

В.К. Финн

Об интеллектуальном анализе данных

Одним из активно развивающихся разделов направления исследований «искусственный интеллект» является анализ данных, получивший названия в англоязычной литературе “data mining” и “knowledge discovery” [1].

Под “knowledge discovery in databases” (обнаружение знаний в базах данных) (КДД) понимают какой-либо нетривиальный процесс идентификации достоверных, новых, потенциально полезных и хорошо понимаемых образцов (структур, patterns) в данных ([1], с. 6).

Под данными понимают множество фактов (разумеется, представленных в соответствии с формулируемыми целями их использования), а под образцами (patterns) понимают некоторые выражения E (формулы языка L), некоторым образом характеризующие подмножество фактов (при этом утверждается, что E не сводится к перечислению подмножества фактов из БД). Предполагается, конечно, что L есть некоторый формальный язык представления знаний.

Под процессом КДД понимают многошаговую систему процедур, включающую подготовку данных, поиск образцов в БД, оценку извлеченного знания, корректировку и итерацию процедур.

Под достоверностью извлеченных образцов понимают сохранение их оценки (некоторой степени достоверности) на новых (расширенных) данных.

Под знанием [1] понимают извлеченный из БД образец E , выраженный в языке L такой, что он имеет оценку $v \geq \rho$, где ρ - заданный порог.

Под Data Mining понимают этап процесса КДД, состоящий в применении специфических алгоритмов порождения образцов E_i , извлеченных из БД (т.е. имеющегося множества фактов). Множество образцов может быть открытым, а их перечисление реализуется специальным алгоритмом.

Основными этапами КДД процесса являются следующие установки и процедуры [1].

1. Выбор предметной области и релевантного знания для реализации целей конечного пользователя компьютерной системы.
2. Выбор исходного множества данных (базы фактов) и подмножества переменных, которые необходимы для извлечения нового знания из базы фактов.
3. Уточнение данных и предпроцессинг: выбор основных операций над данными так, что они могут способствовать уменьшению «шума», определение стратегий для его минимизации.
4. Редукция данных: обнаружение полезных особенностей данных, чтобы представление данных было адекватным решению задач, соответствующих цели КДД.
5. Выбор задачи data mining исследования, т.е. спецификация процесса КДД как классификации, кластеризации и т.д.
6. Выбор алгоритмов, реализующих data mining для поиска образцов (patterns) в данных. Этот выбор должен быть согласован с моделями и параметрами представления данных.

7. Data mining: поиск образцов в форме интересной для пользователя (правила классификации и кластеризации, регрессия, деревья решений и т.д.).
8. Интерпретация порожденных образцов с возможным повторением этапов 1 – 7 для дальнейшей итерации.
9. Обзор и согласование обнаруженного знания.

Таким образом, согласно [1] КДД процесс является использованием data mining алгоритмов для выделения знания (некоторых образцов, извлеченных из баз фактов) в соответствии с формулируемыми критериями принятия результатов при условии необходимого предпроцессинга, формирования выборок из базы фактов и некоторых ее преобразований.

Машинное открытие закономерностей, машинное обучение и, наконец, интеллектуальный анализ данных являются близкими по решаемым задачам и используемым методам направлениями в исследованиях и практических приложениях компьютерной науки. При широком толковании термина «интеллектуальный анализ данных» (ИАД) “data mining” и “knowledge discovery” являются видом ИАД. Это широкое толкование ИАД, по-видимому, состоит в том, что из неупорядоченных и неформализованных данных посредством различных формальных методов, могущих перерабатывать эти данные посредством некоторых алгоритмов в интерпретируемые результаты так, что из них можно извлечь некоторые знания в явном виде такие, что до применения этих методов эти знания были скрыты в массиве данных (т.е. в базе фактов [2]).

Методы извлечения новых знаний из баз фактов, применяемые в ИАД, весьма различны – это и статистические процедуры, генетические алгоритмы, нейронные сети, деревья решений, индуктивное логическое программирование и т.д. Общим обстоятельством в различных реализациях ИАД является то, что данные недостаточно формализованы, но извлекаемость из них посредством компьютерных программ новых полезных знаний возможна.

Относительно возможностей и перспектив развития ИАД возникают следующие естественные вопросы.

1. Можно ли выделить типы предметных областей (моделей) таких, что им будут соответствовать специальные эвристики формализация которых облегчит решение задач ИАД, относящихся к этим предметным областям?
2. Можно ли предложить формальные языки и логические средства ИАД такие, что с их помощью будут формализованы рассуждения, автоматизация которых приведет к программным системам, являющимся инструментом решения задач ИАД для соответствующих областей (моделей)?
3. Существуют ли методы ИАД, способные не только извлекать из данных образцы (в смысле data mining) или некоторые закономерности, но и пригодные для формирования фрагментов теорий, использующих эмпирические данные?¹
4. Существуют ли интеллектуальные системы, пригодные для получения ответов на вопросы 1 – 3, и каково их строение?
5. Каков смысл слова интеллектуальный в названии ИАД – раздела компьютерной науки?

Под словом «интеллектуальный» в названии ИАД будем понимать обнаружение нового (относительно имеющейся базы фактов (БФ) и базы знаний (БЗ)) знания, извлеченного из

¹ Положительный ответ на этот вопрос означает, что возможно автоматизированное формирование фрагментов баз знаний.

БФ и БЗ посредством интеллектуальной системы (ИС). Таким образом, ИАД есть анализ данных посредством ИС. Эта формулировка и есть ответ на вопрос 5.

Однако этот ответ будет информативным лишь при условии уточнения термина «интеллектуальная система» [3].

ИС есть компьютерная система для решения классов задач, которые или не могут быть решены человеком в реальное время, или же их решение требует автоматизированной поддержки, или же их решение дает результаты сопоставимые по информативности с решениями человека. Характеризация компьютерной системы как интеллектуальной будет неполной, если не будут уточнены как природа решаемых задач, так и средства их решения, реализуемые благодаря определенной архитектуре компьютерной системы.

Задачи, решаемые посредством ИС, являются некорректными в том смысле, что они требуют применения формализованной эвристики и не предполагают полноты знаний, являющихся исходными посылками при решении этих задач. Это означает, что применяемые эвристики должны приближенно отображать некоторые аспекты интеллекта [3], а, именно: способности упорядочения данных и знаний с выделением существенных параметров в данных в соответствии с реализуемой целью; способности к рассуждению как синтеза различных познавательных процедур, включающего эмпирическую индукцию, аналогию и дедукцию; способности к выдвижению гипотез («мыслить, следовательно, уметь выдвигать гипотезы» [4] – так считают П. Гаек и Т. Гавранек), способности отвечать на вопрос «почему?», которая может быть реализована посредством абдуктивного объяснения начального состояния базы фактов; способности к обучению на основе позитивных и негативных примеров, способности к фальсификации полученных результатов, и, наконец, способности к адаптации в соответствии с изменением множества фактов и знаний.

Аппроксимация этих способностей в компьютерной системе, которая является интеллектуальной, возможна благодаря ее специальной архитектуре²: ИС=Решатель задач + Информационная среда + Интеллектуальный интерфейс, где

Решатель задач = Рассуждатель + Вычислитель + Синтезатор,

Информационная среда = БФ + БЗ,

Интеллектуальный интерфейс = диалог + представление результатов (включая графику) + научение работе с системой.

Рассуждатель есть подсистема, реализующая логические средства решения задач, посредством которых формализуется соответствующая эвристика. Результатом такой формализации являются различные типы рассуждений, например, приближенные рассуждения, использующие аппарат нечетких множеств (эти рассуждения могут включать генетические алгоритмы); индуктивные рассуждения, использующие антиунификацию, которые осуществляются посредством индуктивного логического программирования, и, наконец, рассуждения, образованные синтезом различных познавательных процедур (например, индукции, аналогии абдукции и дедукции).

Очевидно, что и использование индукции или аналогии достаточно для того, чтобы рассуждения, включающие их, были бы правдоподобными.

² Архитектура ИС не сводится к архитектуре экспертных систем с продукционными правилами.

Эта очевидность вытекает из того факта, что индукция и аналогия не переносят истинность посылок на заключение (перенос истинности с посылок на заключение является принципом дедукции как достоверного вывода). Аналогичное имеет место и для абдуктивного вывода в понимании Ч.С.Пирса [5].

Важно понимать, что правдоподобные рассуждения, формализующие эвристику решения задач, адекватные цели применения ИС, являются основным инструментом ее Решателя реализуемым в Рассуждателе.

В частности, правдоподобные рассуждения являются логическим средством автоматического порождения гипотез на основе БФ и БЗ, что является специфическим видом процесса КДД.

Существенно отметить, что индукция, аналогия и абдукция являются амплиативными выводами в смысле Ч.С.Пирса, т.е. выводами правдоподобными и порождающими новое знание относительно своих посылок, что является необходимым условием когнитивных процедур [3].

Словосочетание «вывод, основанный на знаниях» неявно подразумевает, что имеются в виду либо амплиативные выводы (т.е. «творческие» выводы, реализующие эвристические приемы – некоторые аргументированные догадки), либо дедуктивные выводы, применяемые к информативным посылкам из БФ и БЗ, характеризующим предметную область, благодаря представленным о ней знаниям.

Двумя другими компонентами Решателя являются Вычислитель и Синтезатор. Вычислитель применяется к числовым данным, используя численные методы, релевантные целям ИС (например, таковыми являются различные статистические методы анализа данных, квантовохимические расчеты для прогнозирования биологической активности химических соединений и т.д.).

Синтезатор выбирает стратегии, адекватные не только цели ИС, но и состоянию БФ, и результатам предыдущих применений Решателя.

Второй компонентой ИС является информационная среда, образованная двумя подсистемами – базой фактов (БФ) и базой знаний (БЗ). Попытаемся уточнить эти термины.

БФ – это представление элементарных событий некоторого фрагмента изучаемой предметной области. Каждое элементарное событие – это элемент некоторого отношения. Фрагмент же предметной области характеризуется заданной системой отношений $R_1^{(k_1)}, \dots, R_s^{(k_s)}$, с арностью k_1, \dots, k_s , соответственно.

Факт есть элементарное высказывание p_{ij} языка представления знаний L с некоторой оценкой v_{ij} , представляющее j -ый элемент отношения $R_i^{(k_i)}$, где $i = 1, \dots, s$.

Отметим, что оценка v_{ij} не обязательно является двузначной (истина, ложь), а может быть истинностным значением некоторой неклассической логики (в т.ч. вариантов нечетких логик). Таким образом, БФ есть множество элементарных высказываний p_{ij} с оценкой v_{ij} .

Наличие БФ как подсистемы ИС создает возможность осуществления машинного обучения [6,7], а, следовательно, расширения БЗ.

Если в ИС реализуется машинное обучение на основе БФ (т.е. обучение на положительных и отрицательных примерах изучаемых эффектов), то, возможно, использовать абдуктивное объяснение начального состояния БФ [5,8] не только для принятия порожденных Решателем гипотез, но и для расширения БФ (в случае, если не все факты начального состояния объяснены, ИС применяется к расширенной БФ с целью получить объясняющие БФ гипотезы).

Второй составляющей информационной среды ИС является БЗ – подсистема представления знаний [9].

Объяснение термина «знание» не является ныне сферой лишь философии и философской логики, ибо для создания БЗ требуется характеристика идеи «знание в компьютерной системе» («знание в КС»). Обычно выделяют три типа знаний для КС: декларативные, процедурные и концептуальные.

Под процедурными знаниями понимают задание алгоритмов и их комбинаций, применяемых в Решателе задач для достижения цели. Процедурным знанием являются стратегии решения задач, образованные посредством комбинирования различных видов, рассуждений и вычислений. Под декларативным знанием понимают системы утверждений и, в частности, характеристику предметной области. Таковой являются аксиомы структуры данных (например, булевой) и дескриптивные утверждения, характеризующие предметную область (они могут быть необходимыми условиями корректности результатов применяемых процедур Решателя задач).

Декларативным знанием ИС являются также утверждения, выражающие в имплицитивном виде правила вывода Рассуждателя. Эти утверждения образуют метатеорию ИС и создают возможность исследования на логическом уровне процедур Рассуждателя. Разумеется, это возможно лишь при условии, что имеется алгоритмическая интерпретация предикатов, представляющих процедуры (т.е. правила вывода) и входящих в соответствующие имплицитивные утверждения.

Наконец, концептуальным знанием ИС является множество утверждений и определений понятий, характеризующих принципы создания ИС. Это знание является метатеоретическим, которым руководствуются создатели ИС.

Уточним теперь, три идеи: «представление знаний в ИС», «система знаний для ИС» и «база знаний ИС» (БЗ).

Рассуждения и вычисления, представление знаний и интерфейс являются практическими реализациями принципов функционирования ИС. Посредством этих компонент функционирования ИС осуществляется ИАД.

Под «представлением знаний в ИС» понимают как выбор формы выражения знания посредством некоторого специального языка L , так и содержание, отображающего фрагмент предметной области, введенный в ИС в соответствии с целями, т.е. решаемыми задачами [10].

Наиболее известными формами представления знаний в ИС являются язык логики предикатов 1-го порядка, семантические сети и фреймы [10].

В [11] рассматривался способ представления знаний для машинного обучения в БФ положительных и отрицательных примеров, названный квазиаксиоматическими теориями (КАТ). КАТ состоит из аксиом структуры данных (они необходимы для формализации сходства фактов), аксиом предметной области (они характеризуют ее лишь частично),

процедурных аксиом, представляющих правила правдоподобного вывода; открытого множества фактов (т.е. БФ) и гипотез, порожденных Решателем задач, а также правил вывода, включающих правила правдоподобного вывода. Очевидно, что КАТ есть вид открытой теории, представленной в ИС. Отметим также предложенную Д.А. Пospelовым форму представления знаний, названную им семиотической моделью [12]. Этот способ представлений знаний может быть использован для анализа данных ситуационного управления.

Охарактеризуем теперь идею «система знаний для ИС». Разумеется, что не всякий набор сведений о предметной области пригоден для достижения целей ИС, т.е. для решения множества задач, которое осуществит Решатель. Может быть предложено следующее уточнение идеи «система знаний для ИС».

1⁰. Пусть задана БФ для решения соответствующего класса задач ИС. Каждый факт F из БФ будем называть элементом базового множества знаний для ИС. Если F использовался Решателем (т.е. Рассуждателем или Вычислителем), то будем говорить, что F релевантен цели ИС (относительно данной задачи) (или: реально релевантен цели ИС); Иначе, будем говорить, что F потенциально релевантен цели ИС (относительно класса задач, решаемых в ИС).

2⁰. Пусть заданы множества квантифицированных высказываний Σ^* языка представления знаний L, выразительная сила которого не слабее выразительной силы языка логики предикатов 1-го порядка, и множество Σ^0 неэлементарных бескванторных высказываний L.

$\Sigma = \Sigma^* \cup \Sigma^0$ будем называть базовым множеством знаний для ИС.³

Каждый элемент базового множества знаний Σ по определению релевантен цели ИС.

Если $r \in \Sigma$ и r было использовано Решателем при решении данной задачи, то, будем говорить, что r реально релевантно цели ИС (относительно данной задачи); иначе, будем говорить, что r потенциально релевантно цели (относительно класса задач, решаемых в ИС).

3⁰. Будем называть каркасом ИС пару $\mathbf{R} = \langle \text{БФ} \cup \Sigma, \text{Решатель} \rangle$, а посредством [R] обозначим реализацию каркаса ИС, которой является множество всех результатов применения Решателя для решения данной задачи.

Пусть r высказывание языка представления знаний L такое, что $r \in [R]$, тогда будем говорить, что r релевантно цели ИС.

4⁰. Декларативным знанием ИС будем называть высказывания r такие, что они релевантны цели ИС – это означает, что $r \in \text{БФ}$, или $r \in \Sigma$, или $r \in [R]$.

5⁰. Задание процедуры П Решателя для получения результатов из [R] по определению будем считать релевантным цели ИС.

Задание процедуры П, примененное Решателем в процессе решения задачи ИС будем называть реально релевантным цели ИС; в противном случае будем говорить, что задание процедуры П потенциально релевантно цели ИС.

³ Σ^* или Σ^0 могут быть пустыми множествами, если $\Sigma^* = \emptyset$, то базовое множество для ИС будем называть простым.

Будем говорить, что задание процедуры П Решателя⁴ принадлежит «системе знаний для ИС»; будем говорить также, что задание процедуры П является процедурным знанием ИС. Будем считать, что Решатель ИС определен заданием упорядоченного множества своих процедур П.

6⁰. Знанием в ИС будем называть декларативное и процедурное знание, принадлежащее «системе знаний для ИС». В зависимости от того имеет ли место реальная или потенциальная релевантность элемента «системы знаний для ИС» будем говорить о реальном и потенциальном знании в ИС.

Напомним, что $[R]$ есть реализация каркаса ИС, где $R = \langle \text{БФ} \cup \Sigma, \text{Решатель} \rangle$. Рассматривая Решатель как оператор, естественно положить, что Решатель $(\text{БФ} \cup \Sigma) = [R]$.

Очевидно, что наличие отношения включения между $\text{БФ} \cup \Sigma$ и $[R]$ зависит от типа Решателя. В частности, если БФ является эпистемической базой фактов такой, что она содержит фактоподобные высказывания с оценками, которые являются истинностными значениями некоторой многозначной логики, содержащими оценку «неопределенность», то Решатель преобразует БФ, заменяя оценку «неопределенность» на другие возможные оценки, тогда неверно, что $\text{БФ} \cup \Sigma \subseteq \text{Решатель}(\text{БФ} \cup \Sigma)$, т.е. нарушается условие рефлексивности оператора Решатель. Для класса Решателей, реализующих машинное обучение при порождении всех возможных гипотез (например, для ДСМ - метода автоматического порождения гипотез [8,11]) выполняется условие идемпотентности Решателя:

Решатель (Решатель $(\text{БФ} \cup \Sigma)$) = Решатель $(\text{БФ} \cup \Sigma)$.

Для ИС с Решателем, использующим машинное обучение с индуктивными процедурами и устранением «неопределенности» из БФ не выполняется условие монотонности Решателя: если $\text{БФ}_1 \cup \Sigma_1 \subseteq \text{БФ}_2 \cup \Sigma_2$, то Решатель $(\text{БФ}_1 \cup \Sigma_1) \subseteq \text{Решатель}(\text{БФ}_2 \cup \Sigma_2)$ (о немонотонных операторах вывода см. [13]).

Мы охарактеризовали идею «система знаний для ИС» и с ее помощью уточнили идею «знание в ИС», считая, что неопределяемым понятием является «цель ИС», истолковать которое можно посредством класса задач, решаемого посредством ИС.

Отметим одно важное обстоятельство для понимания природы анализа данных посредством ИС. Если Решатель реализует машинное обучение для БФ с фактами, имеющими оценку «неопределенность», то посредством автоматического порождения гипотез некоторые «неопределенности» будут устранены, и соответствующие им факты из БФ преобразуются в знание из БЗ. Таким образом, исходная БФ уменьшается, а БЗ расширяется.

Применение Решателя к $\text{БФ} \cup \Sigma$, где Σ – базовое множество знаний, породит множество высказываний Σ' , где $\Sigma' \subseteq [R]$. Σ' может быть множеством высказываний не только уточняющих факты с оценкой «неопределенно», но и содержать другие следствия Решателя. В частности, возможно, что имеется $\Sigma'' = \text{Рассуждатель}(\text{БФ} \cup \Sigma)$, где $\Sigma'' \subseteq \Sigma'$.

Пусть $\tilde{\Sigma} = \Sigma' \setminus \Sigma''$, где $\tilde{\Sigma}$ - результаты, полученные Решателем с использованием Вычислителя. Будем говорить, что анализ данных является интеллектуальным, если $\Sigma'' \neq \emptyset$.

⁴ П может принадлежать Рассуждателю, или Вычислителю, или Синтезатору («или» является неисключающим)

Таким образом, под ИАД мы будем понимать анализ данных из БФ с использованием знаний для ИС Σ посредством Решателя такой, что $\Sigma'' \neq \emptyset$, где $\Sigma'' = \text{Рассуждатель} (\text{БФ} \cup \Sigma)$.

ИАД будем называть гибридным, если одновременно выполняются

$$\Sigma'' \neq \emptyset \text{ и } \tilde{\Sigma} \neq \emptyset.$$

Процесс создания информационной среды, т.е. $\text{БФ} \cup \Sigma$ является весьма трудоемким и осуществляется технологически посредством инженерии знаний (систематическое изложение инженерии знаний содержится в [9]). Отметим, что выбор $\text{БФ} \cup \Sigma$ совершается в соответствии с целью ИС – некоторой проблемой P_1 (назовем ее начальной проблемой). Очевидно, что P_1 соответствует некоторый класс задач, решение которых требует автоматизации определенного набора интеллектуальных способностей, упомянутых ранее. Очевидно, что формирование БФ $\cup \Sigma$ является важным этапом процесса КДД.

Обозначим посредством Γ – множество правил вывода, содержащих как правила достоверного вывода, так и правила правдоподобных выводов, посредством которых формализуются различные правдоподобные рассуждения [14]. Посредством C обозначим множество вычислительных процедур, применяемых в ИС для решения проблемы P_1 . Очевидно, что правила из Γ и C реализуются в Рассуждателе и Вычислителе, соответственно, а их комбинирование осуществляет Синтезатор. Напомним, что Решатель = Рассуждатель + Вычислитель + Синтезатор.

Теперь мы можем охарактеризовать строение БЗ для ИС, решающей проблему P_1 :

$$\text{БЗ} = \langle \Sigma, \Sigma', \Gamma, C \rangle, \text{ где } \Sigma' \subseteq [R], \text{ а } [R] = \text{Решатель} (\text{БФ} \cup \Sigma).$$

Таким образом, создание БЗ есть представление знаний «системы знаний для ИС», охарактеризованной выше. Однако мы пока охарактеризовали лишь декларативные и процедурные знания. Теперь сформулируем идею концептуальных знаний. С этой целью рассмотрим следующие принципы.

Принцип I (цель ИС)

Для создания ИС должна быть сформулирована проблема P_1 , которой соответствует класс формализуемых задач.

Примерами P_1 являются диагностика по клиническим данным, предсказание биологической активности химических соединений (в том числе - токсичности), диагностика дефектов изделий, автоматическая классификация текстов по теме, принятие решений на основе предыдущего опыта, предсказание социального поведения (действий, установок, мнений) и распознавание его рациональности, формирование типологии преступлений на основе базы прецедентов с последующим распознаванием типа преступления (например, преступлений рецидивистов) и т.д.

Принцип II (типы «миров»).

Можно выделить три типа предметных областей W («миров») [3]:

(а) предметные области («миры», «универсумы») такие, что факты, принадлежащие W , являются случайными событиями; соответственно правила Решателя (т.е. элементы Γ и

С), применяемые в рассуждениях и вычислениях, используют идеи теории вероятностей (в том числе различные статистические методы анализа данных);

(в) предметные области W такие, что факты, принадлежащие W , причинно обусловлены; соответственно, правила Решателя, применяемые в рассуждениях с W , порождают гипотезы о причинно-следственных зависимостях и основанных на них обобщениях;

(с) предметные области W такие, что факты, принадлежащие W , могут быть как причинно-обусловленными, так и случайными (это означает, что этот «мир» является обобщением миров (а) и (в)); рассуждения, представленные в Решателе, соответствующие этому «миру», должны использовать правила, порождающие гипотезы о зависимостях причинно-следственного типа с учетом вероятностных соображений [15], [16] (например, частоты появления причинно-следственного эффекта).

Принцип III (адекватность предметной области W и Решателя для P_1).

Рассуждатель и Вычислитель должны содержать методы рассуждений и вычислений, соответствующие типам W , т.е. (а), (в) или (с).

Очевидно, что для W типа (а) адекватными будут статистические методы, а для W типа (в) адекватными методами рассуждения будут правдоподобные рассуждения, формализующие анализ отношений причинно-следственного типа (см. в связи с этим [8]).

Принцип IV (условие применимости ИС).

Этот принцип является развитием Принципов II и III. Он состоит в том, что условия применимости Решателя должны быть точно сформулированы. Например, для «миров» типа (в) может быть охарактеризован класс каузальных моделей, к которым применим метод автоматического порождения гипотез [8], использующий правдоподобные рассуждения, осуществляемые Рассуждателем.

БФ, представляющая W типа (в), должна содержать позитивные и негативные примеры изучаемого эффекта. В БФ в неявном виде должны содержаться причинно-следственные зависимости, как позитивные (т.е. вынуждающие наличие эффекта), так и негативные (т.е. вынуждающие отсутствие эффекта). Кроме того, структура данных, используемая для представления фактов из БФ, должна быть пригодна для формализации структурного сходства фактов. Это обстоятельство делает возможным реализацию принципа: сходство фактов влечет сходство эффектов и их повторяемость. Этот принцип лежит в основе формализованного качественного анализа данных.

Принцип V (формализованная эвристика посредством синтеза познавательных процедур для реализации КДД – обнаружения знаний в базах данных (фактов)).

ИС применяются для предметных областей, знания о которых слабо формализованы. Соответственно, решения задач для W требуют формализованных эвристик. В силу этого, актуален следующий принцип: формализованная эвристика для решения задач в ИС, относящихся к проблеме P_1 , должна быть синтезом познавательных процедур, посредством которых из $БФ \cup БЗ$ Решатель извлекает новые знания.

Примером синтеза познавательных процедур является упорядоченное комбинирование индукции, аналогии и абдукции [8].

Синтез познавательных процедур формализуется посредством правдоподобных рассуждений [14]. Он является механизмом извлечения новых знаний из $B\Phi \cup \Sigma$, реализуемым посредством Решателя с использованием, как БФ, так и БЗ.

Легко понять, что изолированное применение индукции, аналогии, абдукции и дедукции не может быть средством эффективной формализацией, имеющей широкое применение.

Индукция, устанавливающая сходство фактов, порождает гипотезы о причинах, аналогия «переносит» обнаруженные закономерности (причины) на факты с оценкой «неопределенно», а абдукция является способом принятия порожденных гипотез посредством объяснения начального состояния БФ.

Принцип VI (фальсифицируемость и аргументируемость результатов работы Решателя)

В ИС должны содержаться средства фальсификации результатов применения Решателя к $B\Phi \cup \Sigma$. Таковыми могут быть утверждения из Σ , которые накладывают ограничения на $[R]$, (где Решатель $(B\Phi \cup \Sigma) = [R]$), или же автоматически порожденные фальсификаторы, которые извлекаются из БФ и запрещают некоторые гипотезы из $[R]$.

Аргументируемость результатов из $[R]$ означает, что порождаемые гипотезы имеют аргументы за их принятие и не имеют контраргументов, их запрещающих.

Принцип VII (синтез теорий истины).

Для ИС, аппроксимирующих базисные способности интеллекта (выделение существенных факторов; синтез познавательных процедур, используемый, в частности, для порождения гипотез посредством правдоподобных рассуждений, а также БФ и БЗ; обучение на примерах из БФ), о которых упоминалось в начале статьи, непригодна аристотелевская теория истины как теория соответствия, формализованная А. Тарским для двузначной логики [17] (см. также [18] глава 9). Дело в том, что порождаемые Решателем гипотезы либо правдоподобны, если порождены Рассуждателем посредством правдоподобных рассуждений [14], либо имеют некоторую вероятность, если порождены Решателем с использованием статистических методов.

Однако и в том и в другом случае имеются критерии принятия гипотез на основе БФ и БЗ, где $BЗ = \langle \Sigma, \Sigma', G, C \rangle$. Следовательно, оценка знаний, полученных с помощью гипотез основана на знании, к которому теория соответствия не применима. Очевидно, что в этом случае в ИС используется теория истины, называемая теорией когерентности, в которой значения истинности основаны на имеющихся знаниях [18], а не на соответствии реальности.

Наконец, результаты работы ИС могут иметь практическую ценность, хотя их истинность не была установлена. В этом случае можно говорить о применимости прагматической теории истины [18]: истинно то, что полезно.

Таким образом, когнитивный процесс порождения нового знания в ИС может быть охарактеризован посредством синтеза трех теорий истины – теории соответствия, теории когерентности и прагматической теории. В самом деле, БФ формируется при соблюдении теории соответствия, гипотезы оцениваются согласно теории когерентности, а результаты работы ИС могут быть оправданы согласно прагматической теории истины. Таков принцип теории истины для ИС, использующих автоматическое порождение гипотез и машинное обучение.

Принцип VIII (инвариантность структуры Рассуждателя относительно варьируемости предметных областей и структур данных).

Пусть r_1, \dots, r_s – множество правил правдоподобного вывода, сформулированные в языке представления знаний L , а $\alpha_{r_1}, \dots, \alpha_{r_s}$ – соответствующие им импликативные аксиомы (т.е. декларативное представление процедур). Пусть σ – сигнатура, используемая для формулировки α_{r_i} , $i = 1, \dots, s$ (т.е. множество функциональных и предикатных символов).

Пусть, далее, посредством α_{r_i} представим синтез познавательных процедур согласно Принципу V, но функциональные символы для выражения сходства фактов могут иметь различные определения. Это означает, что могут варьироваться предметные области и структуры данных (при сохранении σ), однако не изменяются тип правил правдоподобного вывода и тип рассуждения. Например, сохраняется синтез познавательных процедур типа «индукция + аналогия + абдукция» с последующим применением дедукции.

Это означает, что может быть задан класс IS_j , соответствующих подобным проблемам $P^{(j)}$, $j = 1, \dots, k$, решения которых осуществляется одним типом формализованной эвристики. Этот класс эвристик характеризуется одним и тем же типом Рассуждателя и аналогичными условиями его применимости в соответствии с Принципами III и IV, соответственно, адекватности W и условиями применимости Решателя. Будем этот класс формализованных эвристик называть «ядром» Решателя. Очевидно, что конкретная IS_j есть «ядро» + его спецификация относительно предметной области и проблемы $P^{(j)}$.

Класс проблем $P1$, соответствующих «ядру» IS будем называть суперпроблемой (очевидно, что $P^{(j)} \in P1$). Например, суперпроблемой $P1$ является прогнозирование каких-либо эффектов посредством порождения гипотез о причинно-следственных зависимостях в БФ, а конкретными проблемами $P^{(1)}$ и $P^{(2)}$ являются проблемы прогнозирования биологической активности химических соединений посредством гипотез о фармакофорах (предполагаемых причин биологической активности химических соединений – их подструктур) и прогнозирования электорального поведения посредством некоторых детерминант, содержащихся в описаниях индивидуальных социальных субъектов.

Принцип IX (наличие метауровня IS).

Пусть заданы каркас IS $R = \langle B\Phi \cup \Sigma, \text{Решатель} \rangle$,

Решатель $(B\Phi \cup \Sigma) = [R]$ и Рассуждатель $(B\Phi \cup \Sigma) = \Sigma''$ и $B3 = \langle \Sigma, \Sigma', \Gamma, C \rangle$, где $\Sigma' \subseteq [R]$, выраженные посредством языка представления знаний L . Будем полагать, что имеются метаматематические средства ML , посредством которых можно формулировать дедуктивную имитацию Рассуждателя [19] и осуществлять анализ алгоритмов, соответствующих процедурам Решателя. Это означает, что имеется возможность исследования результатов Решателя потенциально релевантных цели IS , т.е. проблеме $P1$. Из этого следует, в частности, возможность планирования выбора стратегий решения задач IS на логическом уровне.

Принцип X (абдуктивное объяснение результатов IS).

Подсистема объяснения результатов компьютерной системы является важной характеристикой ее интеллектуальности. Разумеется, идеальной конструкцией объяснения согласно К. Гемпелю [20] является следующая схема

A_1, \dots, A_k

B

E

где A_1, \dots, A_k – аксиомы, характеризующие предметную область, B – условия реализации события E , которое дедуктивно следует из посылок A_1, \dots, A_k, B .

Очевидно, что эта схема дедуктивного объяснения не охватывает случаев открытых теорий, в которых предметная область может быть охарактеризована аксиомами лишь частично.

Таким образом, если имеется БФ и не имеются A_1, \dots, A_k и условия C , из которых E , принадлежащее БФ, дедуктивно выводимо, то при наличии средств автоматического порождения гипотез (см. в связи с этим [4], [7] и [8]), приходится использовать схему абдукции Ч.С. Пирса [5]:

D – множество фактов

H – множество гипотез

H объясняют D

$\forall h \in H, h$ является правдоподобной

Если D есть БФ ИС, $H \subseteq \Sigma' \subseteq \text{Решатель} (БФ \cup \Sigma)$, отношение « H объясняют БФ» формализовано относительно предметной области W (см. в связи с этим [8]), то абдуктивное принятие гипотез завершает синтез познавательных процедур в соответствии с Принципом VI (H порождаются посредством индукции и аналогии).

Таким образом, объяснение результатов Решателя порождает принятие нового знания, извлеченного из $БФ \cup \Sigma$ в соответствии с целями КДД.

Принцип XI (эволюционная эпистемология решения задач в ИС)

$P1 \rightarrow TT \rightarrow EE \rightarrow P2$ является известной схемой роста знания К.Р. Поппера [21], где $P1$ – решаемая проблема, TT – пробная теория для ее решения, EE – устранение ошибок и коррекция TT после ее применения, а $P2$ – вновь возникшая проблема после анализа результатов измененной (и более корректной) TT .

Принцип эволюционной эпистемологии для ИС с Рассуждателем, реализующим машинное обучение, может быть представлен следующим образом:

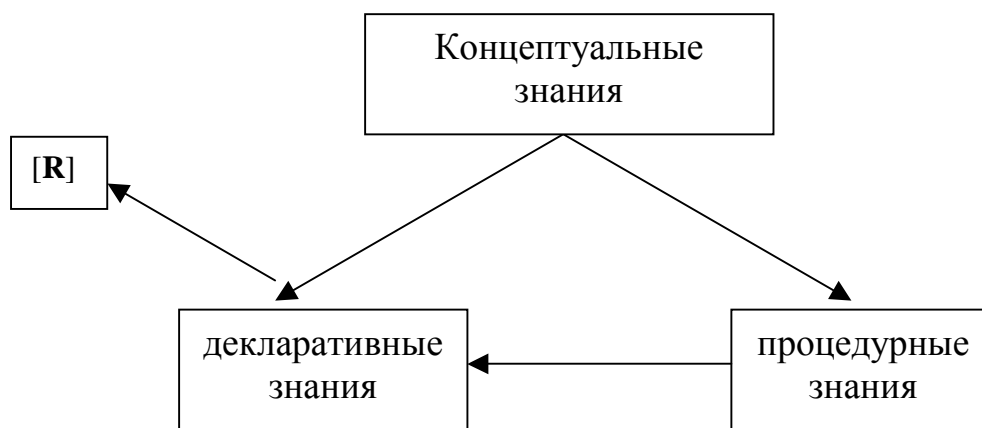
$P^{(1)} \rightarrow \text{Решатель (БФ} \cup \Sigma) \rightarrow \text{ЕЕ} \rightarrow P^{(2)}$,

где ЕЕ – коррекция и расширение БФ \cup Σ (например, с использованием абдукции – см. Принцип X и [8]) и, быть может, выбор другой стратегии для Рассуждателя, $P^{(2)}$ – новая возникшая проблема, требующая продолжения процесса КДД в рамках суперпроблемы P1 (или, быть может, с переходом к другой суперпроблеме).

Следствием Принципа XI является необходимость включения ИС в реальный процесс исследования, управления или принятия решений, а также открытость БФ \cup Σ , что означает, что ИС должна быть человеко-машинной (партнерской) компьютерной системой.

Следующие важные следствия вытекают из Принципов I-XI, которые могут быть предложены в качестве «кодекса интеллектуальности компьютерной системы».

- (1) ИС должна быть партнерской человеко-машинной системой, способной функционировать и как человеко-машинная система, применяемая к открытым предметным областям («мирам»), и как система, работающая в автономном режиме, применяемая к «замкнутым мирам».
- (2) Принципы I-XI образуют систему концептуальных знаний, управляющую созданием как подсистемы декларативных знаний, так и подсистемы процедурных знаний, т.е. «системы знаний для ИС». Следующая диаграмма представляет это управление:



Напомним, что [R] – множество результатов Решателя, применяемого к БФ \cup Σ , т.е. Решатель (БФ \cup Σ) = [R].

- (3) Цель ИС является одновременно и суперпроблемой P1, и множеством ее конкретизаций – проблемами $P^{(1)}$, относящимися к конкретным предметным областям W и соответствующим им БФ. Эта цель осуществляется посредством формализованной эвристики, которая способна не только имитировать интеллектуальные способности человека (в смысле теста А. Тьюринга), но и

усиливать их. Это означает, что Рассуждатель реализует классы рассуждений таких, что аппроксимируются важные познавательные возможности человеческого интеллекта: выделение существенных особенностей в фактах посредством установления их сходства (в особенности структурного для сложноорганизованных объектов), осуществление правдоподобных рассуждений, являющихся синтезом различных познавательных процедур (в том числе: индукции, аналогии и дедукции), использование фальсификации на основе контрпримеров, аргументирование принимаемых решений, объяснение результатов посредством абдукции, автоматическое порождение гипотез с эффективно приписываемой им оценкой правдоподобия, обучение на примерах из БФ, выбор стратегий после коррекции результатов (процесс ЕЕ в схеме роста знания К.Р. Поппера).

В [8] правдоподобные рассуждения, аппроксимирующие указанные выше интеллектуальные способности, были названы СРЕ-рассуждениями – когнитивными правдоподобными эмпирическими рассуждениями.

Таким образом, Рассуждатели, которые осуществляют СРЕ-рассуждения, применяются для суперпроблем P1 таких, что для решения их задач используются исходные данные, которые являются изменяемыми и пополняемыми множествами высказываний (под изменением высказывания мы понимаем пересмотр его истинностного значения, а БФ для таких высказываний ранее были названы эпистемическими).

Очевидно, что СРЕ-рассуждения являются недедуктивными: из истинности их посылок не вытекает истинность следствия. Следствия СРЕ-рассуждений являются лишь правдоподобными высказываниями. СРЕ-рассуждения содержат амплиативные выводы (термин Ч.С. Пирса), следствия которых есть результат некоторой догадки, оформленной в виде применения специальных правил. Примерами таких правил являются правила для индукции, аналогии и абдукции.

Отметим, что класс когнитивных рассуждений (СР-рассуждений), содержащих амплиативные выводы, шире класса СРЕ-рассуждений. Класс СР-рассуждений подразделяется на три известных в литературе подкласса [14]: вероятностные (например, использующие байесовские правила), приближенные (например, использующие аппарат нечетких множеств), правдоподобные рассуждения, являющиеся организацией различных взаимодействующих процедур (например, аргументации и пересмотра истинностных значений). СРЕ-рассуждения являются примером последнего типа рассуждений.

Третье следствие Принципов I – XI состоит в том, что, говоря метафорически, «мозгом» ИС является взаимодействие Рассуждателя, реализующего СР-рассуждения, с информационной средой ИС (т.е. с БФ и БЗ) при условии открытости БФ и БЗ, создающих возможность их коррекции (ЕЕ – в схеме роста знания), благодаря которой человеко-машинная система осуществляет адаптацию к изменившемуся состоянию ИС. Это создает возможность сформулировать новую суперпроблему (P2 в схеме роста знания).

Следовательно, строение и возможности Рассуждателя определяют тип анализа данных, компьютерной системы (КС), обладающей Решателем который используется для анализа данных (в том числе ДМ и КДД). Можно выделить три основных типа Решателей для анализа данных:

1°. Решатель = Вычислитель + Синтезатор, Рассуждатель отсутствует, Вычислитель может осуществлять различные процедуры над числовыми данными, Синтезатор же объединяет различные вычислительные методы, реализуемые Вычислителем. Компьютерную систему с таким Решателем будем называть вычислительной.

2°. Решатель = Рассуждатель + Вычислитель + Синтезатор, где Рассуждатель осуществляет комбинаторные алгоритмы [22], а Синтезатор объединяет вычислительные методы Вычислителя и комбинаторные алгоритмы Рассуждателя. Компьютерную систему с таким Решателем будем называть комбинаторно-вычислительной.

3°. Решатель = Рассуждатель + Вычислитель + Синтезатор, где Рассуждатель реализует СР-рассуждения, а Синтезатор формирует стратегии решения задач, объединяющие рассуждения, вычисления и комбинаторные алгоритмы. Однако рассуждения являются управляющей процедурой выполняемой стратегии решения задач.

КС с Решателем типа 3° и БЗ= $\langle \Sigma, \Sigma', \Gamma, C \rangle$ будем называть интеллектуальной, и, соответственно, анализ данных посредством этой ИС будем называть интеллектуальным анализом данных.

ИС реализует анализ данных посредством взаимодействия трех типов знания – концептуального (в соответствии с Принципами I – XI), декларативного и процедурного (схема этого взаимодействия представлена в приведенной выше диаграмме).

Примерами анализа данных в вычислительных системах являются КС, использующие различные статистические методы и нейронные сети.

Примерами КС, которые были названы выше комбинаторно-вычислительными, являются системы, использующие деревья решений (см. в связи с этим [7]), а также системы формирования знаний методами локальной геометрии [23, 24].

Обратим внимание еще раз на одно важное обстоятельство: для интеллектуального анализа данных необходимо определить тип предметной области W , соответствующей суперпроблеме P_1 и проблеме $P_1^{(1)}$ согласно Принципу III (адекватности W и P_1).

- (4) Четвертое следствие Принципов I – XI состоит в интерпретируемости результатов работы Рассуждателя Σ'' в БФ, где $\Sigma'' = \text{Рассуждатель} (\text{БФ} \cup \Sigma)$. Эта интерпретируемость основана на Принципах I (Цель ИС P_1), II (Выделение и характеристика предметной области W), III (адекватность W и P_1 , где P_1 – суперпроблема), IV (условие применимости ИС), V (синтез познавательных процедур для КДД), VI (фальсифицируемость и аргументируемость $[R]$ – результатов работы Решателя) и X (принцип абдуктивного объяснения результатов ИС) и, наконец, XI (эволюционная эпистемология решения задач в ИС, включающая EE – коррекцию результатов и исправление ошибок).

Интерпретируемость результатов ИС средствами БФ означает их «экологическую чистоту» относительно W , минимально искажаемую формальными средствами Решателя.

В связи с условием интерпретируемости результатов работы Решателя отметим сомнительность в возможности интерпретируемости результатов нейронных сетей, о чем справедливо сказано в [23].

- (5) Пятым следствием упомянутых выше Принципов I – XI является возможность извлекать из результатов работы ИС не только «образцы», добываемые из данных в смысле «data mining», но и формулировать гипотезы о закономерностях, обнаруженных в расширяемых последовательностях БФ в интерактивном режиме работы ИС.

Пусть I – множество имен реализаций ИС, тогда множество реализаций ИС обозначим посредством R_i , где $R_i = [R]_{i \in I}$. Будем говорить, что p – устойчивый результат, если $p \in R_i$. Устойчивый результат будем называть гипотезой о закономерности относительно R_i , если p – универсально квантифицированное высказывание (т.е. кванторная приставка начинается с квантора общности, а p имеет предваренную нормальную форму).

Если предметная область W имеет тип (\forall) , т.е. БФ, соответствующая ей, содержит в неявно заданном виде причинно-следственные отношения, то гипотезы о закономерностях будут выражать зависимости причинно-следственного типа, порожденные Рассуждателем [8] (о возможностях машинного открытия закономерностей см. в [25]).

Очевидно, что не каждый результат ДМ и КДД является гипотезой о закономерностях, однако существует возможность порождения гипотез о закономерностях посредством ИС, применяемых к открытым БФ \cup Σ . Этот случай применения компьютерных систем, разумеется, является интеллектуальным анализом данных.

Сделаем теперь некоторые замечания о терминологии, относящейся к анализу данных. Термины “data mining” (ДМ) и “knowledge discovery” (КД) могут быть, соответственно, переведены как «добыча из данных» и «открытие знания» (или «обнаружение знания»). Аналогично термин “knowledge discovery in data bases” (КДД) переводится как «открытие знаний в базах данных». Недостаточно четкое раскрытие смысла терминов ДМ, КД и КДД было отмечено в [23]. Можно предположить, что источником трудностей уточнения смысла этих терминов является отсутствие соответствующих уточнений таких идей искусственного интеллекта как «интеллектуальная система», «Решатель задач», «Синтез познавательных процедур», «знание для интеллектуальной системы», «когнитивное рассуждение». Рассмотрение же идей ДМ, КД и КДД в системе перечисленных выше уточненных идей создает возможность различать возможные виды анализа данных. Обратим также внимание на тот факт, что ДМ в смысле [1] трудно отличить от распознавания образов, в котором из данных извлекается «образец». По-видимому, наиболее плодотворным путем развития методов анализа данных в рамках исследований ИИ является определенное выше «интеллектуальный анализ данных» посредством ИС и анализ данных посредством «комбинаторно-вычислительных систем».

Современные компьютерные технологии сделали возможным анализ клинических данных осуществлять посредством различных точно определенных процедур, которые не только извлекают из данных явно содержащиеся в них «образцы» в смысле ДМ, но и порождают обобщение для систем БФ, могущие быть интерпретируемыми как гипотезы о закономерностях относительно соответствующих семейств БФ. В [26] авторы рассматривают способы анализа клинических данных, полагая, что они могут быть только количественными. Более того, они считают, что осуществляемая диагностика должна характеризоваться посредством вероятностей. Они, конечно, рассматривают клинические прогнозы как гипотезы, которые могут использовать предполагаемые причины диагноза. Авторы область клинической эпидемиологии характеризуют как evidence based medicine. К сожалению, этот термин переведен как «доказательная медицина», а не как «медицина, основанная на очевидных фактах». Смысл же этого термина состоит в том, что решения, принимаемые врачами, должны быть аргументированными (но не доказанными – ведь авторы говорят о вероятных прогнозах, т.е. о гипотезах!), что означает, что решения используют аргументы, извлеченные из клинических данных. Однако данные могут быть не только количественными, они могут включать также качественные параметры. Следовательно, анализ данных в медицине должен (при условии использования компьютерных технологий) совершаться в рамках сложившейся культуры ИИ. Сформулированные типы анализа данных посредством вычислительных систем, комбинаторно-вычислительных систем и интеллектуальных систем (т.е. интеллектуальный

анализ данных) необходимы «медицине, основанной на очевидных фактах». А это означает, что выбор таких систем должен быть адекватен предметной области и соответствующей проблеме $P^{(j)}$. Более того, анализ клинических данных не может быть лишь количественным, он должен быть и качественным, использующим адекватные БЗ и Решатели задач.

Следует обратить внимание также и на другое важное обстоятельство – измерение (т.е. применение количественных методов) должно предполагать первоначальную формализацию данных как систем отношений. Само же измерение есть отображение реляционной системы в числовую [27]. Следовательно, без качественного предпроцессинга вряд ли можно обойтись (это есть один из аспектов КДД).

Настоящий номер «Новостей искусственного интеллекта» посвящен проблеме анализа данных в ИИ. В нем имеются два раздела «Методы машинного обучения и анализа данных» и «Анализ данных в медицине».

В первом разделе журнала публикуются статьи С.О. Кузнецова «Методы теории решеток и анализа формальных понятий в машинном обучении», М.О. Корляковой «Исследование алгоритмов обобщения для организации описания предметной области» и С.М. Гусаковой «Подход к решению задач атрибуции исторических источников с помощью ДСМ - метода».

В статье С.О. Кузнецова представлен обзор применения алгебраических методов для машинного обучения, используемого для обнаружения знаний в БФ. Следует подчеркнуть тот факт, что машинное обучение является необходимым инструментом КДД как для комбинаторно-вычислительных систем, так и для систем для интеллектуального анализа данных.

Статья М.О. Корляковой посвящена исследованию КДД – точнее, его алгоритмическим аспектам. Рассматриваемые автором алгоритмы предназначены для использования в обнаружении знаний в БФ.

В статье С.М. Гусаковой предложен метод интеллектуального анализа данных посредством некоторой модификации ДСМ – метода автоматического порождения гипотез [8] для задачи атрибуции исторических источников. Следует отметить важность применения методов ИИ к проблемам гуманитарных наук.

Во втором разделе журнала рассмотрены проблемы анализа медицинских данных.

В статье «Технология Data Mining в медико-биологических исследованиях В.А. Дюк представил обстоятельный обзор обнаружения знаний в базах данных» (КДД). Автор приводит пример системы для прогнозирования времени жизни пациентов, перенесших сердечный приступ, по данным эхокардиограммы. Таким образом, рассматриваемая система по нашей классификации является системой с Рассуждателем, реализующим комбинаторно-вычислительный анализ БФ с последующим предсказанием исследуемого эффекта. В силу того, что Рассуждатель использует правила продукционного типа, автор обоснованно называет эту систему экспертной. В статье также содержатся критические замечания относительно анализа данных посредством нейронных сетей, деревьев решений и генетических алгоритмов.

В статье Б.А. Кобринского рассмотрены проблемы инженерии медицинских знаний для формирования баз знаний с участием множества экспертов.

Статья О.Ю. Ребровой «Применение методов интеллектуального анализа данных для решения задачи медицинской диагностики» является отчетом о результатах применения

статистических методов и нейронных сетей для диагностики трех типов инсульта по клиническим данным. Автор относит применяемые методы к сфере «доказательной медицины». В [28] представлен обзор применения статистических методов в медицине.

В статье Е.С. Панкратовой формулируются принципы представления знаний для задач клинической и лабораторной диагностики. Предложенные средства предназначены для использования в интеллектуальных системах с Рассуждателем, реализующим синтез индукции, аналогии и абдукции, образующий ДСМ – рассуждения [7,8]. Анализ данных осуществляется в БФ посредством Решателя с Рассуждателем, реализующим ДСМ – рассуждения.

В начале данной статьи были сформулированы пять вопросов об анализе данных как инструменте ИИ и перспективах его развития как средства интеллектуального анализа данных (ИАД). Введенная в статье терминология есть попытка формулирования системы идей, характеризующей различные типы анализа данных, соответствующие природе предметных областей, благодаря которой возможно уточнение смысла термина ИАД. Кратким резюме было бы следующее утверждение: ИАД есть процесс извлечения новых знаний из БФ и БЗ посредством ИС, имеющий Решатель с подсистемой Рассуждателя, реализующей когнитивные рассуждения.

Анализ данных является необходимым этапом любой познавательной деятельности в сферах науки, управления, медицины и юридической деятельности. В силу этого отображение познавательной деятельности в ИС, использующей БФ со сложно структурированными объектами, является постоянной потребностью компьютерной науки, которая адекватно характеризуется схемой эволюционной эпистемологии $P^{(1)}1 \rightarrow TT \rightarrow EE \rightarrow P^{(2)}$, где $P^{(1)}$ и $P^{(2)}$ - исходная и новая проблемы относительно предметной области W .

Автор выражает глубокую благодарность Д.В. Виноградову и М.А. Михеенковой за ценные замечания.

Литература

- [1] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. Advances in Knowledge Discovery and Data Mining, The AAAI Press, 1996.
- [2] Финн В.К. Об особенностях ДСМ–метода как средства интеллектуального анализа данных, НТИ, сер.2, №5, 2001, с. 1-4.
- [3] Финн В.К. Искусственный интеллект: Идеальная база и основной продукт, 9-ая национальная конференция по искусственному интеллекту, Труды конференции, Т.1, М., Физматлит, 2004, с.11-20.
- [4] Гаек П., Гавранек Т., Автоматическое образование гипотез. М.: Наука, 1984.
- [5] Josephson J.R., Josephson S.G. (Eds.) Abductive Inference: Computation, Philosophy, Technology. New York: Cambridge University Press, 1994.
- [6] Jain S., Osherson D., Royer J.S., Sharma A. Systems That Learn. An Introduction to Learning Theory, second edition. The MIT Press, Cambridge, Massachusetts, London, England, 1999.
- [7] Кузнецов С.О. Методы теории решеток и анализа формальных понятий в машинном обучении. Настоящий номер журнала «Новости искусственного интеллекта».

- [8] Финн В.К. Синтез познавательных процедур и проблема индукции. НТИ, сер.2, № 1-2, 1999, с. 8-44.
- [9] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. Питер-Пресс, 2000.
- [10] Nilsson N.J. Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publishers, Inc., San Francisco, California, 1998.
- [11] Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ. Итоги науки и техники. Информатика, Т.15, М.: 1991, с. 54-101.
- [12] Поспелов Д.А. Ситуационное управление. Теория и практика. М.: Наука, главная редакция физико-математической литературы, 1986.
- [13] Antoniou G. Nonmonotonic reasoning. MIT Press, 1997.
- [14] Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах. М.: Физматлит, 2004.
- [15] Аншаков О.М. Обобщенные кванторы, определяемые с помощью шаблонов. Ч. I. НТИ, сер.2, № 11, 2000, с. 5-17.
- [16] Аншаков О.М. Обобщенные кванторы, определяемые с помощью шаблонов. Ч. II. НТИ, сер. 2, № 5, 2001, с. 35-48.
- [17] Тарский А. Понятие истины в языках дедуктивных наук. Философия и логика Львовско-Варшавской школы. М.: РОССПЭН, 1999, с. 19-155.
- [18] Поппер К.Р. Объективное знание. Эволюционный подход. М.: УРСС, 2002.
- [19] Аншаков О.М., Скворцов Д.П., Финн В.К. Логические средства экспертных систем типа ДСМ. Семиотика и информатика. Выпуск 28, М.: 1986, с. 65-101.
- [20] Гемпель К.Г. Логика объяснения. Дом интеллектуальной книги, М.: 1998.
- [21] Эволюционная эпистемология и логика социальных наук. Карл Поппер и его критики. Эдиториал УРСС. М.: 2000.
- [22] Рейнгольд Э., Нивергельт Ю., Део Н. Комбинаторные алгоритмы. Теория и практика. Мир, М.: 1998.
- [23] Дюк В.А. Технологии Data Mining в медико-биологических исследованиях. Настоящий номер журнала «Новости искусственного интеллекта».
- [24] Дюк В. Обработка данных на ПК в примерах. Питер, Санкт-Петербург, 1997.
- [25] Simon H.A. Machine Discovery. Foundations of Science, vol. 1, №2, 1995/96, pp. 171-200.
- [26] Флетчер Р., Флетчер С., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. Медиа Сфера, М.: 2004.

[27] Пфанцагль И. Теория измерений. М.: Мир, 1976.

[28] Реброва О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. Медиа Сфера, М.: 2002.