^\sigma = \frac{|\tilde{Б}\tilde{Ф}^\sigma|}{|БФ^\sigma|}$ (где $\sigma = +,-,$), которая выражает степень объяснения (σ)-примеров из БФ (степень абдуктивного объяснения (σ)-примеров).

Если $\rho^\sigma = 1$, что означает, что все (σ)-факты из БФ имеют объяснение, то для БФ выполняется важное условие: все (σ)-факты имеют причину h^σ, где $\sigma = +,-$. Это означает, что для всех историй болезней C с диагнозом «хронический панкреатит с последующим развитием сахарного диабета (Q^+)» обнаружена h^+, являющаяся причиной Q^+ (аналогичное имеет место для h^- и Q^-). Очевидно, что $h^\sigma \subset C$, С – соответствующая история болезни, а $\sigma = +,-$.

Если $\rho^\sigma = 1$, то будем говорить, что для БФ имеет место аксиома каузальной полноты $\left(АКП^\sigma\right)$: в каждом (σ)-факте («С – история болезни с исходом Q^σ») С содержит причину h^σ для Q^σ, где $\sigma = +,-$ ((+) – для случая развития сахарного диабета панкреатогенного генеза, (–) – в противном случае).

Заметим, что $АКП^\sigma$ выполняется далеко не всегда, т.е. возможно, что $\rho^\sigma \le 1$. Тогда следует ввести порог для принятия порожденных гипотез посредством ослабления процедуры абдуктивного принятия гипотез, а именно: потребуем, чтобы $\rho^\sigma \ge \overline{\rho}^\sigma$, где $\overline{\rho}^\sigma$ – вводимый допустимый порог для принятия гипотез (обычно в задачах распознавания образов полагают, чтобы $0{,}8 \le \rho^\sigma \le 1$, т. е. $\overline{\rho}^\sigma = 0{,}8)$.

Если ρ^σ <1, то следует рассмотреть последовательность расширяемых БФ: $БФ_0 \subset БФ_1 \subset ... \subset БФ_s$ такую, что $\rho^\sigma(0) \le \rho^\sigma(1) \le ... \le \rho^\sigma(s)$, где $\overline{\rho}^\sigma \le \rho^\sigma(s) \le 1$. Таким образом, при расширении $БФ_0$ посредством добавления (+)- и (–)-фактов достигается заданный порог абдуктивного объяснения $БФ_0$

$$\overline{\rho}^\sigma : \tilde{\rho}^\sigma \le \frac{|\tilde{Б}\tilde{Ф}^\sigma(s)|}{|БФ^\sigma(s)|} \le 1, \text{ где } \rho^\sigma(s) = \frac{|\tilde{Б}\tilde{Ф}^\sigma(s)|}{|БФ^\sigma(s)|}.$$

Монотонное возрастание функции $\rho^\sigma(n)$, где n = 0, …, s, означает улучшение объяснения БФ посредством порожденных гипотез о (±)-причинах ($\sigma = +,-$) и большее правдоподобие порожденных гипотез из Н (гипотез о (±)-причинах и гипотез о предсказании эффектов Q^σ посредством выводов по аналогии).

Факты из базы фактов выражают отношение «история болезни больных хроническим панкреатитом (С) – развитие сахарного диабета (Q)». Таким образом, база фактов (массив историй болезней с двумя исходами – развитие сахарного диабета **панкреатогенного генеза** или отсутствие такового) образована парами двух типов $\langle C, Q^+ \rangle$ и $\langle C, Q^- \rangle$, представляющих наличие развития сахарного диабета (Q^+) в истории болезни (С) и отсутствие такового (Q^-) в истории болезни (С), соответственно.

Пары $\langle C, Q^+ \rangle$ и $\langle C, Q^- \rangle$ будем называть (+)-примерами и (–)-примерами, соответственно.

Итак, ДСМ-рассуждение состоит из последовательного применения правил индуктивного вывода, правил вывода по аналогии и заключительного применения абдукции – принятия порожденных гипотез из Н посредством объяснения БФ гипотезами о (±)-причинах их H_2.

Охарактеризуем правила индуктивного вывода, посредством которых распознается сходство (+)-фактов и, соответственно, (–)-фактов такое, что оно используется для порождения гипотез о (+)- и (–)-причинах.

Правила индуктивного вывода были сформулированы английским мыслителем XIX в. Д.С. Миллем, который рассмотрел пять правил и назвал их методами сходства, различия, сходства-различия, остатков и сопутствующих изменений. Эти правила были представлены как некоторые схемы вывода, характеризуемые словесно с помощью вспомогательных символов [2]. Далее приведем правила индуктивного вывода для методов сходства и различия, которые впоследствии были формализованы посредством языка современной математической логики [3, 4].

Метод сходства

Если два или более случаев подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство – в котором только и согласуются все эти случаи – есть причина (или следствие) данного явления.

Метод различия

Если случай, в котором исследуемое явление наступает, и случай, в котором оно не наступает, сходны во всех обстоятельствах, кроме одного, встречающегося лишь в первом случае, то это обстоятельство, в котором одном только и разнятся эти два случая, есть следствие, или причина, или необходимая часть причины явления.

Формализация индуктивных методов Д.С. Милля в [3, 4] реализует **первый принцип** качественного анализа данных: сходство фактов влечет наличие (отсутствие) исследуемого эффекта и его повторяемость в исходных данных.

Вторым принципом качественного анализа данных является предварительное выяснение типа исходных данных, представляющих некоторую предметную область Д:

(1) если предметная область Д1 является множеством случайных событий, то следует применять статистические методы анализа данных;

(2) если предметная область Д2 является множеством событий таких, что как наличие результирующих эффектов, так и их отсутствие вынуждаются комбинациями определенных факторов, то следует применять логико-комбинаторные методы анализа данных;

(3) третий тип предметных областей ДЗ является соединением типов (1) и (2).

Третьим принципом качественного анализа данных является принятие решений относительно имеющихся экспериментальных данных посредством аргументов «за», оправдывающих наличие изучаемого эффекта и аргументов «против» его наличия.

В соответствии с этими тремя принципами формализуются ДСМ-рассуждения и их начальная процедура – правила индуктивного вывода.

Средствами формализации этих правил являются предикаты сходства – M^{σ}-предикаты, где $\sigma = +,-$. Эти предикаты являются функциями такими, что область их определения – это множество пар $\langle C, Q^{\sigma} \rangle$, извлекаемые из баз фактов, а область значений – истинностные значения двузначной логики – t (истина) и f (ложь). M^{σ}-предикаты – это функции двух переменных V и W, где V – переменная для сходств (σ)-фактов ($\sigma = +,-$), а W – переменная для исследуемого эффекта. Таким образом, M^{σ}-предикатами являются $M^{+}(V,W)$ и $M^{-}(V,W)$ такие, что $M^{+}(C,Q^{+})$ и $M^{-}(C,Q^{-})$ – истинны, т. е., $M^{+}(C,Q^{+})$= t и $M^{-}(C,Q^{-})$ = t, где C – история болезни, а Q^{+} и Q^{-} – наличие и отсутствие развития сахарного диабета панкреатогенного генеза, соответственно.

Важным обстоятельством, соответствующим первому принципу качественного анализа данных ((+)-фактов и (–)-фактов), является средство установления сходства (σ)-фактов (соответственно, (+)-фактов из $БФ^{+}$ и (–)-фактов из $БФ^{-}$, где $БФ^{+} \subset БФ$ и $БФ^{-} \subset БФ$).

Сходство (+)-фактов и, соответственно, (–)-фактов устанавливается посредством M^{σ}-предикатов для того, чтобы обнаружить причину h^{+} для Q^{+} в истории болезни C (аналогично для h^{-}, Q^{-} и C). В силу этого $M^{+}(V,W)$ содержит пять условий для того, чтобы в $БФ^{+}$ обнаружить сходство (+)-фактов такое, чтобы оно было **кандидатом** для извлечения **причины** h^{+} для Q^{+} из C (истории болезней). Аналогичное имеет место для h^{-} для Q^{-} из C.

Сформулируем теперь пять условий, содержащихся в $M^{+}(V,W)$ для извлечения из $БФ^{+}$ сходств (+)-фактов, которые могут быть кандидатами (+)-причин h^{+} для Q^{+} в C.

(1) Экзистенциальное условие (ЭУ$^{+}$).

В БФ должны существовать (+)-факты $\langle C_1, Q^{+} \rangle, ..., \langle C_k, Q^{+} \rangle$.

(2) Собственное сходство (СХ).

$C_1 \cap ... \cap C_k = h^{+}$ и $h^{+} \neq \emptyset$ – это означает, что общая часть историй болезней $C_1, ..., C_k$ есть h^{+} и h^{+} не пусто (имеет множество факторов).

(3) Эмпирическая зависимость (ЭЗ$^{+}$).

Если $\langle X, Q^{+} \rangle$ есть любой (+)-факт (X – переменная, значение которой есть история болезни C) и V содержится в X ($V \subset X$), то X имеет диагноз Q^{+} (развитие сахарного диабета панкреатогенного генеза у X).

(4) Условие исчерпываемости (УИ).

Все (+)-факты в БФ $\langle C_i, Q^{+} \rangle$ такие, что h^{+} содержится в $C_i (h^{+} \subset C_i)$ **исчерпаны** (других таких прецедентов нет).

(5) Нижняя граница обнаруженных сходств (+)-фактов, содержащих h^{+} $k \geq 2$, где k – число всех сходных фактов из (1) и (2), содержащих h^{+} (h^{+} – кандидат в (+)-причины).

Предикат $M^{+}(V,W)$ определяется следующим образом: $M^{+}(V,W)$ истинен для пары $\langle C, Q^{+} \rangle$, где C – значение V, а Q^{+} – значение W тогда и только тогда, когда одновременно выполняются условия (1), (2), (3), (4), (5). Это означает, что $M^{+}(C,Q^{+}) = t$. Аналогично определяется $M^{-}(V,W)$, заданный для (–)-фактов на БФ.

Заметим, что определенные $M^{+}(V,W)$ и $M^{-}(V,W)$ выражают **минимальные** условия обнаружения сходств в БФ для (+)- и (–)-фактов, соответственно. Для определенных выше M^{σ}-предикатов введем обозначение $M_a^{\sigma}(V,W)$, где «а» – соответствующее им имя.

Сформулируем теперь правила индуктивного вывода (п.п.в.-1)[1], используя для этого переменные V и W: V – для причин, W – для эффектов (следствий причин). Посредством $X \Rightarrow_1 Y$ и $V \Rightarrow_2 W$ обозначим предикаты (логические функции): «X обладает множеством свойств Y» и «V есть причина W», соответственно.

Областью определения предиката $X \Rightarrow_1 Y$ будет множество пар $\langle C, Q \rangle$, где C – значение X и Q – значение Y (напомним, что C – история болезни, а Q – развитие сахарного диабета панкреатогенного генеза). Областью значений $X \Rightarrow_1 Y$ будут истинностные значения $\overline{v} = \langle v, n \rangle$, где v – типы истинностных значений – фактическая истина, фактическая ложь, фактическое противоречие и неопределенность, соответственно, а n – число применений правил правдоподобного вывода, выражающее степень правдоподобия утверждений. Чем больше n, тем меньше степень правдоподобия: степень правдоподобия n-фактов из БФ и порожденных гипотез такова, что, соответственно, n = 0 (для фактов) и n > 0 (для гипотез).

Если $\langle C, Q^{+} \rangle$, $\langle C, Q^{-} \rangle$ – факты, то это, соответственно, означает, что $C \Rightarrow_1 Q = \langle 1,0 \rangle$, $C \Rightarrow_1 Q = \langle -1,0 \rangle$, т.е., утверждение «C есть история болезни, содержащая развитие сахарного диабета панкреатогенного генеза» является **фактически истинным** (для

[1] Индуктивный вывод является правдоподобным выводом [3], так как его результаты не всегда бывают истинными, но их правдоподобие оправдано специальными процедурами (п.п.в.-1 – правилами правдоподобного вывода для индукции).

$\langle C,Q^+\rangle$) и, соответственно, – **фактически ложно** (для $\langle C,Q^-\rangle$).

Таким образом, $\langle 1,0\rangle$ и $\langle -1,0\rangle$ – истинностные значения фактов из БФ (n = 0 – правила правдоподобного вывода не применялись). Следовательно, $C \Rightarrow {}_1Q = \langle 1,0\rangle$ и $C \Rightarrow {}_1Q = \langle -1,0\rangle$ – другие обозначения для $\langle C,Q^+\rangle$ и $\langle C,Q^-\rangle$, соответственно. Если пара $\langle C,Q\rangle$ представляет неопределенность, предлагаемую для предсказания, то $C \Rightarrow {}_1Q = \langle \tau,0\rangle$, где τ – тип истинностного значения «неопределенно» [3].

Таким образом, факты из базы фактов имеют типы истинностных значений 1, -1, τ, а сами истинностные значения (оценки утверждений) представимы как $\overline{v} = \langle v,n\rangle$, где $v = 1,-1,0,\tau$, а n = 0,1,2,… Напомним, что n – число шагов применения правил правдоподобного вывода, выражающего степень правдоподобия утверждений (n > 0 – для гипотез, n = 0 – для фактов).

Четыре типа истинностных значений (1, -1, 0, τ) соответствуют третьему принципу качественного анализа экспериментальных данных: утверждение принимается с типом истинностного значения «1» (фактически истинно), если есть аргументы «за» и нет аргументов «против»; утверждение принимается с типом истинностного значения «-1» (фактически ложно), если нет аргументов «за» и есть аргументы «против»; утверждение не принимается, если есть аргументы «за» и есть аргументы «против», тогда оно имеет тип истинностного значения «0» (фактически противоречиво); утверждение остается неопределенным, если нет аргументов «за» и нет аргументов «против», тогда тип истинностного значения утверждения есть «τ» (неопределенность) [3].

Далее определим правила правдоподобного вывода для индукции (п.п.в.-1), используя предикаты сходства $M_a^+(V,W)$ и $M_a^-(V,W)$, называемые предикатами простого (минимального) сходства. С этой целью рассмотрим логическую операцию $J_{\overline{v}}$ – такую, что она характеризует приписывание истинностных значений утверждениям «”С обладает Q” и имеет истинностное значение $\overline{v}$», где $\overline{v} = \langle v,n\rangle$, $v = 1,-1,0,\tau$, а n = 0,1,2,… :

$$J_{\overline{v}}(C \Rightarrow {}_1Q) = \begin{cases} t, если\, C \Rightarrow {}_1Q = \overline{v} \\ f, если\, C \Rightarrow {}_1Q \neq \overline{v} \end{cases},$$

где t, f – истинностные значения двузначной логики (истинно, ложно).

Таким образом, $J_{\langle 1,0\rangle}(C \Rightarrow {}_1Q)$ – факт: в истории болезни С представлено развитие сахарного диабета панкреатогенного генеза у больного хроническим панкреатитом; $J_{\langle -1,0\rangle}(C \Rightarrow {}_1Q)$ – факт: в истории болезни С не имеется развития сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом.

Введем в рассмотрение также предикат $V \Rightarrow {}_2W$ – «V есть причина W» [3]. В нашем исследовании значениями переменной V является C', такие, что C' содержится в истории болезни С ($C' \subset C$), а значениями W являются Q^+ и Q^- – наличие развития сахарного диабета панкреатогенного генеза и отсутствие развития сахарного диабета панкреатогенного генеза, соответственно.

Правила индуктивного вывода (п.п.в.-1) применяются к БФ, а результатом их применения являются гипотезы

$$J_{\langle 1,n\rangle}(C' \Rightarrow {}_2Q), J_{\langle -1,n\rangle}(C' \Rightarrow {}_2Q), J_{\langle 0,n\rangle}(C' \Rightarrow {}_2Q)$$
$$\text{и } J_{\langle \tau,n\rangle}(C' \Rightarrow {}_2Q),$$

где $n \geq 1$, а C' содержится в С ($C' \subset C$), где С – история болезни. Эти гипотезы утверждают, что C' есть причина развития сахарного диабета панкреатогенного генеза ($J_{\langle 1,n\rangle}(C' \Rightarrow {}_2Q) = t$), C' есть причина отсутствия сахарного диабета панкреатогенного генеза ($J_{\langle -1,n\rangle}(C' \Rightarrow {}_2Q) = t$), фактически противоречиво, что C' есть причина сахарного диабета панкреатогенного генеза (есть аргументы «за» и «против» одновременно) ($J_{\langle 0,n\rangle}(C' \Rightarrow {}_2Q) = t$), если нет аргументов «за» и нет аргументов «против», то сохраняется оценка $C' \Rightarrow {}_2Q$ посредством типа истинностного значения τ (неопределенность) – ($J_{\langle \tau,n\rangle}(C' \Rightarrow {}_2Q) = t$).

Сформулируем теперь четыре правила индуктивного вывода (п.п.в.-1), где «&» – логическое «И», а «¬» – логическое «не» (отрицание).

$$(I)^+ \frac{J_{\langle \tau,n\rangle}(V \Rightarrow {}_2W), M_a^+(V,W) \& \neg M_a^-(V,W)}{J_{\langle 1,n+1\rangle}(V \Rightarrow {}_2W)},$$

$$(I)^- \frac{J_{\langle \tau,n\rangle}(V \Rightarrow {}_2W), \neg M_a^+(V,W) \& M_a^-(V,W)}{J_{\langle -1,n+1\rangle}(V \Rightarrow {}_2W)},$$

$$(I)^0 \frac{J_{\langle \tau,n\rangle}(V \Rightarrow {}_2W), M_a^+(V,W) \& \mathrm{M}_a^-(V,W)}{J_{\langle 0,n+1\rangle}(V \Rightarrow {}_2W)},$$

$$(I)^\tau \frac{J_{\langle \tau,n\rangle}(V \Rightarrow {}_2W), \neg M_a^+(V,W) \& \neg M_a^-(V,W)}{J_{\langle \tau,n+1\rangle}(V \Rightarrow {}_2W)}.$$

$(I)^+$: имеется сходство на (+)-примерах, обнаруженное посредством $M_a^+(V,W)$ и не имеется соответствующего сходства на (–)-примерах, что представимо посредством $\neg M_a^-(V,W)$ (неверно, что $M_a^-(V,W)$). Поэтому $(I)^+$ порождает гипотезу $J_{\langle 1,n+1\rangle}(V \Rightarrow {}_2W)$: V есть (+)-причина W (аргументом «за» является сходство (+)-примеров, распознаваемое посредством $M_a^+(V,W)$).

$(I)^-$: имеется сходство на (–)-примерах, обнаруженное посредством $M_a^-(V,W)$ и не имеется соответствующего сходства на (+)-примерах, что представимо посредством $\neg M^+(V,W)$ (неверно, что $M_a^+(V,W)$).

Поэтому $(I)^-$ порождает гипотезу $J_{\langle -1,n+1\rangle}(V \Rightarrow {}_2W)$: V есть (–)-причина W (аргументом «против» является сходство (–)-примеров, распознаваемое посредством $M_a^-(V,W)$.

$(I)^0$: имеется сходство на (+)-примерах и на (–)-примерах такое, что оно совпадает (т.е. существуют совпадающие аргументы «за» и «против» того, чтобы V была причиной W). Поэтому $(I)^0$ порождает гипотезу $J_{\langle 0,n+1\rangle}(V \Rightarrow {}_2W)$: утверждение, что V есть причина W является фактически противоречивым.

$(I)^\tau$: не имеется сходство как на (+)-примерах, так и на (–)-примерах (следовательно, $M_a^+(V,W)$ и $M_a^-(V,W)$ – ложны). Это означает, что нет аргументов «за» то, чтобы V была причиной W, и нет аргументов «против» того, чтобы V была причиной W. Поэтому $(I)^\tau$ порождает гипотезу о том, что утверждение «V есть причина W» и является неопределенным (сохраняется тип истинностного значения τ (неопределенность)).

Заметим, что в правилах индуктивного вывода $(I)^+,(I)^-,(I)^0,(I)^\tau$ степень правдоподобного следствия есть n + 1 (n + 1 – число шагов правдоподобного вывода п.п.в.-1), но чем больше шагов, тем меньше правдоподобия гипотезы, ибо максимальное правдоподобие имеет факт, относительно которого число шагов вывода n = 0.

Напомним, что вторым принципом качественного анализа данных является принцип сходства: сходство фактов влечет наличие (отсутствие) эффекта и его повторяемость[2]. Следовательно, от условий **распознавания** сходства (+)- и (–)-примеров зависит обоснованность принятия обнаруженных гипотез о (±)-причинах. Таким образом, структурное сходство (±)-примеров является условием каузального вынуждения наличия (отсутствия) изучаемых эффектов. Поэтому усиление условий распознавания сходств (±)-примеров посредством добавления к M_a^σ-предикатам сходства дополнительных условий повышает обоснованность порождаемых гипотез о (±)-причинах, что соответствует первому и второму принципу качественного анализа данных, характерному для медицинских исследований.

Далее рассмотрим усиления для распознавания сходств (±)-примеров, используемых для порождения гипотез о (±)-причинах изучаемых эффектов. В нашем исследовании таковыми являются наличие и отсутствие сахарного диабета панкреатогенного генеза.

Рассмотрим усиление распознавания сходств (±)-примеров, которое добавляется к предикатам для простого сходства $M_a^+(V,W)$ и $M_a^-(V,W)$, которые как уже отмечалось, являются **минимальными** средствами обнаружения сходств в БФ.

Усиление $M_a^\sigma(V,W)$, где $\sigma = +-$, посредством условий запрета на контрпримеры.

Рассмотрим случай для $M_a^+(V,W)$, применяемый к (+)-примерам.[3] Добавим к $M_a^+(V,W)$ условие: кандидат в причины V эффекта W содержится только в (+)-примерах или (τ)-примерах с эффектом W [3]. Это означает, что для каждого X и каждого Y, если V содержится в X и W содержится в Y, то имеет место, что $X \Rightarrow_1 Y$ есть (+)-пример или (τ)-пример. Это означает, что V и W не могут содержаться в (–)-примерах и (0)-примерах. Следовательно, в этом случае гипотезы о (+)-причинах порождаются только (+)-примерами.

Ещё раз сформулируем условие запрета на контрпримере для $M_a^+(V,W)$: для всех X и всех Y, если $V \subset X$ и $W \subseteq Y$, то имеет место $J_{\langle 1,n\rangle}(X \Rightarrow_1 Y)$ или $J_{\langle \tau,n\rangle}(X \Rightarrow_1 Y)$.

Аналогично формулируется условие запрета на контрпримеры для $M_a^-(V,W)$: для всех X и всех Y, если $V \subset X$ и $W \subseteq Y$, то имеет место $J_{\langle -1,n\rangle}(X \Rightarrow_1 Y)$ или $J_{\langle \tau,n\rangle}(X \Rightarrow_1 Y)$.

Это означает, что V и W не могут содержаться в (+)-примерах.

Введем обозначение для предикатов (±)-сходства с запретами на контрпримеры:

$$M_{ab}^+(V,W) = M_a^+(V,W)\,\&\,b^+(V,W),$$
$$M_{ab}^-(V,W) = M_a^-(V,W)\,\&\,b^-(V,W),$$

где $b^+(V,W)$ и $b^-(V,W)$ представляют условия запрета на контрпримеры для (+)- и (–)-примеров, соответственно.

Усиление $M_a^\sigma(V,W)$, где $\sigma = +-$, посредством условия различия [2, 3, 5].

Идея индуктивного метода различия принадлежит Д. Гершелю [6], а в виде правила индуктивного вывода она была сформулирована Д.С. Миллем [2].

Метод различия формализован в качестве одного из правил индуктивного вывода ДСМ-рассуждений посредством предикатов $M_{ad_0}^\sigma(V,W)$, где $\sigma = +-$, для порождения сходств (σ)-примеров. Эти предикаты **усиливают** предикаты простого сходства $M_a^\sigma(V,W)$ добавлением условий $d_0^\sigma(V,W)$, где $\sigma = +-$, [5].

Определим условие $d_0^+(V,W)$ для случая (+)-примеров: для любых X,Y,Z, и таких, что если $X \Rightarrow_1 Y$ есть (+)-пример, V содержится в X, W содержится в $Y(W \subseteq Y)$, а из X удалено V $(X - V)$ и $X - V$ не пусто $(X - V \neq \varnothing)$ и $X - V$ содержится в $Z((X - V) \subset Z)$ и неверно, что V содержится в $Z(\neg(V \subset Z)$, где "¬" –

[2] Отметим, что в случае предметных областей, образованных случайными событиями, сходство фактов определяется посредством частоты их повторения. Тогда естественно применение статистических методов.

[3] (±)-факты и (±)-гипотезы, выразимые посредством предиката $X \Rightarrow_1 Y$ (объект X обладает множеством свойств Y), называются (±)-примерами.

отрицание «не»), то неверно, что $Z \Rightarrow_1 U$ есть (+)-пример и W содержится в $Y(\neg(J_{\langle 1,n\rangle}(Z \Rightarrow_1 U)\,\&\,(W \subseteq Y)))$[4].

В нашем исследовании Y есть Q – развитие сахарного диабета панкреатогенного генеза, поэтому получим упрощение формулы $d_0^+(V,W)$, так как $W = Y = Q$:

$$d_0^+(V,Q) = \forall X \forall Z((J_{\langle 1,n\rangle}(X \Rightarrow_1 Q)\,\&$$
$$\&(V \subset X)\,\&\,((X-V) \subset Z)\,\&\,((X-V) \neq \varnothing)\,\&$$
$$\&\neg(V \subset Z)) \rightarrow \neg J_{\langle 1,n\rangle}(Z \Rightarrow_1 Q)),$$

где $\varnothing$ – пустое множество элементов.

Аналогично определяется условие $d_0^-(V,Q)$ для (–)-примеров.

Таким образом, получаем определения $M_{ad_0}^+$- и $M_{ad_0}^-$-предикатов для индуктивного метода различия [2, 5]: $M_{ad_0}^+(V,W) = M_a^+(V,W)\,\&\,d_0^+(V,W)$, $M_{ad_0}^-(V,W) = M_a^-(V,W)\,\&\,d_0^-(V,W)$. Очевидно, что для нашего исследования имеем: $M_{ad_0}^\sigma(V,Q) = M_a^\sigma(V,Q)\,\&\,d_0^\sigma(V,Q)$, где $\sigma = +,-$.

Итак, мы имеем предикаты простого сходства $M_a^\sigma(V,W)$, где $\sigma = +,-$ и их усиления посредством условий $b^\sigma(V,W)$, $d^\sigma(V,W)$. Следовательно, получаем следующие возможные предикаты для распознавания сходства (кандидатов в гипотезы о (±)-причинах) в БФ: $M_a^\sigma(V,W)$, $M_{ab}^\sigma(V,W)$, $M_{ad_0}^\sigma(V,W)$, $M_{ad_0b}^\sigma(V,W)$, где $M_{ad_0b}^\sigma(V,W) = M_a^\sigma(V,W)\,\&\,b^\sigma(V,W)\,\&\,d_0^\sigma(V,W)$, где $\sigma = +,-$.

Для нашего исследования имеем $M_a^\sigma(V,Q)$ и условия $d_0^\sigma(V,Q)$, $b^\sigma(V,Q)$.

§2. СТРАТЕГИИ ДСМ-РАССУЖДЕНИЙ И ДСМ-ИССЛЕДОВАНИЯ: ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

Рассмотренные четыре предиката могут быть упорядочены по «логической силе» посредством дедуктивной выводимости [5]. Это упорядочение можно представить в виде алгебраических решеток [7, 5]. С этой целью заменим M^σ-предикаты на соответствующие им представления (имена) a^σ, $(ab)^\sigma$, $(ad_0)^\sigma$, $(ad_0b)^\sigma$, где $\sigma = +,-$, и получим следующие решетки M^σ-предикатов (рис. 1 и рис. 2).

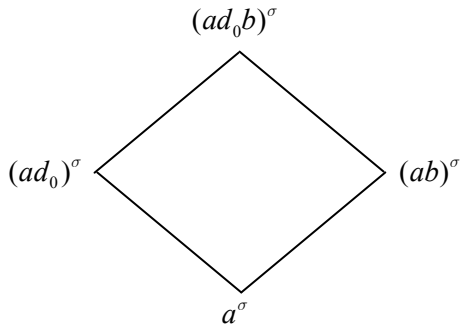


Рис.1

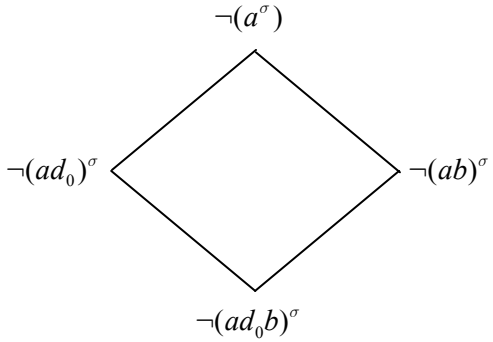


Рис.2

Используя определения M^σ-предикатов, расширяется множество правил индуктивного вывода (оно, ранее, было сформулировано для $(I)_a^+$, $(I)_a^-$, $(I)_a^0$, $(I)_a^\tau$, которым соответствовали комбинации M^σ-предикатов $(\sigma = +,-)$

$$M_a^+(V,W)\,\&\,\neg M_a^-(V,W), \neg M_a^+(V,W)\,\&$$
$$\&M_a^-(V,W), M_a^+(V,W)\,\&\,M_a^-(V,W), \neg M_a^+(V,W)\,\&$$
$$\&\ \neg M_a^-(V,W).$$

Теперь имеется 16 комбинаций M^σ-предикатов из $M_x^+(V,W)$, $M_y^-(V,W)$ и их отрицаний $\neg$ таких, что x принадлежит множеству

$I^+ = \{a^+, (ab)^+, (ad_0)^+, (ad_0b)^+\}$ (рис.1),

$\neg x$ принадлежит множеству

$\neg I^+ = \{\neg a^+, \neg(ab)^+, \neg(ad_0)^+, \neg(ad_0b)^+\}$ (рис.2),

а y и $\neg y$ принадлежат, соответственно, множествам

$I^- = \{a^-, (ab)^-, (ad_0)^-, (ad_0b)^-\}$ (рис.1)

и $\neg I^- = \{\neg a^-, \neg(ab)^-, \neg(ad_0)^-, \neg(ad_0b)^-\}$ (рис.2).

$a^\sigma, (ab)^\sigma, \ldots, \neg(ad_0)^\sigma, \neg(ad_0b)^\sigma$, где $\sigma = +,-$, представляют соответствующие M^σ-предикаты и их отрицания $(\neg M^\sigma)$. Тогда получаем множество п.п.в.-1 $(I)_{x,\neg y}^+$, $(I)_{\neg x,y}^-$, $(I)_{x,y}^0$ и $(I)_{\neg x,\neg y}^\tau$.

В частности, для $(I)_{x,\neg y}^+$ имеем п.п.в.-1$^{(+)}$: $(I)_{a,\neg a}^+$, $(I)_{a,\neg(ab)}^+$, …, $(I)_{ad_0b,\neg(ad_0b)}^+$ (всего 16 комбинаций п.п.в.-1$^{(+)}$).

[4] Представление $d_0^+(V,W)$ на языке логики:
$\forall$ – операция «для всех», & - логическая связка «и»,
$\neg$ – логическая связка отрицания («не»),
$\rightarrow$ – логическая связка «если…, то …», тогда $d_0^+(V,W) = \forall X \forall Y \forall Z \forall U((J_{\langle 1,n\rangle}(X \Rightarrow_1 Y)\,\&\,(W \subseteq Y)\,\&\,(V \subset X)\,\&\,\&((X-V) \subset Z)\,\&\,((X-V) \neq \varnothing\,\&\,\neg(V \subset Z)) \rightarrow \neg(J_{\langle 1,n\rangle}(Z \Rightarrow_1 U)\,\&\,(W \subseteq Y)))$,
где $X-V$ – разность множеств X и V («X без V»)

Например:

$$(I)^{+}_{ab,\neg(ad_0)}\frac{J_{\langle\tau,n\rangle}(V\Rightarrow_2 W),M^{+}_{ab}(V,W)\,\&\,\neg M^{-}_{ad_0}(V,W)}{J_{\langle 1,n+1\rangle}(V\Rightarrow_2 W)}.$$

Общий вид для правил $(I)^{+}$:

$$(I)^{+}_{x,\neg y}\frac{J_{\langle\tau,n\rangle}(V\Rightarrow_2 W),M^{+}_{x}(V,W)\,\&\,\neg M^{-}_{y}(V,W)}{J_{\langle 1,n+1\rangle}(V\Rightarrow_2 W)}.$$

Аналогичное имеет место для всех п.п.в.-$1^{(\sigma)}$, где $\sigma=-,0,\tau$.

Заметим, что правила индуктивного (правдоподобного) вывода сформулированы как схемы конкретных выводов, так как они используют переменные V и W (V – для гипотез о (±)-причинах, W – для эффекта, которым является наличие или отсутствие сахарного диабета). Реализацией этих правил являются подстановки C' и Q, соответственно, вместо V и W, где C' – фрагменты историй болезней, а Q – эффект.

Напомним, что ДСМ-рассуждение образовано тремя процедурами – индуктивными выводами (п.п.в.-1), выводами по аналогии и абдуктивным принятием порожденных гипотез (схема принятия гипотез посредством абдукции была сформулирована выше).

Сформулируем теперь правила вывода по аналогии, посредством которых осуществляется предсказание сведений из БФ, имеющих начальную оценку неопределенности (τ). Эту часть БФ обозначают посредством $БФ^{\tau}$ ($БФ^{\tau}\subset БФ$). Элементы $БФ^{\tau}$ представимы утверждениями $J_{\langle\tau,0\rangle}(C\Rightarrow_1 Q)$, выражающими неопределенность относительно того будет ли у больного хроническим панкреатитом (история болезни С) развиваться сахарный диабет панкреатогенного генеза (Q). Таким образом, в результате вывода по аналогии на основе имеющихся прецедентов тип истинностного значения τ может быть заменен на «1» (тип истинностного значения «фактическая истина» – возникнет сахарный диабет панкреатогенного генеза), или на «-1» (тип истинностного значения «фактическая ложь» – сахарный диабет панкреатогенного генеза не возникает), или на «0» (тип истинностного значения «фактическое противоречие» – имеются как аргументы «за», так и аргументы «против» возникновения сахарного диабета панкреатогенного генеза), или на «τ» (неопределенность сохраняется – нет аргументов ни «за», ни «против»).

Правила правдоподобного вывода по аналогии (п.п.в.-2) формулируются посредством предикатов $\Pi^{\sigma}(V,W)$, где $\sigma=+,-,0,\tau$.

Охарактеризуем смысл $\Pi^{+}(V,W)$. Пара $\langle C,Q\rangle$ выполняет $\Pi^{+}(V,W)$, то есть, делает истинным $\Pi^{+}(C,Q)(\Pi^{+}(C,Q)=t)$, если существует такое X, что $J_{\langle 1,n\rangle}(X\Rightarrow_2 Q)=t$ и X содержится в C ($X\subset C$), но не существует Z такого, что $J_{\langle -1,n\rangle}(Z\Rightarrow_2 Q)=t$ и Z содержится в C ($Z\subset C$).

Таким образом, если $\Pi^{+}(C,Q)=t$, то это означает, что существует гипотеза о (+)-причине C' такая, что $C'\Rightarrow_2 Q$ фактически истинна (на шаге n) и C' содержится в C ($C'\subset C$), C' есть значение X; но не существует (–)-причины Z такое, что $Z\Rightarrow_2 Q$ фактически ложно (т.е. $\langle -1,n\rangle$ истинностное значение (–)-причины $Z\Rightarrow_2 Q$), а Z содержится в с ($Z\subset C$). Говоря неформально C имеет (+)-причину для Q (аргумент «за») и не имеет каких-либо (–)-причин в качестве аргументов «против».

Правило правдоподобного вывода по аналогии (п.п.в.-2) для предсказания (+)-фактов представимо следующим образом:

$$(II)^{+}\frac{J_{\langle\tau,n\rangle}(V\Rightarrow_1 W),\Pi^{+}(V,W)}{J_{\langle 1,n+1\rangle}(V\Rightarrow_1 W)}.$$

Так как гипотеза (+)-причины $C'\Rightarrow_2 Q$ порождена сходными (+)-примерами, выразимыми посредством предиката $X\Rightarrow_1 Y$, то результат п.п.в.-1 $(II)^{+}C\Rightarrow_1 Q$ аналогичен этим сходным примерам, а их сходством является $C'\Rightarrow_2 Q$, где $C'\subset C$.

Аналогично формулируется п.п.в.-2 для предсказания (–)-примеров посредством предиката $\Pi^{-}(V,W)$, определяемого посредством гипотез о (–)-причинах.

Правила для выводов по аналогии, порождающие гипотезы с оценками «фактически противоречиво», формулируются посредством предиката $\Pi^{0}(V,W)$. Пара $\langle C,Q\rangle$ выполняет $\Pi^{0}(V,W)$ (т. е., $\Pi^{0}(C,Q)=t$), если существуют гипотезы (+)-причине и (–)-причине $X_1\Rightarrow_2 Q$ и $X_2\Rightarrow_2 Q$ такие, что X_1 и X_2 содержатся в C ($X_1\subset C$ и $X_2\subset C$). Таким образом, гипотезы $J_{\langle 1,n\rangle}(X\Rightarrow_2 Q)$ и $J_{\langle -1,n\rangle}(X\Rightarrow_2 Q)$ конфликтуют, а потому $п.п.в.-2^{0}$

$$(II)^{0}\frac{J_{\langle\tau,n\rangle}(V\Rightarrow_1 Q),\Pi^{0}(V,Q)}{J_{\langle 0,n+1\rangle}(V\Rightarrow_1 Q)}$$

порождает предсказание о фактической противоречивости $V\Rightarrow_1 Q$.

Правило вывода по аналогии, сохраняющее неопределенность при предсказании, формулируется посредством предиката $\Pi^{\tau}(V,W)$. $\Pi^{\tau}(V,W)$ определяется посредством отрицания $\Pi^{+}(V,W),\Pi^{-}(V,W)$ и $\Pi^{0}(V,W)$, а именно:

$$\Pi^{\tau}(V,W)=\neg\Pi^{+}(V,W)\,\&\,\neg\Pi^{-}(V,W)\,\&\,\neg\Pi^{0}(V,W)$$

$$(II)^{\tau}\frac{J_{\langle\tau,n\rangle}(V\Rightarrow_1 Q),\Pi^{\tau}(V,Q)}{J_{\langle\tau,n+1\rangle}(V\Rightarrow_1 Q)}.$$

Общие формулировки п.п.в.-2 для $(II)^{+},(II)^{-},(II)^{0}$ и $(II)^{\tau}$ используют предикаты $\Pi^{\sigma}(V,W)$, где $\sigma=+,-,0,\tau$, а в случае нашего исследования W = Q.

Заметим, что, так как $\Pi^{\sigma}(V,W)$ определяются посредством предиката $X\Rightarrow_2 Y$ (X есть причина Y), то правила вывода по аналогии ($п.п.в.-2^{(\sigma)}$) используют $п.п.в.-1^{(\sigma)}$ $(I)^{+}_{x,\neg y},(I)^{-}_{\neg x,y},(I)^{0}_{x,y}$ и $(I)^{\tau}_{\neg x,\neg y}$, по-

средством которых порождаются гипотезы о (±)-причинах $J_{\langle 1,n\rangle}(X \Rightarrow {}_2Y)$ и $J_{\langle -1,n\rangle}(X \Rightarrow {}_2Y)$. Поэтому имеется четыре вида п.п.в.-2 $(II)^{+}_{x,\neg y}, (II)^{-}_{\neg x,y}, (II)^{0}_{x,y}$ и $(II)^{\tau}_{\neg x,\neg y}$ такие, что каждому $n.n.в.-2^{(\sigma)}$ $(II)^{\sigma}$, где $\sigma = +,-,0,\tau$, соответствует 16 вариантов, что следует из рис. 1 и рис. 2.

Например, $(II)^{+}_{a,a}, (II)^{+}_{a,ab}, (II)^{+}_{ab,a}, (II)^{+}_{ab,ab}$ и т.д.

Стратегией ДСМ-рассуждения [8] будем называть выбор п.п.в.-1 и, соответственно п.п.в.-2 для соответствующих x и y (и их отрицаний).

Таким образом, стратегия $Str_{x,y}$ образована

$$(I)^{+}_{x,\neg y}, (II)^{+}_{x,\neg y}; (I)^{-}_{\neg x,y}, (II)^{-}_{\neg x,y};$$
$$(I)^{0}_{x,y}, (II)^{0}_{x,y}; (I)^{\tau}_{\neg x,\neg y}, (II)^{\tau}_{\neg x,\neg y}.$$

$Str_{x,y}$ есть последовательное применение п.п.в.-1 и п.п.в.-2.

Приводим перечень стратегий, использованных в настоящем исследовании [9]:

(1) $Str_{a,a}$ (9) Str_{a,ad_0}

(2) $Str_{ab,a}$ (10) Str_{ab,ad_0}

(3) $Str_{ad_0,a}$ (11) Str_{ad_0,ad_0}

(4) $Str_{ad_0b,a}$ (12) Str_{ad_0b,ad_0}

(5) $Str_{a,ab}$ (13) Str_{a,ad_0b}

(6) $Str_{ab,ab}$ (14) Str_{ab,ad_0b}

(7) $Str_{ad_0,ab}$ (15) Str_{ad_0,ad_0b}

(8) $Str_{ad_0b,ab}$ (16) Str_{ad_0b,ad_0b}.

Множество используемых стратегий $Str_{x,y}$ ДСМ-рассуждений (1) – (16) обозначим посредством $\overline{Str}$. Применение различных $Str_{x,y}$ из $\overline{Str}$ согласуется с первым принципом качественного анализа данных: сходство фактов влечет наличие (отсутствие) эффекта и его повторяемость. Элементы $\overline{Str}$ частично упорядочены [5, 8] по «логической силе», если они сравнимы. Их сравнимость или несравнимость обусловлена их расположением в решетках для M^{σ}-предикатов и их отрицаний ($\neg M^{\sigma}$), представленных на рис. 1 и рис. 2 (множество $\overline{Str}$ является множеством прямых произведений указанных решеток [5, 7]).

Элементы $\overline{Str}$ представляют разные виды сходств фактов, а эти виды сходств являются различными **механизмами каузального вынуждения** рассматриваемых эффектов (в нашем исследовании это – наличие или отсутствие развития сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом [9]). Таким образом, первый принцип качественного анализа данных специфицируется посредством распознавания вида каузального вынуждения изучаемого эффекта посредством обнаружения гипотез о (±)-причинах в БФ.

Для $n.n.в.-1^{(+)}$ наибольшим элементом является пара $\langle M^{+}_{ad_0b}, \neg M^{-}_{a}\rangle$, наименьшим – $\langle M^{+}_{a}, \neg M^{-}_{ad_0b}\rangle$, соответствующие этому правилу [5, с. 26]. Для $n.n.в.-1^{(-)}$ наибольшим элементом является пара $\langle \neg M^{+}_{a}, M^{-}_{ad_0b}\rangle$, а наименьшим – $\langle \neg M^{+}_{ad_0b}, M^{-}_{a}\rangle$, соответствующие этому правилу [5, с. 27]. Для $n.n.в.-1^{(0)}$ наибольшим элементом является пара $\langle M^{+}_{ad_0b}, M^{-}_{ad_0b}\rangle$, а наименьшим – $\langle M^{+}_{a}, M^{-}_{a}\rangle$ [5, с. 28]. Для $n.n.в.-1^{(\tau)}$ наибольшим элементом является пара $\langle \neg M^{+}_{a}, \neg M^{-}_{a}\rangle$, а наименьшим – $\langle \neg M^{+}_{ad_0b}, \neg M^{-}_{ad_0b}\rangle$, соответствующие этому правилу [5, с. 29].

В представлении этих правил в соответствующих четырех решетках [5] имеем следующие пары наибольших и наименьших элементов:

$$\left\{\langle (ad_0b)^{+}, \neg a^{-}\rangle, \langle a^{+}, \neg (ad_0b)^{-}\rangle\right\},$$
$$\left\{\langle \neg a^{+}, (ad_0b)^{-}\rangle, \langle \neg (ad_0b)^{+}, a^{-}\rangle\right\},$$
$$\left\{\langle (ad_0b)^{+}\rangle, \langle (ad_0b)^{-}\rangle, \langle a^{+}, a^{-}\rangle\right\},$$
$$\left\{\langle \neg a^{+}, \neg a^{-}\rangle, \langle \neg (ad_0b)^{+}, \neg (ad_0b)^{-}\rangle\right\}.$$

Заметим, что из наибольших элементов решеток логически следуют все остальные 15 элементов, а наименьшие элементы решеток логически следуют из всех других элементов (они в решетках расположены над наименьшими элементами [5]) и, что в оценку качества гипотезы (её «логическая сила») входит вид каузального вынуждения эффекта, представленный парами M^{σ}-предикатов и их отрицаний из $Str_{x,y}$ и соответствующих им $n.n.в.-1^{(\sigma)}$, где $\sigma = +,-,0,\tau.$, а также, что упорядоченность оценок качества гипотез по их «логической силе» распространяется и на гипотезы о предсказаниях наличия (отсутствия) изучаемого эффекта, порождаемых посредством п.п.в.-2 (правилами вывода по аналогии), использующих предикаты $\Pi^{\sigma}(V,W)$, где $\sigma = +,-,0,\tau$.

Проясним теперь, термин «рассуждение», который использовался для характеризации ДСМ-рассуждений, применяемых для порождения гипотез о (±)-причинах и гипотез, предсказывающих наличие или отсутствие изучаемого эффекта.

Рассуждение есть:

1^0. Поиск и выбор посылок, релевантных цели.

2^0. Применение правил вывода к соответствующим посылкам из баз фактов и баз знаний.

3^0. Контроль за выводом (оценка и принятие следствий).

Спецификация этого определения для ДСМ-рассуждений состоит в использовании правил индуктивного вывода (п.п.в.-1) и правил вывода по аналогии (п.п.в.-2), использующих гипотезы о причинах, полученные посредством п.п.в.-1. Последовательность применения правил вывода ДСМ-рассуждения есть

$(n.n.в.-1 \rightarrow n.n.в-2)_1 \rightarrow ... \rightarrow (n.n.в.-1 \rightarrow n.n.в-2)_n$,

где $(n.n.в.-1 \rightarrow n.n.в-2)_i$ – i-й этап ДСМ-рассуждения (i = 1,…, n), а результаты n-го этапа совпадают с результатами (n + 1)-го этапа, представляя **стабилизацию** ДСМ-рассуждения: новые гипотезы не возникают.

Принятие же результатов ДСМ-рассуждения после достигнутой стабилизации осуществляется посредством абдукции, рассмотренной выше и вычисления функций $\rho^{\sigma}(i)$, где i = 1,…, s, а s – номер последней базы фактов, полученной после соответствующих расширений начальной $БФ_0$: $БФ_0 \subset БФ_1 \subset ... \subset БФ_s$.

Таким образом, ДСМ-рассуждение есть **синтез трех познавательных процедур** – индукции (п.п.в.-1), аналогии (п.п.в.-2) и абдукции (принятие гипотез посредством объяснения БФ). Целью же ДСМ-рассуждения является устранение неопределенности посредством предсказания наличия (отсутствия) эффекта у прецедентов, относительно которых неизвестно его наличие или отсутствие (их называют τ-примерами, где τ – тип истинностного значения «неопределенность»).

В результате применения ДСМ-рассуждений τ-примеры получают типы истинностных значений: «1» (есть эффект), в силу наличия аргументов «за» и отсутствия аргументов «против»; «-1» (нет эффекта) в силу отсутствия аргументов «за» и наличия аргументов «против»; «0» (фактическое противоречие) в силу наличия как аргументов «за», так и аргументов «против», и, наконец, возможно сохранение неопределенности «τ» в силу отсутствия как аргументов «за», так и аргументов «против». Предсказание реализуется правилами вывода по аналогии (п.п.в.-2), а аргументами являются гипотезы о (±)-причинах, порожденные правилами индуктивного вывода (п.п.в.-1).

Принятие же гипотез о предсказаниях и гипотез о причинах осуществляется посредством процедуры абдукции, объясняющей базу фактов (полностью, когда $\rho^{\sigma}=1$, или частично, когда $\overline{\rho}^{\sigma} \le \rho^{\sigma} < 1$, где $\sigma = +,-$, а $\rho^{-\sigma}$ – выбранный порог, например, 0.8).

ДСМ-рассуждение для представления знаний использует язык, содержащий как дескриптивные средства для распознавания сходств (±)-примеров, так и аргументативные средства для формализации правил индуктивного вывода и вывода по аналогии, а также для принятия гипотез посредством абдукции.

Эвристика есть рассуждение такое, что оно применяется для достижения цели с использованием правил правдоподобного вывода. Следовательно, ДСМ-рассуждение есть **формализованная эвристика**, такая что её содержанием является реализация познавательного процесса, состоящего из анализа данных (применение индукции), предсказаний (применение аналогии), объяснения результатов (применение абдукции). Эта формализация и автоматизация познавательного процесса извлечения нового знания из баз фактов является инструментом **качественного анализа** медицинских данных.

Однако ДСМ-рассуждения являются лишь первым этапом реализации ДСМ-метода автоматизированной поддержки научных исследований (в настоящей работе – медицинских исследований).

Вторым этапом является собственно **ДСМ-исследование**, реализующее поиск и обнаружение **эмпирических закономерностей** в расширяемых последовательностях баз фактов – $БФ_0$, $БФ_1$, …, $БФ_s$, таких что $БФ_0 \subset БФ_1 \subset \ldots \subset БФ_s$.

Три принципа качественного анализа данных (сходства, выбора типа предметной области, аргументации) должны быть дополнены четвертым принципом обнаружения регулярностей (эмпирических закономерностей) в последовательностях расширяемых баз фактов для исследовательских областей, таких, что их массивы данных являются **открытыми** множествами, пополняемыми в результате производимых экспериментальных исследований. Очевидно, что этот принцип весьма актуален для анализа постоянно пополняемых клинических данных, на основе которых принимаются врачебные решения. Ниже охарактеризуем ДСМ-исследования, используя результаты [8, 9].

Идея осуществления ДСМ-исследований состоит в обнаружении сохраняющихся регулярностей в порождении гипотез при последовательном расширении начальной базы фактов $БФ_0$, таком что $БФ_0 \subset БФ_1 \subset \ldots \subset БФ_s$.

Рассмотрим $БФ_i$, где i=0,1,…,s. $БФ_i$ представима посредством утверждений $J_{(\nu,0)}\left(C \Rightarrow {}_1Q\right)$, где ν=1,-1,τ, а $БФ_i = БФ_i^{+} \cup БФ_i^{-} \cup БФ_i^{\tau}$; $J_{(1,0)}\left(C_j \Rightarrow {}_1Q\right)$, $J_{(-1,0)}\left(C_k \Rightarrow {}_1Q\right)$, $J_{(\tau,0)}\left(C_h \Rightarrow {}_1Q\right)$ – представления (описания) элементов $БФ_i^{+}$, $БФ_i^{-}$, $БФ_i^{\tau}$, соответственно. Будем считать, что $БФ_0^{\tau} = БФ_1^{\tau} = \ldots = БФ_S^{\tau}$, а число элементов $БФ_0^{\tau}$ есть $\left|БФ_0^{\tau}\right| = m_0$. Пусть $\Omega^{\sigma}(p)$ – множество представлений $БФ_i^{\sigma}$, где $\sigma = +,-,\tau$, а p=0,1,…, s.

Пусть, далее, $\Delta^{\sigma}(p)$, где $\sigma = +,-,\tau$, а i=0,1,…,s, есть множество формул вида $J_{(\nu,n)}\left(C' \Rightarrow {}_2Q\right)$, где ν=1,-1,0,τ, $n \ge 1$, а C' есть подмножество сведений некоторых историй болезней, вынуждающее эффект Q. Если имеется сахарный диабет панкреатогенного генеза, то ν=1; если не имеется такового, то ν=–1; если имеются аргументы «за» и «против», то ν=0 (неопределенность).

Пусть пара $\left\langle C_i', Q\right\rangle$ соответствует гипотезе $J_{(\nu,n)}\left(C_i' \Rightarrow {}_2Q\right)$, где $\upsilon = 1,-1,0,\tau$, $n \ge 1$, а Δ(0), … , Δ(s) – множества гипотез, порожденные для $БФ_0$, $БФ_1$, …, $БФ_s$, соответственно. Пусть k – число всех гипотез о причинах, представимых посредством $V \Rightarrow {}_2W$, порожденных для всех $БФ_p$, тогда каждому множеству Δ(p), где $0 \le p \le s$ принадлежат гипотезы $J_{(\nu,n_p)}\left(C_i' \Rightarrow {}_2Q\right)$, где $\upsilon = 1,-1,0,\tau$, а i=1,…,k для пар $\left\langle C_i', Q\right\rangle$, или же в Δ($p$) не имеется этой гипотезы для $\left\langle C_i', Q\right\rangle$.

Определим теперь функцию $\sigma_i(p)$, такую, что $\sigma_i(p)=\nu$, если $J_{(\nu,n_p)}\left(C_i' \Rightarrow {}_2Q\right)$ принадлежит Δ(p); положим, что $\sigma_i(p)=*$, если $J_{(\nu,n_p)}\left(C_i' \Rightarrow {}_2Q\right)$ не принадлежит Δ(p) (будем говорить тогда, что для соответст-

вующего значения p $\sigma_i(p)$ не определена). Таким образом, имеем

$$\sigma_i(p)=\begin{cases}\nu, \textit{если } J_{(\nu,n_p)}\left(C_i'\Rightarrow {}_2Q\right) \textit{принадлежит } \Delta(p)\\ *, \textit{если } J_{(\nu,n_p)}\left(C_i'\Rightarrow {}_2Q\right) \textit{не принадлежит } \Delta(p)\end{cases},$$

где ν = 1,-1,0,τ, а i = 1,…,k.

Каждой паре $\langle C_i',Q\rangle$ соответствует упорядоченное множество $\pi_i^{(2)}=\langle\sigma_i(0),\sigma_i(1),...,\sigma_i(p),...,\sigma_i(s)\rangle$, где $1\le i\le k$, $0\le p\le s$, а $\sigma_i(p)$ имеет значения 1,-1, 0,τ,*. Множество $\pi_i^{(2)}$ образует табл. 1.

Таблица 1

	Δ(0)	…	Δ(p)	…	Δ(s)
$\langle C_1',Q\rangle$	$\sigma_1(0)$	…	σ_1(p)	…	σ_1(s)
⋮	⋮		⋮		⋮
$\langle C_i',Q\rangle$	$\sigma_i(0)$	…	σ_i(p)	…	σ_i(s)
⋮	⋮		⋮		⋮
$\langle C_k',Q\rangle$	$\sigma_k(0)$	…	σ_k(p)	…	σ_k(s)

В каждой строке табл. 1 для пар $\langle C_i',Q\rangle$, i = 1,…, s, содержатся истинностные значения $\overline{\nu}=\langle\nu,n_p\rangle$, где ν = 1,-1,0,τ, а n_p – число применений правил вывода для получения гипотезы о причине $J_{(\nu,n_p)}\left(C_i'\Rightarrow {}_2Q\right)$, соответствующей паре $\langle C_i',Q\rangle$ (C_i' – «тело» причины, Q – эффект).

Необходимым условием распознавания эмпирической закономерности (ЭЗК) является сохранение типов истинностных значений «1» (фактическая истина: сахарный диабет панкреатогенного генеза возникает) или «-1» (фактическая ложь: сахарный диабет панкреатогенного генеза не возникает) для гипотез о (±)-причинах [9].

Если $\sigma_i(p)$=1 или $\sigma_i(p)$=-1 для всех p, где p=0,1,…,s, то для пары $\langle C_i',Q\rangle$ имеет место необходимое условие обнаружения **эмпирического закона** (ЭЗ$^+$ или ЭЗ$^-$). Это означает, что $\pi_i^{(2)}=\langle 1,...,1,...,1\rangle$ или $\pi_i^{(2)}=\langle -1,...,-1,...,-1\rangle$.

Для определения основания (причин) эмпирического закона (ЭЗ$^+$, ЭЗ$^-$) [10] введем в рассмотрение вспомогательную функцию ξ_2^σ **несохранения** типов истинностных значений «1» и «-1» при расширении баз фактов $БФ_0$, $БФ_1$, …, $БФ_s$, соответствующих Δ(0), …, Δ(s).

Определим множества $B_2^+(i)$ и $D_2^+(i)$, где i=1,…,k, таких, что они сохраняют «1» ($\sigma_i(p)$=1) и не сохраняют «1» ($\sigma_i(p)\ne 1$), соответственно:

$$B_2^+(i)=\left\{\sigma_i(p)\middle|\left(\sigma_i(p)=1\right)\&\left(1\le i\le k\right)\&\left(0\le p\le s\right)\right\},$$

$$D_2^+(i)=\left\{\sigma_i(p)\middle|\left(\sigma_i(p)\ne 1\right)\&\left(1\le i\le k\right)\&\left(0\le p\le s\right)\right\}.$$

Аналогично определяются для $\nu=-1$ $B_2^-(i)$ и $D_2^-(i)$ – условия сохранения и несохранения «-1», соответственно.

Определим теперь функции несохранения типов истинностных значений «1» и «-1» для последовательности расширений $БФ_0$, $БФ_1$, …, $БФ_s$

$$\xi_{2,i}^+(V,W)=\frac{\left|D_2^+(i)\right|}{\left|B_2^+(i)\right|},$$ где i=1,…,k,

а $\left|D_2^+(i)\right|$ и $\left|B_2^+(i)\right|$ – числа сохранения и несохранения «1». Аналогично определим $\xi_{2,i}^-(V,W)=\frac{\left|D_2^-(i)\right|}{\left|B_2^-(i)\right|}$.

Определим теперь функции сохранения типов истинностных значений «1» и «-1», соответственно, для последовательности расширений $БФ_0$:

$$\tilde{\xi}_{2,i}^\sigma(V,W)=\begin{cases}1-\xi_{2,i}^\sigma(V,W), \textit{если } \left|B_2^\sigma(i)\right|>\left|D_2^\sigma(i)\right|\\ \#, \textit{если } \left|B_2^\sigma(i)\right|\le\left|D_2^\sigma(i)\right|\end{cases},$$

где σ=+,–; i=1,…,k, а # - указание на то, что функции $\tilde{\xi}_{2,i}^\sigma(V,W)$ не определены для соответствующих значений i и $\langle V,W\rangle$.

Определим также функции сохранения истинностных значений для последовательности расширений $БФ_0$ $БФ_0\subset БФ_1\subset\ldots\subset БФ_s$ для предиката («Z – обладает множеством свойств W») $\tilde{\xi}_{1,i}^\sigma(Z,W)$. Напомним, что $\Omega^\sigma(i)$ – множество представлений $БФ_i^\sigma$ посредством высказываний, где σ=+,–,0,τ, i = 0,1,…,s, а $\Omega^\sigma(i)$ есть множество формул вида $J_{(\nu,n)}\left(C\Rightarrow {}_1Q\right)$, где ν = 1,-1,0,τ, n≥0, а C – история болезни.

Пусть пара $\langle C_i,Q\rangle$ соответствует факту или гипотезе $J_{(\nu,n)}\left(C_i\Rightarrow {}_1Q\right)$, где ν = 1,-1,0,τ, n≥2, а Ω(0), …, Ω(s) – множество гипотез, порожденных для $БФ_0$, …, $БФ_s$, соответственно.

Пусть, далее, $\langle C_1,Q\rangle$, …, $\langle C_{m_0},Q\rangle$ – множество пар, соответствующее $БФ_0^\tau\left(БФ_0^\tau\subset БФ\right)$, тех элементов $БФ_0$, которые предложены для предсказания. Следовательно, их представления $J_{(\tau,0)}\left(C_j\Rightarrow {}_1Q\right)$, где j=1,…,m_0 образуют множество Ω^τ, а $\left|БФ_0^\tau\right|=\left|\Omega^\tau(0)\right|=m_0$.

Рассмотрим множество гипотез, порожденных посредством правил вывода по аналогии (п.п.в.-2) для последовательности расширений $БФ_0$:

$$БФ_0\subset БФ_1\subset\ldots\subset БФ_s.$$

Эти гипотезы являются предсказаниями наличия или отсутствия эффекта Q – развития сахарного диабета у больных хроническим панкреатитом. Они представимы посредством формул $J_{(\nu,0)}\left(C_j\Rightarrow {}_1Q\right)$, где ν = 1,-1,0,τ[5].

Каждой $БФ_p$, где p=1,…,s соответствует множество гипотез (предсказаний) Ω(p) таких, что их элемен-

[5] Если ν=τ, то гипотеза $J_{(\tau,0)}\left(C_j\Rightarrow {}_1Q\right)$, где τ – тип истинностного значения «неопределенность», выражает отсутствие предсказаний.

тами является $J_{(\nu,n_p)}\left(C_j \Rightarrow {}_2Q\right)$, где ν = 1,-1,0,τ, а j=1,…,m_0.

Определим функции $\theta_j(p)$, где j=1,…,m_0, а p=1,…,s. Областью определения $\theta_j(p)$ являются гипотезы из Ω(p), порожденные посредством п.п.в.-2 для $БФ_p$. $\theta_j(p) = \nu$, если $J_{(\nu,n_p)}\left(C_j \Rightarrow {}_2Q\right)$ принадлежит Ω(p).

Таким образом, каждой паре $\langle j,p\rangle$ соответствует некоторое ν, где ν = 1,-1,0,τ, для $J_{(\nu,n_p)}\left(C_j \Rightarrow {}_2Q\right)$ (n_p – число применений правил вывода ДСМ-рассуждения). Получаем табл. 2.

Таблица 2

	Ω (0)	…	Ω (p)	…	Ω (s)
$\langle C_1,Q\rangle$	$\theta_1(0)$	…	$\theta_1(p)$ θ$_1$(p)	…	$\theta_1(s)$
⋮	⋮		⋮		⋮
$\langle C_j,Q\rangle$	$\theta_j(0)$	…	$\theta_j(p)$	…	$\theta_j(s)$
⋮	⋮		⋮		⋮
$\langle C_{m_0},Q\rangle$	$\theta_{m_0}(0)$	…	$\theta_{m_0}(p)$	…	$\theta_{m_0}(s)$

Для определения функции сохранения предсказания эффекта [10] введем в рассмотрение вспомогательную функцию $\xi^{\sigma}_{1,j}$ **несохранения** типов истинностных значений «1» и «–1» при расширении баз фактов $БФ_0$, …, $БФ_s$, соответствующих Ω(0), …, Ω(s).

Аналогично $B^{\sigma}_2(i)$ и $D^{\sigma}_2(i)$ определим $B^{\sigma}_1(j)$ и $D^{\sigma}_1(j)$ всех θ$_j$(p), где j=1,…,m$_0$, таких, что они сохраняют «1» (θ$_j$(p)=1) и не сохраняют «1» (θ$_j$(p)≠1), соответственно:

$$B^{+}_1(j)=\left\{\theta_j(p)\middle|\left(\theta_j(p)=1\right)\&\left(1\le j\le m_0\right)\&\left(0\le p\le s\right)\right\},$$

$$D^{+}_1(j)=\left\{\theta_j(p)\middle|\left(\theta_j(p)\ne 1\right)\&\left(1\le j\le m_0\right)\&\left(0\le p\le s\right)\right\}.$$

Аналогично определяются для $\nu=-1$ $B^{-}_1(j)$ и $D^{-}_2(j)$ условия сохранения и несохранения «–1», соответственно.

Определим функции $\xi^{\sigma}_{1,j}(V,W)$:

$$\xi^{+}_{1,j}(Z,W)=\frac{\left|D^{+}_1(j)\right|}{\left|B^{+}_1(j)\right|},\ \text{где } \sigma=+,-,\ j=1,\ldots,m_0.$$

Функции $\tilde{\xi}^{\sigma}_{1,j}(Z,W)$ сохранения типов истинностных значений «1» и «–1» определим для последовательности расширяющихся баз фактов $БФ_0$, …, $БФ_s$ следующим образом:

$$\tilde{\xi}^{\sigma}_{1,i}(Z,W)=\begin{cases}1-\xi^{\sigma}_{1,i}(Z,W), \textit{если } \left|B^{\sigma}_1(j)\right|>\left|D^{\sigma}_1(j)\right|\\ \#, \textit{если } \left|B^{\sigma}_1(j)\right|\le\left|D^{\sigma}_1(j)\right|\end{cases},$$

где σ=+,–, j = 1,…,m_0.

Вторым необходимым условием распознавания эмпирической закономерности является сохранение типов истинностных значений «1» и «–1» для гипотез, предсказывающих, соответственно, наличие или отсутствие сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом.

Аналогично табл. 1 формируется табл. 2 со строками $\pi^{(i)}_j=\left\langle\theta_j(0),\theta_j(1)...,\theta_j(p),...,\theta_j(s)\right\rangle$, где j = 1,…,m_0.

Эмпирическим законам соответствуют строки табл. 2 такие, что θ$_j$(p)=1 или θ$_j$(p)= –1 для всех p=1,…,s:

$$\pi^{(1)}_j=\langle 1,...,1,...,1\rangle \text{ или } \pi^{(1)}_j=\langle -1,...,-1,...,-1\rangle$$

Обозначения $\tilde{\xi}^{\sigma}_{2,i}(V,W)$ и $\tilde{\xi}^{\sigma}_{1,i}(Z,W)$ использовались для обозначения выразимости этих функций посредством табл. 1 и 2. Опустив индексы j и i, получим общие обозначения функций сохранения «1» и «–1» $\tilde{\xi}^{+}_1(Z,W)$, $\tilde{\xi}^{+}_2(V,W)$ и $\tilde{\xi}^{-}_1(Z,W)$, $\tilde{\xi}^{-}_2(V,W)$.

Определим теперь понятие эмпирических законов (позитивного и негативного) посредством предикатов $L^{+}_{1,2}(\mathrm{V},Z,W)$ и $L^{-}_{1,2}(\mathrm{V},Z,W)$, соответственно. С этой целью определим вспомогательные предикаты сохранения типов истинностных значений для гипотез о (±)-причинах и гипотез о предсказании изучаемых эффектов (в случае нашего исследования им является возникновение (невозникновение) сахарного диабета панкреатогенного генеза.

$$L^{+}_2(V,W,s)=\left(\tilde{\xi}^{+}_2(V,W)=1\right)\&\left(\overline{\rho}^{+}\le\rho^{+}(s)\le 1\right),$$

$$L^{-}_2(V,W,s)=\left(\tilde{\xi}^{-}_2(V,W)=1\right)\&\left(\overline{\rho}^{-}\le\rho^{-}(s)\le 1\right),$$

где $\overline{\rho}^{\sigma}$ – выбранные пороги (обычно $\overline{\rho}^{\sigma}$=0.8), а σ=+,–.

$L^{\sigma}_2(V,Z,s)$ выражают условия сохранения типов истинностных значений «1(–1)» для гипотез о (±)-причинах, представленные в табл. 1.

Определим также

$$L^{+}_1(Z,W,s)=\left(\tilde{\xi}^{+}_1(Z,W)=1\right)\&\left(\overline{\rho}^{+}\le\rho^{+}(s)\le 1\right),$$

$$L^{-}_1(Z,W,s)=\left(\tilde{\xi}^{-}_1(Z,W)=1\right)\&\left(\overline{\rho}^{-}\le\rho^{-}(s)\le 1\right).$$

$L^{\sigma}_1(Z,W,s)$ выражает условие сохранения типов истинностных значений для гипотез о предсказаниях наличия (отсутствия) изучаемого эффекта Q (в нашем исследовании – это наличие или отсутствие развития сахарного диабета панкреатогенного генеза).

Введем теперь определения для позитивных и негативных эмпирических законов, используя предикаты сохранения типов истинностных значений $L^{\sigma}_{1,2}(V,Z,W)$, где $\sigma=+,-$.

Необходимым и достаточным условием обнаружения эмпирического закона является одновременное сохранение типов истинностных значений «1» («–1») и выполнимость неравенств $\overline{\rho}^{\sigma}\le\rho^{\sigma}(s)\le 1$ для последней базы фактов $БФ_s$, где $\sigma=+,-$, $\rho^{-\sigma}$ – выбранный порог, а $\rho^{\sigma}(s)$ – степень абдуктивного объяснения (σ)-примеров из $БФ_s$ такая, что $\rho^{\sigma}(s)=\dfrac{\left|\widetilde{БФ}^{\sigma}_S\right|}{\left|БФ^{\sigma}_S\right|}$, где $\left|\widetilde{БФ}^{\sigma}_S\right|$ – число объясненных (σ)-примеров из $БФ_s$, а $\left|БФ^{\sigma}_S\right|$ – число всех (σ)-примеров в $БФ_s$ [3,4,2].

Выполнимость этих неравенств означает, что рассматриваемая последовательность расширяемых баз фактов $БФ_0$, $БФ_1$, …, $БФ_s$ существенным образом содержит неявно знания о каузальном вынуждении изучаемого эффекта (он является значением переменной *W*) посредством некоторой причины (она является значением переменной *V*) при реализации ДСМ-рассуждения для фиксированной стратегии $Str_{x,y}$ из множества возможных стратегий $\overline{Str}$.

Итак, определим предикаты $L^{\sigma}_{1,2}(V,Z,W)$ ($\sigma=+,–$) – «V –причина эффекта W такая, что она содержится в носителе эффекта Z; одновременно как гипотеза о причине, так и гипотеза о предсказании эффекта W сохраняют тип истинностного значения «1»(« –1»)» для последовательности расширения $БФ_0$, $БФ_1$, …, $БФ_s$ ($БФ_0 \subset БФ_1 \subset \ldots \subset БФ_s$)

$$L^{\sigma}_{1,2}(V,Z,W) = \exists s\left(L^{\sigma}_{2}(V,W,s)\,\&\,L^{\sigma}_{2}(Z,W,s)\,\&\,(V\subset Z)\right),$$

где $\sigma=+,–$, а $\exists s$ означает, что «существует s» такое, что имеет место условие, выразимое посредством одновременной выполнимости $L^{\sigma}_{2}(V,W,s)$, $L^{\sigma}_{1}(V,W,s)$ и $(V\subset Z)$.

Определения эмпирических законов ($ЭЗ^{\sigma}$) может быть усилено посредством добавления условия **монотонного возрастания** функций абдуктивного объяснения базы фактов $\rho^{\sigma}(i)$, соответствующих $БФ_i$,

где $БФ_0 \subset \ldots \subset БФ_i \subset \ldots \subset БФ_s$, $\rho^{\sigma}(i) = \dfrac{\left|\widetilde{БФ}^{\sigma}\right|}{\left|БФ^{\sigma}\right|}$,

где i=0,1,…,s:

$(*)^{\sigma}$ $\rho^{\sigma}(0) \le \rho^{\sigma}(1) \ldots \le \rho^{\sigma}(s)$, где $\sigma=+,–$. Тогда получим следующие определения для предикатов, представляющих $ЭЗ^{\sigma}$, где $\sigma=+,–$:

$$L^{\sigma}_{1,2}(V,Z,W) = \exists s\left(\left(\rho^{\sigma}(0)\le\ldots\le\rho^{\sigma}(s)\right)\&\left(\overline{\rho}^{\sigma}\le\rho^{\sigma}(s)\le 1\right)\&\right.$$
$$\left.\&\left(\tilde{\xi}^{\sigma}_{1}(V,W)=1\right)\&\left(\tilde{\xi}^{\sigma}_{2}(V,W)=1\right)\&(V\subset Z)\right).$$

Иначе:

$$\hat{L}^{\sigma}_{1,2}(V,Z,W) = \exists s\left(L^{\sigma}_{2}(V,W,s)\,\&\,L^{\sigma}_{1}(Z,W,s)\,\&\,(V\subset Z)\,\&\right.$$
$$\left.\&\left(\rho^{\sigma}(0)\le\ldots\le\rho^{\sigma}(s)\right)\right),$$

где $\sigma=+,–$.

Следовательно, определение $ЭЗ^{\sigma}$ посредством $L^{\sigma}_{1,2}(V,Z,W)$ усиливается добавлением монотонного возрастания функций $\rho^{\sigma}(i)$ степени абдуктивного объяснения $БФ_i$,

Таким образом, $L^{\sigma}_{2}(V,W,s)$, $L^{\sigma}_{1}(V,W,s)$ и $(V\subset Z)$ выражают регулярности, сохраняющиеся при расширении базы фактов $БФ_0$, обнаруженные посредством ДСМ-рассуждения, реализующего стратегию $Str_{x,y}$.

Рассмотрим случай позитивного предиката $L^{+}_{1,2}(V,Z,W)$, значениями переменных *V*, *Z*,*W* которого являются, соответственно, *C'*,*C*,*Q*, где $C'\subset C$, *C* – история болезни, а *Q* – наличие развития сахарного диабета панкреатогенного генеза.

(1) $\exists V \exists Z L^{+}_{1,2}(V,Z,Q)$ – форма представления эмпирического закона «развитие сахарного диабета панкреатогенного генеза»,

(2) $\exists Z L^{+}_{1,2}(C',Z,Q)$ – *Q* эмпирический закон, причиной которого является *C'*;

(3) $L^{+}_{1,2}(C',C,Q)$ – реализация эмпирического закона Q посредством причины *C'* и её носителя *C* ($C\subset C'$)

(4) Упорядоченное множество $\left\langle C',Q,Str_{x,y}\right\rangle$ назовем механизмом каузального вынуждения эффекта *Q* посредством причины *C'* и стратегии ДСМ-рассуждения $Str_{x,y}$.

Введенные определения характеризуют сильный вариант эмпирических закономерностей (ЭЗК) – **эмпирические законы (ЭЗ)**, их реализации и механизмы каузального вынуждения изучаемых эффектов. Ослабленным вариантом ЭЗК являются эмпирические тенденции, которые формулируются заменой условий $\tilde{\xi}^{\sigma}_{j}(V,W)=1$, где j=1,2, σ=+,–, в определениях $L^{\sigma}_{1}(Z,W,s)$ и $L^{\sigma}_{2}(V,W,s)$, входящих в определение $ЭЗ^{\sigma}$ посредством $L^{\sigma}_{1,2}(V,Z,W)$, на условия $\overline{\rho}^{\sigma}\le\tilde{\xi}^{\sigma}_{j}(V,W)<1$, где $\overline{\rho}^{\sigma}$ – выбранный порог (например, 0.8).

Таким образом, эмпирические тенденции (ЭТ) $L^{\sigma}_{3,4}(V,Z,W)$ определяются посредством предикатов $L^{\sigma}_{3}(Z,W,s)$ и $L^{\sigma}_{4}(V,W,s)$, образованных из $L^{\sigma}_{1}(Z,W,s)$ и $L^{\sigma}_{2}(V,W,s)$, соответственно, заменой $\tilde{\xi}^{\sigma}_{1}(V,W)=1$ на $\overline{\rho}^{\sigma}\le\tilde{\xi}^{\sigma}_{1}(V,W)<1$ и $\tilde{\xi}^{\sigma}_{2}(V,W)=1$ на $\overline{\rho}^{\sigma}\le\tilde{\xi}^{\sigma}_{2}(V,W)<1$. Тогда ЭТ определяется посредством $L^{\sigma}_{3,4}(V,Z,W)$ следующим образом:

$$L^{\sigma}_{3,4}(V,Z,W) = \exists s\left(L^{\sigma}_{4}(V,W,s)\,\&\,L^{\sigma}_{3}(Z,W,s)\,\&\,(V\subset Z)\right).$$

Заметим, что для определения ЭТ посредством соответствующих замен в определениях ЭЗ необходимо и достаточно сделать замену на $\overline{\rho}^{\sigma}\le\tilde{\xi}^{\sigma}_{1}(V,W)<1$ или на $\overline{\rho}^{\sigma}\le\tilde{\xi}^{\sigma}_{2}(V,W)<1$ (три возможных варианта).

Таким образом, применение ДСМ-метода автоматической поддержки научных исследований состоит из двух стадий – **ДСМ-рассуждений** и **ДСМ-исследований**. Последние посредством ДСМ-рассуждений реализуют процедуры обнаружения эмпирических закономерностей – эмпирических законов (ЭЗ) или эмпирических тенденций (ЭТ).

В нашем исследовании были обнаружены только эмпирические законы с соответствующими каузальными механизмами – гипотезами о причинах наличия (отсутствия) развития сахарного диабета панкреатогенного генеза у больных хроническим панкреатитом. Кроме того, было предсказано [11], что при изменении негативных фактов исключением примеров с наличием диабета непанкреатогенного генеза порождаются $ЭЗ^{\sigma}$ посредством $\hat{L}^{\sigma}_{1,2}(V,Z,W)$, выражаю-

щего монотонное возрастание $\rho^{\sigma}(i)$. Следовательно, отрицательные примеры с диабетом непанкреатогенного генеза порождают «шум», препятствуя обнаружению ЭЗ посредством усиленных предикатов $\hat{L}^{\sigma}_{1,2}(V,Z,W)$, распознающих усиленный механизм каузального вынуждения эффекта развития сахарного диабета у больных хроническим панкреатитом.

Научное исследование достигает уровня зрелости, если его результаты содержат обнаруженные эмпирические закономерности. Они же могут быть извлечены из расширяемых массивов данных посредством компьютерных интеллектуальных систем, применение которых поддерживает научные исследования, использующие автоматизированные рассуждения.

ДСМ-метод автоматизированной поддержки научных исследований (ДСМ-метод АПНИ) одной из своих компонент имеет компьютерные интеллектуальные системы (ИС-ДСМ), имеющие следующее строение: ИС-ДСМ=(БФ,БЗ) – Решатель задач – Комфортный интерфейс, где БФ – расширяемая база фактов, БЗ – пополняемая база знаний. В нашем исследовании это порожденные гипотезы о (±)-причинах, предсказания и обнаружения ЭЗ: установленное сохранение типов истинностных значений (ν=1, ν=–1), представленные посредством $L^{+}_{1,2}(V,Z,W)$ и $L^{-}_{1,2}(V,Z,W)$.

Пятым принципом качественного анализа данных является несингулярное оценивание результатов исследований посредством шкал оценки результатов для сравнения рассуждений и гипотез, порожденных ДСМ-рассуждениями и ДСМ-исследованиями [8]. Параметрами, образующими шкалы оценок качества рассуждений являются вид стратегий ДСМ-рассуждений $\mathrm{Str}_{x,y}$ (они упорядочены по «логической силе»), наличие (отсутствие) эмпирических закономерностей (ЭЗ или ЭТ), степень достоверности предсказания, три вида ошибок определяемые далее.

Посредством m_0 и l_0 обозначаются число высказываний, предложенных на прогноз и верных предсказаний, соответственно. Тогда $\left|БФ_0^{\tau}\right| = m_0$, где $БФ_0^{\tau}$ – фрагмент исходной $\mathrm{БФ}_0$, предложенный для предсказания, m_0 – число его элементов, а $\frac{l_0}{m_0}$ – степень достоверности предсказаний посредством ДСМ-метода АПНИ [8, 9].

Посредством *a*,*b*,*c* обозначим типы ошибок, упорядоченные по «вредности».

(1) a – число грубых ошибок таких, что истинностное значение факта есть $\langle 1,0\rangle$, а п.п.в-2 (правило вывода по аналогии) предсказывает его как $\langle -1,n\rangle$; соответственно, истинностное значение факта есть $\langle -1,0\rangle$, а п.п.в.-2 предсказывает его как $\langle 1,n\rangle$, где n>0, «1», «–1», соответственно, – типы истинностных значений «фактическая истина» и «фактическая ложь» (в нашем исследовании они есть оценки наличия и отсутствия сахарного диабета панкреатогенного генеза, соответственно).

(2) b – число ошибок таких, что истинностное значение факта есть $\langle 1,0\rangle$ или $\langle -1,0\rangle$, а п.п.в.-2 предсказывают его как $\langle 0,n\rangle$, где «0» – тип истинностного значения «фактическое противоречие», а n>0.

(3) c – число ошибок таких, что истинностное значение факта есть $\langle 1,0\rangle$ или $\langle -1,0\rangle$, а п.п.в.-2 предсказывают его как $\langle \tau,n\rangle$, где «τ» – тип истинностного значения «неопределенность» (это означает **отказ** от предсказаний).

Заметим, что $\bar{\nu} = \langle \nu, 0\rangle$ – истинностное значение, ν=1,-1,0,τ, а n – число применений п.п.в.-1 и п.п.в.-2, выражающих степень правдоподобия гипотез (n=0 – означает представление факта), а $Str_{x,y}$ – вид стратегии ДСМ-рассуждений. Получаем **шкалу оценок качества рассуждений** [8]:

(*)

$Str_{x,y}$	ЭЗ	ЭТ	$\frac{l_0}{m_0}$	a	b	c	ρ^{+}	ρ^{-}

Имеется также **шкала оценок качества гипотез о причинах (для $V \Rightarrow {}_2W$) и для предсказаний** $(Z \Rightarrow {}_1W)$:

(**)

$Str_{x,y}$	$\bar{\nu}$	k	ЭЗ	ЭТ	$\Rightarrow_1$	$\Rightarrow_2$

Здесь k – число примеров, породивших гипотезы о (±)-причинах, $\Rightarrow_1$ и $\Rightarrow_2$ – предикаты, используемые для представления гипотез о предсказании и гипотез о причинах, соответственно; а ρ^{+} и ρ^{-} – функции степени принятия гипотез посредством процедуры абдукции, объясняющей $\mathrm{БФ}_0$.

Сформулированные выше шкалы оценок качества ДСМ-рассуждений и порождаемых гипотез являются средствами оценивания результатов ДСМ-исследования и выбора интересных гипотез для экспертных заключений, совпадающих или несовпадающих с оценками, полученными применением интеллектуальной системы ИС-ДСМ к последовательности созданных в нашем исследовании баз фактов $\mathrm{БФ}_0$, $\mathrm{БФ}_1$ и $\mathrm{БФ}_2$, где $\mathrm{БФ}_0 \subset \mathrm{БФ}_1 \subset \mathrm{БФ}_2$.

Далее приведем результаты проведенного исследования.

§3. ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ДЛЯ ОБНАРУЖЕНИЯ РАЗВИТИЯ САХАРНОГО ДИАБЕТА ПАНКРЕАТОГЕННОГО ГЕНЕЗА

Для возможности применения интеллектуальной системы (ИС-ДСМ) к конкретной задаче необходима настройка на предметную область, включающая:

1) разработку языка представления данных;
2) определение понятия «объект» и «свойство» в терминологии ДСМ-метода;
3) задание операции сходства;
4) задание отношения вложения;

Первым этапом была разработка подсистемы представления знаний и создание базы фактов – одной из составных частей ИС-ДСМ. Атрибутами этой БФ являются сведения о больных, представленные в

медицинских картах и описанные в соответствие с разработанным языком представления данных.

Врач перечисляет все возможные признаки описания больного или описания конкретного анализа больного, по которым надо сделать прогноз наличия или отсутствия сахарного диабета панкреатогенного генеза. Задача врача – предоставить достаточное количество примеров положительного и отрицательного типа в терминологии ДСМ-метода, а также не пропустить значимых для диагноза признаков. Задача разработчиков системы – структурировать эти признаки таким образом, что к ним было возможно применять операцию сходства и определить отношение вложения.

(1) Для задачи прогнозирования сахарного диабета панкреатогенного генеза были выделены следующие признаки-атрибуты: возраст больного, индекс массы тела, вредные привычки (алкоголь и табакокурение), длительность заболевания, наличие хронического кальцифицирующего панкреатита, вирсунголитиаза, панкреатическая гипертензия, резекционные или дренирующие операции в анамнезе, панкреонекрозы в анамнезе, гликемический профиль, уровень с-пептида, изменение размеров поджелудочной железы, концентрации гастроинтестинальных гормонов в периферической крови (ГИП, ВИП, соматостатин, секретин, гастрин), наличие развёрнутой клинической картины сахарного диабета и т.д. (всего 39 признаков):

1. Возраст пациента (20-40 лет, 41-60лет, 61 лет и более)
2. Индекс массы тела (16 и менее – выраженный дефицит массы; 16,1-18,5 – недостаточная (дефицит) масса тела; 18,6-25 – норма; 25,1-30 – избыточная масса тела (предожирение); 30,1-35 – ожирение первой степени; 35,1-40 – ожирение второй степени; 40,1 и более – ожирение третьей степени (морбидное))
3. ХКП (хронический кальцифицирующий панкреатит)
4. ХГ,контроль (хронический гастрит)
5. ХХ (хронический холецистит),контроль
6. СРК (синдром раздраженного кишечника), контроль
7. ПН (панкреонекроз)
8. ХП (рак)
9. ЖКБ (желчекаменная болезнь)
10. ДД (дуоденальная дистрофия)
11. IPMN (доброкачественная внутрипротоковая опухоль поджелудочной железы)
12. Цистоаденома
13. Вирсунголитиаз
14. Деформация панкреатического протока
15. Увеличение размеров поджелудочной железы (головка, тело, хвост)
16. Резекционные операции в анамнезе (ГПДР, ПДР, дистальная резекция, операция по Бегеру)
17. Инсулинотерапия
18. Алкоголь в анамнезе
19. Табакокурение в анамнезе
20. Длительность заболевания (до 5 лет, 5-10, 10-15, более 15 лет)
21. Дренирующие операции в анамнезе
22. Боль (0, 1, 2,3)
23. Кисты
24. Кальцинаты
25. Панкреатическая гипертензия
26. Гликемический профиль ммоль/л (меньше 3, 3-4,5, 4,5-6,5, 6,6-8, 8,1-10, больше 10)
27. Нарушение толерантности к глюкозе
28. Экзокринная недостаточность (0, 1, 2)
29. ХЦК (холецистокинин) (меньше 0,15, 0,16-3, 3,1-7, 7,1-10, больше 10)
30. Секретин (меньше 0,023, 0,023-0,223, 0,223-0,423, 0,424-0,623, 0,624-0,823, 0,824-1,023, 1,023-1,26, больше 1,26)
31. ГИП (глюкозозависимый инсулинотропный пептид) (менее 0,002, 0,002-2,002, 2,0021-4,002, 4,0021-6,002, 6,0021-7,9, более 7,9)
32. ВИП (вазоактивный инсулиновый гормон) (менее 0,001, 0,001-4,0, 4,001-8,0, 8,01-12,0, 12,01-16,0, 16,01-20,0, 20,01-24,0, больше 24)
33. соматостатин (меньше 0,02, 0,02-0,42, 0,421-0,62, 0,621-0,82, 0,821-10,2, 10,21-12,2, больше 12,2)
34. гастрин (меньше 0,02, 0,02-20, 20,01-40, 40,01-60, 60,01-80, 80,01-91,5, больше 91,5)
35. с-пептид (0, 0,1-2, 2,1-4, 4,1-6, больше 6)
36. SETR (да, нет)
37. PRSS
38. SPINK
39. ХБК-хронический билиарный панкреатит

В базе фактов выполняются требования структурированности данных. Эффективность анализа результатов различных исследований зависит от максимально полного описания обследований больного.

При анализе этих многочисленных медицинских фактов, представленных в БФ, был выбран один тип данных (№1) [12], а именно: указывается один из возможных качественных признаков, представленных списком:

El(1)= Ati, $i \in \{1,\ldots,n\}$, n-число признаков в списке.

Пример 1.

Увеличение размеров поджелудочной железы

At_1 – головка

At_2 – тело

At_3 – хвост

Пример 2. ХЦК (холецистокинин)

At_1 – меньше 0,15

At_2 –0,16-3

At_3 – 3,1-7

At_4 – 7,1-10

At_5 – больше 10

Пример 3.

Экзокринная недостаточность (0, 1, 2)

At_1 – 0

At_2 – 1

At_3 – 3

Таким образом, был создан язык представления медицинских данных, разработаны программы ввода медицинской информации. Полученная система содержит автоматизированные истории болезни пациентов. Форма ввода медицинских данных очень удобна для врача. Можно просматривать БФ, добавлять информацию о заболевании, по мере ее получения, дополнительно включать результаты различных

исследований. При необходимости возможно наполнение БФ новыми терминами и понятиями.

(2) В нашей задаче объект представляет собой упорядоченный набор из 39 клинических, инструментальных и лабораторных признаков.

$$Oi = <El_1, \ldots, El_{39}>$$

Каждый элемент этого кортежа соответствует конкретному признаку.

Свойство в данном случае одно – сахарный диабет панкреатогенного генеза.

(3) Операция сходства определяется поэлементно.

$O' = <El'_1, El'_2, \ldots, El'_{39}>$

$O'' = <El''_1, El''_2, \ldots, El''_{39}>$,

$O' \prod O'' = <El'_1 \cap El''_1, El'_2 \cap El''_2, \ldots, El'_{39} \cap El''_{39}>$

Для всех перечисленных признаков был выбран тип данных (№1). Результатом операции сходства является совпадение признаков.

$El' = At_i$

$El'' = At_j$

$$El' \cap El'' = \begin{cases} At_i, если\ At_i = At_j \\ \Lambda\ иначе \end{cases}$$

(4) Отношение вложения также определяется поэлементно:

$O' = <El'_1, El'_2, \ldots, El'_{39}>$

$O'' = <El''_1, El''_2, \ldots, El''_{39}>$,

$O' \subseteq O'' \rightleftharpoons El'_1 \subseteq El''_1 \& El'_2 \subseteq El''_2 \& \ldots \& El'_{39} \subseteq El''_{39}$

Отношение вложения для элементов зависит от типа данных. Так как был выбран тип данных №1 вложение равносильно совпадению значений атрибутов

$El' = At_i$, $El'' = At_j$

$El' \subseteq El'' \rightleftharpoons At_i = \Lambda\ V\ At_i = At_j$

Первоначально массив данных БФ включал 91 историю болезни (|БФ|=91), т.е. состоял из 91 факта, из которых 20 были (+)-примерами, 61 – (–)-примерами и 10 – (τ)-примерами. В ходе исследования массив данных пополнился и были составлены два расширения исходного массива (исходный массив будем обозначать $БФ_0$):

1. $БФ_1$, $|БФ_1|$=133 (33 (+)-примера, 90 (–)-примеров, 10 (τ)-примеров),

2. $БФ_2$, $|БФ_2|$=165 (38 (+)-примеров, 115 (–)-примеров, 10 (τ)-примеров).

На этом расширении было проведено первое исследование по анализу эмпирических законов по следующей схеме [11]:

Этап 1. Проводим ДСМ-рассуждения с использованием каждой стратегии $Str_{x,y}$ из $\overline{Str}$ для каждого из расширений $БФ_i$, $i = 0,1,2$;

Этап 2. Проводим ДСМ-исследование для обнаружения ЭЗ

1. Для каждого рассуждения подсчитываем значения объясняемости (ρ^σ) исходного массива (приведены в *Приложениях 1* и *2* (табл. 1) для первого и второго исследования соответственно) и проверяем подформулу из определения ЭЗ: $\tilde{\rho}^\sigma \le \rho^\sigma(s) \le 1$;

2. Для каждого рассуждения все полученные гипотезы (гипотезы о причинах и гипотезы о предсказаниях) группируем в последовательности полученных ими истинностных значений $\nu, \nu \in \{+1, -1, 0, \tau\}$, где + и − – есть сокращения для типов истинностных значений «+1» и «-1». Эти последовательности назовем «цепочками» (множество $\pi_i^{(2)}$). Для каждой из полученных «цепочек» вычисляем значение функций сохранения истинностных значений $\tilde{\xi}_i^\sigma$, $\sigma \in \{+,-\}$, $i = 1,2$.

3. С помощью таблиц, составленных на шаге 2, получаем ЭЗ следующим образом:

Для гипотез о предсказаниях отбираем те гипотезы, которым соответствуют «цепочки» с одинаковыми истинностными значениями на всех расширениях массива (значение $\tilde{\xi}_1^\sigma$ для соответствующего σ равняется 1). Эти гипотезы выполняют предикат $L_1^\sigma(Z,Q)$ ($\sigma \in \{+,-\}$), Q – константа, означающая возникновение сахарного диабета для (+)-примеров и невозникновение сахарного диабета для (–)-примеров);

Для гипотез о причинах отбираем те гипотезы, которым соответствуют цепочки с одинаковыми истинностными значениями на всех расширениях массива (значение $\tilde{\xi}_2^\sigma$ для соответствующего σ равняется 1). Эти гипотезы выполняют предикат $L_2^\sigma(V,Q)$ (), Q – константа, означающая возникновение сахарного диабета для (+)-примеров и невозникновение сахарного диабета для (-)-примеров);

4. Находим пары, выполняющие квазикомпозицию

$$\exists Z\ L_{1,2}^\sigma(V,Z,Q) \rightleftharpoons \exists Z \left(L_1^\sigma(Z,Q) \& L_2^\sigma(V,Q) \& (V \subset Z) \right),$$

где $\sigma \in \{+,-\}$. Получаем множество причин

$$\boldsymbol{V} = \left\{ V \middle| \exists Z L_1^\sigma(Z,W) \& L_2^\sigma(V,W) \& (V \subset Z) \right\}.$$

5. Для каждой стратегии $Str_{x,y}$ из $\overline{Str}$ можем указать множество причин $\boldsymbol{V}'$ ($\boldsymbol{V}' \subseteq \boldsymbol{V}$), которые порождаются в рассуждении с использованием этой стратегии, т.е. построить отображение g: $\overline{Str} \to 2^V$. Это отображение представлено в *Приложениях 1*, *2* (табл. 2 и 3).

6. Для каждой стратегии $Str_{x,y}$ из $\overline{Str}$ подсчитываем значения l_0, a, b, c.

7. Оцениваем причины (механизмы ЭЗ), отобранные на шаге 3, и предсказания, осуществленные на последнем из расширений БФ – $БФ_3$, Все стратегии $Str_{x,y}$ из $\overline{Str}$ ссортируем по количеству правильных предсказаний и количеству ошибок

Проанализировав результаты первого исследования, мы предположили, что, так как (–)-примерами выступали примеры, имеющие разную природу: пациенты, у которых не был поставлен диагноз диабет и пациенты, у которых был поставлен диагноз диабет непанкреатогенного генеза (причем количество примеров со вторым диагнозом составляло небольшой процент от общего количества всех примеров), то удаление этого «шума» из исходного массива может улучшить результаты проводимого исследования.

Таким образом, для второго исследования получили следующую последовательность расширяющихся массивов данных:

1. $БФ_0$, $|БФ_0|$=85 (21 (+)-пример, 54 (–)-примера, 10 (τ)-примеров),

2. $БФ_1$, $|БФ_1|$=124 (34 (+)-примера, 80 (–)-примеров, 10 (τ)-примеров),

3. $БФ_2$, $|БФ_2|$=148 (39 (+)-примеров, 99 (–)-примеров, 10 (τ)-примеров).

Второе исследование было проведено по той же схеме.

И в первом и во втором исследовании наилучшие результаты показала группа, включающая стратегии $Str_{ab,ab}$ и $Str_{ad0b,ab}$. Эти стратегии дают наибольшее количество правильных доопределений l_0(8 из 10), при этом **все правильные предсказания входят в состав ЭЗ**. Во втором исследовании **появились улучшения** для группы стратегий $Str_{ab,ad0b}$ и $Str_{ad0b,ad0b}$: добавилось одно правильное (–)-предсказание (увеличилось значение правильных предсказаний l_0 и, соответственно, уменьшилось значение количества ошибок типа c).

Список гипотез о причинах V был предъявлен для анализа экспертам. Каждую гипотезу предлагалось оценить по надежности числовой оценкой от «4» до «1», где наибольшая оценка означала, что гипотеза представляется наиболее состоятельной, средние оценки – что отобранные признаки могут иметь место, но неоднозначно свидетельствуют о наличии диабета с осложнением, а наименьшая оценка означала, что гипотеза не имеет смысла. В первом исследовании почти все гипотезы были оценены как «содержательные», всего одна (–)-гипотеза (№117182315 в *Приложении* 3) получила наименьшую оценку. **Во втором эксперименте эта гипотеза ушла из списка V**.

Для объясняемости (ρ^σ), полученной на *Шаге 1* ДСМ-исследования, было проверено условие возрастания $\tilde{\rho}^\sigma \le \rho^\sigma(s) \le 1$. В обоих исследованиях объясняемость (ρ^σ) возрастает. Во втором эксперименте выполняется условие монотонного возрастания: $\tilde{\rho}^\sigma \le \rho^\sigma(0) \le \ldots \le \rho^\sigma(n) \le 1$ для всех стратегий из $\overline{Str}$, в то время как в первом исследовании такое условие не выполнялось для $Str_{a,ad0}$ и $Str_{ad0,ad0}$ (наблюдались колебания для $БФ_2^-$ с 1 до 0.99).

Во втором исследовании появились две положительные гипотезы о причинах (№154401(73051), №694862(233212) в *Приложениях* 1, 2, 3), которые входят в состав ЭЗ для **всех** стратегий из $\overline{Str}$, в то время как в первом исследовании таких гипотез не было. Таким образом, во втором исследовании содержатся улучшенные результаты.

Результаты исследований приведены в *ПРИЛОЖЕНИЯХ* в следующем порядке:

В *Приложении 1* и *Приложении 2* представлены результаты первого и второго исследований по одинаковой схеме:

1) В табл. 1 приведены значения абдуктивной сходимости ρ для всех стратегий из $\overline{Str}$;

2) табл. 2 иллюстрирует отображение g: $\overline{Str} \to 2^V$ для позитивных эмпирических зависимостей;

3) табл. 3 иллюстрирует отображение g: $\overline{Str} \to 2^V$ для негативных эмпирических зависимостей;

4) табл. 4 дает информацию о количестве правильных предсказаний и ошибок (по видам) для стратегий из $\overline{Str}$.

В *Приложении 3* приведены все эмпирические зависимости для первого эксперимента, а в *Приложении 4* – экспертная оценка эмпирических зависимостей для первого эксперимента.

СПИСОК ЛИТЕРАТУРЫ

1. Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. – М.: Медиа Сфера, 1998; Fletcher Robert H., Fletcher Suzanne W., Wagner Edward H. Clinical Epidemiology. The Essensials // Williams And Wilkins, 1996.
2. Милль Д.С. Система логики силлогистической и индуктивной. Издание пятое. – М.: ЛЕНАНД, 2011; Mill J.S. A System of Logic Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and The Methods of Scientific Investigation. – London: Parker, Son and Bowin, 1843.
3. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть I // Искусственный интеллект и принятие решений. – 2010. – №3. – С. 3-21; Finn V.K. J.S. Mill's inductive methods in artificial intelligence system. Part 1 // Scientific and Technical Information. – 2011. – Vol. 38, № 6. – P. 385-402.
4. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта. Часть II // Искусственный интеллект и принятие решений. – 2010. – №4. – С. 14-40; Finn V.K. J.S. Mill's inductive methods in artificial intelligence system. Part 2 // Scientific and Technical Information. – 2012. – Vol. 39, № 5. – P. 241-260.
5. Финн В.К. Дистрибутивные решетки индуктивных процедур // Научно-техническая информация. Сер. 2. – 2014. – №11. – С. 1-30; Finn V.K. Distributive Lattices of Inductive JSM Procedures // Automatic Documentation and Mathematical Linguistics. – 2014. – Vol. 48, №6. – P. 265-295.
6. Гершель Дж. Философия естествознания. Издание второе. – М.: Книжный дом «ЛИБРОКОМ», 2011; Herschel J.F.W. Preliminary Discourse on the Study of Natural Philosophy. – London, 1851.
7. Гретцер Г. Общая теория решеток. – М.: «Мир», 1982; Grätzer G. General Lattice Theory // Akademic – Verlag. Berlin,1978.
8. Финн В.К. О классе ДСМ-рассуждений, использующих изоморфизм правил вывода // Искусственный интеллект и принятие решений. – 2016. – №3. – С. 48-61.
9. Шестерникова О.П., Агафонов М.А., Винокурова Л.В., Панкратова Е.С., Финн В.К. Интеллектуальная система прогнозирования развития сахарного диабета у больных хроническим панкреатитом // Искусственный интеллект и принятие решений. – 2015. – № 4. – С. 12-50.
10. Финн В.К. Обнаружение эмпирических закономерностей в последовательностях баз фактов посредством ДСМ-рассуждений // Научно-техни-

ческая информация. Сер. 2. – 2015. – №8. – С.1-29; Finn V.K. Detecting Empirical Regularities in Bases of Facts Using JSM Reasoning // Automatic Documentation and Mathematical Linguistics. – 2015. – Vol. 49, № 4. – P. 122-151.

11. Шестерникова О.П. О применении интеллектуальной системы прогнозирования развития диабета у больных хроническим панкреатитом // Искусственный интеллект и принятие решений. – 2016. – №3. – С. 62-71.

12. Панкратова Е.С, Виноградов Д.В. Формальное описание настройки интеллектуальных ДСМ-систем // Научно-техническая информация. Сер.2. – 2011. – № 9. – С.1-5; Pankratova E.S., Vinogradov D.V. Formal Description of Adaptation of Intelligent JSM-Systems for Clinical and Laboratory Data Analysis // Automatic Documentation and Mathematical Linguistics. – 2011. – Vol. 45, № 5. – P. 213-217.

ПРИЛОЖЕНИЕ 1

Результаты первого исследования

Таблица 1

Значения объясняемости (абдуктивной сходимости ρ^{σ})*

$Str_{x,y}$	Расширения базы фактов					
	$БФ_{0,0}$		$БФ_{0,1}$		$БФ_{0,2}$	
	$БФ^{+}_{0,0}$	$БФ^{-}_{0,0}$	$БФ^{+}_{0,1}$	$БФ^{-}_{0,1}$	$БФ^{+}_{0,2}$	$БФ^{-}_{0,2}$
1	1	1	1	1	1	1
2	0.85	1	0.97	1	0.97	1
3	1	1	1	1	1	1
4	0.85	1	0.97	1	0.97	1
5	1	0.79	1	0.89	1	0.90
6	0.85	0.79	0.97	0.89	0.97	0.90
7	1	0.79	1;	0.89	1	0.96
8	0.85	0.79	0.97	0.89	0.97	0.90
9	1	0.87	1	1	1	0.99
10	0.85	0.93	0.97	1	0.97	1
11	1	0.87	1	1	1	0.99
12	0.85	0.93	0.97	1	0.97	1
13	1	0.66	1	0.68	1	0.68
14	0.85	0.59	0.97	0.68	0.97	0.68
15	1	0.59	1	0.68	1	0.68
16	0.85	0.59	0.97	0.68	0.97	0.68

** В таблице представлены значения объясняемости (абдуктивной сходимости ρ) для $Str_{x,y} \in \overline{Str}$ (по вертикали – порядковыми номерами стратегии) и $БФ_{0,1}$, $БФ_{0,2}$, $БФ_{0,3}$ (по горизонтали – отдельно для положительных и отри-*

Таблица 2

Таблица соответствия положительных эмпирических зависимостей стратегиям (g: $\overline{Str} \to 2^{V}$)*

Идентификаторы ЭЗ$^{+}$ \ $Str_{x,y}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
154401	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
694862	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
239, 1509, 5028, 124738, 126721, 131411, 131880, 167074, 347776, 358062, 372242, 690520, 692023, 693860, 694194, 696198, 697367	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0
3668, 4858, 4943, 130942, 688015, 688182, 688349, 688683, 689017, 693693, 695196	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0

** В клетках таблицы содержатся булевские значения: «1» – «Истина», «0» – «Ложь». Значение «Истина» в клетке на пересечении строки с ЭЗ и столбца со стратегией означает, что эта причина реализована этой стратегией. Идентификаторы есть имена(номера) представления причин (V), см. Приложение 3.*

Таблица 3

Таблица соответствия отрицательных эмпирических зависимостей стратегиям (g: $\overline{Str} \to 2^V$)*

Идентификаторы ЭЗ⁻ \ $Str_{x,y}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
98862, 99640, 100418, 108198,	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
1526151	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2370477	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0
100807, 516144, 539344, 564151, 1549329, 2036199, 2148924, 2161369, 2258245, 2272340, 2282419, 2365792, 3161644, 6674224, 22783570, 22980952, 22987123, 23289353, 23388089, 23394260, 23400431, 26308390, 26525757, 40951161, 117153607, 117182315, 117185119, 117188923, 117923651	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0

** В клетках таблицы содержатся булевские значения: «1» – «Истина», «0» - «Ложь». Значение «Истина» в клетке на пересечении строки с ЭЗ и столбца со стратегией означает, что эта причина реализована этой стратегией. Идентификаторы есть имена(номера) представления причин (V), см Приложение 3.*

Таблица 4

Таблица для применяемого множества стратегий (типы ошибок)*

$Str_{x,y}$	l_0		*a*		*b*		*C*	
	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
6, 8	3	5			1	1		
14, 16	4	2		1				3
2, 4, 10, 12		5			4	1		
13, 15	3	1			1	5		
5, 7	4			3		2		1
1, 3, 9,11		1			4	5		

** По вертикали расположены стратегии, представленные порядковыми номерами. По горизонтали указано количество правильных предсказаний (l_0) и ошибок типа a, b и c отдельно для примеров, верифицирующихся как положительные и как отрицательные. Стратегии сгруппированы и отсортированы по количеству правильных предсказаний и ошибок. Таким образом, наилучшими стратегиями являются $Str_{ab,ab}$ и $Str_{ad0b,ab}$.*

ПРИЛОЖЕНИЕ 2

Результаты второго исследования

Таблица 1

Значения объясняемости (абдуктивной сходимости ρ^{σ})

$Str_{x,y}$	Расширения базы фактов					
	$БФ_{0,0}$		$БФ_{0,1}$		$БФ_{0,2}$	
	$БФ^{+}_{0,0}$	$БФ^{-}_{0,0}$	$БФ^{+}_{0,1}$	$БФ^{-}_{0,1}$	$БФ^{+}_{0,2}$	$БФ^{-}_{0,2}$
1	1	1	1	1	1	1
2	0.95	1	0.97	1	0.97	1
3	1	1	1	1	1	1
4	0.95	1	0.97	1	0.97	1
5	1	0.8	1	0.86	1	0.95
6	0.95	0.8	0.97	0.86	0.97	0.95
7	1	0.8	1;	0.86	1	0.95
8	0.95	0.8	0.97	0.86	0.97	0.95
9	1	0.85	1	1	1	1
10	0.95	0.93	0.97	1	0.97	1
11	1	0.85	1	1	1	1
12	0.95	0.93	0.97	0.99	0.97	1
13	1	0.57	1	0.67	1	0.75
14	0.9	0.57	0.97	0.67	0.97	0.75
15	1	0.57	1	0.67	1	0.75
16	0.9	0.57	0.9	0.68	0.97	0.75

**В таблице представлены значения объясняемости (абдуктивной сходимости ρ) для $Str_{x,y} \in \overline{Str}$ (представлены номерами по вертикали) и $БФ_{0,1}$, $БФ_{0,2}$, $БФ_{0,3}$ (по горизонтали, отдельно для положительных и отрицательных примеров).*

Таблица 2

Таблица соответствия положительных эмпирических зависимостей стратегиям (g: $\overline{Str} \to 2^V$)*

$Str_{x,y}$ / Идентификаторы ЭЗ$^+$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
73051, 233212	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
111406, 123598, 208129, 215778, 228270, 892299, 892705, 897577, 901637, 912968	0	1	0	1	1	1	1	1	0	1	0	1	0	0	0	0
224, 239, 9610, 9868, 9997, 13609, 76812, 91428, 92571, 123979, 213900, 216404, 218908, 219534, 224514, 228896, 891284, 891487, 891893, 893720, 895547, 899607, 899810, 900013, 900825, 902043	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
9352, 9481, 890878, 891690, 892908	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0

** В клетках таблицы содержатся булевские значения: «1» – «Истина», «0» - «Ложь». Значение «Истина» в клетке на пересечении строки с ЭЗ и столбца со стратегией означает, что эта причина реализована этой стратегией*

Таблица 3

Таблица соответствия отрицательных эмпирических зависимостей стратегиям (g: $\overline{Str} \to 2^V$)*

$Str_{x,y}$ / Идентификаторы ЭЗ$^-$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
13465, 306925, 322701, 324091, 954885, 968138	1	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0
303545	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
595216	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
311657, 472011, 476985, 588594, 841042, 1716322, 1717931, 8270657, 8489900, 8497634, 8669678, 8716082, 8750885, 10008740, 10103614, 10183375, 10214311, 10222045, 10318720, 10326454, 14361602, 14905007, 15300932, 17306555, 17747920, 17997826, 18117435, 20886645, 29469330, 30511979, 30784658, 86476672, 86480230, 86484091, 86486046, 86488001, 86489956, 86516190, 86523421, 86526010, 86527965, 86529920, 86531875, 86533830, 86537325, 86539280, 86541235, 86543190, 86548710, 86552291, 86554246, 86572318, 87665560, 87775717, 87831824, 87839628, 88072880	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0

** В клетках таблицы содержатся булевские значения: «1» – «Истина», «0» - «Ложь». Значение «Истина» в клетке на пересечении строки с ЭЗ и столбца со стратегией означает, что эта причина реализована этой стратегией.*

Таблица 4

Таблица для применяемого множества стратегий (типы ошибок)

$Str_{x,y}$	l_0		a		b		c	
	+	-	+	-	+	-	+	-
6, 8	3	5			1	1		
14, 16	4	3		1				2
2, 4, 10, 12		5			4	1		
13, 15	4			2		3		1
5, 7	3	1			1	5		
1, 3, 9,11		1			4	5		

** По вертикали расположены стратегии, представленные порядковыми номерами, по горизонтали указано количество правильных предсказаний (l_0) и ошибок типа a, b и c отдельно для примеров, верифицирующихся как положительные и как отрицательные. Стратегии сгруппированы и отсортированы по количеству правильных предсказаний и ошибок. Таким образом, наилучшими стратегиями являются $Str_{ab,ab}$ и $Str_{ad0b,ab}$.*

Список эмпирических зависимостей

Список позитивных эмпирических зависимостей

Каждая ЭЗ представлена идентификатором (именем, номером) и значением переменных V для причин, включающих признаки, которые характеризуют эту причину.

154401

33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

694862

28- Экзокринная недостаточность	2
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

239

29- ХЦК (холецистокинин)	0,16-3
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
35- с-пептид	0,1-2

1509

29- ХЦК (холецистокинин)	0,16-3
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

3668

29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
35- с-пептид	0,1-2

4858

29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

4943

29- ХЦК (холецистокинин)	0,16-3
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

5028

25- Панкреатическая гипертензия	да
29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

124738

2- Индекс массы тела	16,1-18,5 недостаточная (дефицит) масса тела
29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

126721

28- Экзокринная недостаточность	2
29- ХЦК (холецистокинин)	0,16-3
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

130942

14- Деформация панкреатического протока	есть
29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

131411

14- Деформация панкреатического протока	есть
29- ХЦК (холецистокинин)	0,16-3
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

131880

29- ХЦК (холецистокинин)	0,16-3
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

167074

29- ХЦК (холецистокинин)	0,16-3
34- гастрин	0,02-20
35- с-пептид	0,1-2

347776

18- Алкоголь в анамнезе	да
19- Табакокурение в анамнезе	да
29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

358062

18- Алкоголь в анамнезе	да
19- Табакокурение в анамнезе	да
29- ХЦК (холецистокинин)	0,16-3
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

372242

14- Деформация панкреатического протока	есть
18- Алкоголь в анамнезе	да
19- Табакокурение в анамнезе	да
29- ХЦК (холецистокинин)	0,16-3
35- с-пептид	0,1-2

688015

32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
35- с-пептид	0,1-2

688182

25- Панкреатическая гипертензия	да
35- с-пептид	0,1-2

688349

35- с-пептид	0,1-2

688683

33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

689017

32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

690520

25- Панкреатическая гипертензия	да
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

692023

26- Гликемический профиль ммоль/л	8,1-10
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

693693

14- Деформация панкреатического протока	есть
35- с-пептид	0,1-2

693860

14- Деформация панкреатического протока	есть
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

694194

28- Экзокринная недостаточность	2
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

695196

22- Боль	2
35- с-пептид	0,1-2

696198

34- гастрин	0,02-20
35- с-пептид	0,1-2

697367

33- соматостатин	0,02-0,42
34- гастрин	0,02-20
35- с-пептид	0,1-2

Список негативных эмпирических зависимостей

Каждая ЭЗ представлена идентификатором (именем, номером) и значением переменных V для причин, включающих признаки, которые характеризуют эту причину.

98862

32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

99640

34- гастрин	0,02-20

100418

33- соматостатин	0,02-0,42

108198

29- ХЦК (холецистокинин)	0,16-3

1526151

36- SETR	нет

2370477

35- с-пептид	0,1-2

100807

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42
34- гастрин	0,02-20

516144

20- Длительность заболевания	5-10
28- Экзокринная недостаточность	1
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

539344

26- Гликемический профиль ммоль/л	4,5-6,5
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42

564151

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42

1549329

26- Гликемический профиль ммоль/л	4,5-6,5
33- соматостатин	0,02-0,42

2036199

26- Гликемический профиль ммоль/л	4,5-6,5

2148924

26- Гликемический профиль ммоль/л	4,5-6,5
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
34- гастрин	0,02-20

2161369

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
34- гастрин	0,02-20

2258245

26- Гликемический профиль ммоль/л	4,5-6,5
28- Экзокринная недостаточность	1
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

2272340

26- Гликемический профиль ммоль/л	4,5-6,5
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

2282419

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

2365792

2- Индекс массы тела	18,6 25-норма
26- Гликемический профиль ммоль/л	4,5-6,5

3161644

26- Гликемический профиль ммоль/л	4,5-6,5
36- SETR	нет

6674224

26- Гликемический профиль ммоль/л	4,5-6,5
35- с-пептид	0,1-2

22783570

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42
34- гастрин	0,02-20

22980952

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42

22987123

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

23289353

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
34- гастрин	0,02-20

23388089

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
31- ГИП (глюкозозависимый инсулинотропный пептид)	0,002-2,002
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

23394260

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

23400431

26- Гликемический профиль ммоль/л	4,5-6,5
29- ХЦК (холецистокинин)	0,16-3
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
35- с-пептид	0,1-2

26308390

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
33- соматостатин	0,02-0,42
35- с-пептид	0,1-2

26525757

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
35- с-пептид	0,1-2

40951161

26- Гликемический профиль ммоль/л	4,5-6,5
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0
34- гастрин	0,02-20
35- с-пептид	0,1-2

117153607

20- Длительность заболевания	5-10
28- Экзокринная недостаточность	1
34- гастрин	0,02-20

117182315

20- Длительность заболевания	5-10
28- Экзокринная недостаточность	1

117185119

20- Длительность заболевания	5-10
28- Экзокринная недостаточность	1
29- ХЦК (холецистокинин)	0,16-3

117188923

20- Длительность заболевания	5-10
28- Экзокринная недостаточность	1
33- соматостатин	0,02-0,42

117923651

2- Индекс массы тела	18,6 25-норма
28- Экзокринная недостаточность	1
32- ВИП (вазоактивный инсулиновый гормон)	0,001-4,0

ПРИЛОЖЕНИЕ 4

Экспертная оценка эмпирических зависимостей

Для позитивных ЭЗ

# п/п	Идентификатор $ЭЗ^{+}$	Оценка эксперта
1	154401	3
2	694862	4
3	239	2
4	1509	3
5	3668	3
6	4858	2
7	4943	2
8	5028	3
9	124738	2
10	126721	3
11	130942	2
12	131411	4
13	131880	3
14	167074	2
15	347776	3
16	358062	3
17	372242	3
18	688015	2
19	688182	3
20	688349	2
21	688683	2
22	689017	3
23	690520	3
24	692023	4
25	693693	2
26	693860	4
27	694194	3
28	695196	2
29	696198	2
30	697367	2

Для негативных ЭЗ

# п/п	Идентификатор $ЭЗ^{-}$	Оценка эксперта
1	98862	2
2	99640	2
3	100418	2
4	108198	2
5	1526151	2
6	2370477	2
7	100807	4
8	516144	3
9	539344	3
10	564151	2
11	1549329	2
12	2036199	2
13	2148924	3
14	2161369	3
15	2258245	2
16	2272340	2
17	2282419	2
18	2365792	2
19	3161644	2
20	6674224	2
21	22783570	4
22	22980952	3
23	22987123	3
24	23289353	3
25	23388089	2
26	23394260	2
27	23400431	3
28	26308390	3
29	26525757	2
30	40951161	3
31	117153607	2
32	117182315	1
33	117185119	2
34	117188923	2
35	117923651	2

Материал поступил в редакцию 22.08.16.

Сведения об авторах

АГАФОНОВ Максим Александрович – младший научный сотрудник отделения патологии поджелудочной железы и желчевыводящих путей Центрального научно-исследовательского института гастроэнтерологии Московского клинического научного центра
e-mail: referent509@yandex.ru

ШЕСТЕРНИКОВА Ольга Павловна – начальник отдела разработки ООО «Прогтек», Москва.
e-mail: oshesternikova@gmail.com

ВИНОКУРОВА Людмила Васильевна – доктор медицинских наук, ведущий научный сотрудник отделения патологии поджелудочной железы и желчевыводящих путей ГБУЗ Московского клинического научного центра
e-mail: vinokurova52@mail.ru

ПАНКРАТОВА Елена Сергеевна – кандидат технических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), Москва
e-mail: elena_pankr@inbox.ru

ФИНН Виктор Константинович – доктор технических наук, профессор, заслуженный деятель науки РФ, главный научный сотрудник Института проблем информатики ФИЦ ИУ РАН; руководитель Отделения интеллектуальных систем в гуманитарной сфере Российского государственного университета
e-mail: ira.finn@gmail.com

ИНФОРМАЦИОННОЕ ПИСЬМО И ПРИГЛАШЕНИЕ

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ К 65-ЛЕТИЮ ВИНИТИ РАН

«ИНФОРМАЦИЯ В СОВРЕМЕННОМ МИРЕ»

Москва, 25-26 октября 2017 г.

подробная информация на сайте: http://www.viniti.ru

Главный организатор:

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Соорганизаторы:

Российская академия наук
Федеральное агентство научных организаций
Российский фонд фундаментальных исследований
Министерство образования и науки РФ

Проблемно-тематическое направление конференции: современный издательский процесс, интеллектуальная собственность, научные библиотеки, информационное обеспечение научной и инновационной деятельности, информационные технологии для научной и библиотечной отрасли, информационная безопасность, международное сотрудничество и информационный обмен, инфометрия, классификации, стандартизация, образование для отрасли, экономика информации

Основные вопросы, предлагаемые к обсуждению:

● Популяризация научных знаний: Новые модели распространения научной информации
● Редакционно-издательская деятельность в цифровой среде: продукты и сервисы
● Издательские стандарты и технологии
● Перспективы развития книжного дела. Проекты и программы
● Взаимодействие цифровых и печатных ресурсов в научно-технической библиотеке
● Информационно-библиотечное обслуживание: сервисный подход
● Управление данными и навигация в современной научной библиотеке
● Научные библиотечные консорциумы – основные подписчики на научную литературу
● Перспективы развития национальных систем научно-технической информации
● Государственные проекты и программы поддержки информационного обеспечения научно-образовательной деятельности
● Тенденции развития региональных аналитических центров
● Информационное обеспечение экспертной деятельности. Использование информационно-аналитических систем для управления наукой и образованием
● Формальные и неформальные каналы развития современных научных коммуникаций

● Современные агрегаторы научной литературы открытого доступа как источник научной информации
● Машинная обработка данных и аналитические исследования: Приоритеты и сотрудничество
● Использование специальных сервисов компании CrossRef для идентификации научных публикаций
● Роль поисковых систем в современном издательском процессе
● Защита данных от несанкционированного использования. Маркеры безопасности. Политика безопасности открытых систем
● Вопросы достоверности и доверенности при обработке информационного потока
● Межгосударственный обмен научно-технической информацией на евразийском пространстве
● Информационное взаимодействие в рамках СНГ
● Международное партнерство при хранении и обработке больших массивов данных
● Современное состояние систем классификации знаний как инструмента индексирования и поиска данных по перспективным направлениям науки и критическим технологиям
● Современные библиометрические методы определения научных лидеров: Новые математические модели
● Анализ читательской аудитории научной литературы путем вебометрического анализа
● Подготовка специалистов в сфере научно-информационной деятельности
● Мастер-класс по работе с классификационными системами (УДК, ГРНТИ)
● Информация как источник цифрового капитала и фактор социальных изменений
● Информационная деятельность как фактор национальной экономики
● Новейшие бизнес-модели для публикаций открытого и закрытого доступа

На конференции планируются доклады представителей ведущих информационных центров и научно-технических библиотек России, СНГ и дальнего зарубежья.

В рамках юбилейной конференции состоится научно-практический семинар по классификационным системам «Перспективные направления научных исследований и критические технологии в классификационных системах». Предполагается проведение специализированных обучающих мероприятий по УДК индексированию. Запланировано заседание методического совета пользователей ГРНТИ и УДК. Участники конференции получат свидетельства о повышении квалификации.

Материалы конференции будут опубликованы в журнале **«Научно-техническая информация» Серия 1 и 2,** а также в Трудах конференции. Труды Конференции издаются на CD-ROM и в виде публикаций.

Доклады

Принимаются оригинальные работы, имеющие научное и прикладное значение, соответствующие тематическим направлениям конференции и НЕ ОПУБЛИКОВАННЫЕ ГДЕ-ЛИБО РАНЕЕ.

Предлагаемый доклад должен отвечать следующим требованиям:

1. Необходимо указать название доклада, фамилию, имя, отчество (полностью) авторов/соавторов, название организации, город, страну, выделить автора, который будет представлять доклад.

2. Необходимо наличие аннотации, раскрывающей содержание доклада. Размер аннотации - не более 850 знаков (включая пробелы).

3. Доклады принимаются только в электронной форме; тексты – в формате MS Word; схемы, диаграммы, фотографии, сканированные виды экранов и т. п. - в формате JPG. Объем доклада вместе с аннотацией, рисунками, приложениями и т.п. не более 10 страниц формата А4.

4. Доклад необходимо выслать по электронной почте до 11 сентября 2017 г. в адрес оргкомитета: conf@viniti.ru

Доклады, не соответствующие вышеуказанным требованиям,
НЕ РАССМАТРИВАЮТСЯ.

Программный комитет оставляет за собой право определять статус доклада (пленарный доклад, доклад, стендовый доклад), включать принятые доклады в те или иные секции.

Время для выступления: пленарные доклады – 15–20 мин., доклады на отдельных мероприятиях – до 10 мин. Доклады включаются в Труды на основании решения экспертов оргкомитета.

Контакты: 125190, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 8 (499) 152 61 13, 8 (499) 155 42 52, 8 (499) 151 02 61. Факс 8 (499) 943 00 60
Интернет-сайт: http://www.viniti.ru Эл. почта: conf@viniti.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru

Центр (Отдел) научно-информационного обслуживания (ЦНИО) ВИНИТИ РАН

Информационные услуги, предоставляемые ЦНИО ВИНИТИ РАН:

– проведение тематического поиска и консультации поисковых экспертов;
– подготовка списков научной литературы;
– подбор, копирование полнотекстовых материалов из первоисточников на бумажном носителе и в электронном виде;
– библиометрическая оценка публикационной активности исследователей и научных организаций с использованием российских и зарубежных баз данных;
– информационное обеспечение информационно-аналитической деятельности по подготовке и предоставлению аналитических обзоров и других научных материалов.

ВИНИТИ РАН располагает следующими информационными ресурсами:

– фондом НТЛ, включающим более 2,5 млн. отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1991 года;
– базами данных и Интернет-ресурсами: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД Questel (патенты) и другими реферативными ресурсами;
– полнотекстовыми электронными ресурсами (статьи, патенты, материалы конференций).

Ознакомиться с информацией о доступных полнотекстовых и реферативных ресурсах можно на сайте ВИНИТИ www.viniti.ru

К услугам пользователей – **Электронный Каталог ВИНИТИ http://catalog.viniti.ru** и **служба электронной доставки документов.**

Осуществляется платное информационное обслуживание по разовым заказам и на договорной основе с предоставлением всех необходимых финансовых документов.

Проводится индивидуальное обслуживание пользователей в читальном зале ЦНИО ВИНИТИ.

Обращаться в ЦНИО ВИНИТИ:
- адрес: 125190, Россия, г. Москва, ул. Усиевича, 20;
- телефоны: 8(499) 155 -42 -43, 8(499) 155 -42 -17;
- эл. почта cnio@viniti.ru, fdk@viniti.ru;
- факс 8(499) 930 -60 -00 (для ЦНИО).