МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Т. В. Афанасьева
А. Н. Афанасьев

Введение в проектирование систем интеллектуального анализа данных

Учебное пособие

Ульяновск
УлГТУ
2017

УДК 004.8 (075)
ББК 32.813 я7
А 94

Рецензенты:
канд. техн. наук, доцент, генеральный директор ООО «РИТГ»
Игонин А. Г.;
кафедра систем автоматизированного проектирования ЮФУ.

Утверждено редакционно-издательским советом университета в качестве учебного пособия

Афанасьева, Татьяна Васильевна

А 94 Введение в проектирование систем интеллектуального анализа данных : учебное пособие / Т. В. Афанасьева, А. Н. Афанасьев. – Ульяновск : УлГТУ, 2017. – 64 с.

ISBN 978-5-9795-1686-8

Содержание пособия включает изложение основ направления интеллектуального анализа (Knowledge Discovery in Databases&Data Mining), которое в настоящее время объединяет статистические, нейросетевые и нечеткие модели, необходимое при проектировании систем анализа данных. Приводятся постановки основных задач интеллектуального анализа данных, примеры систем и стандарты в этой области. Описание задач, примеры и стандартов интеллектуального анализа данных базируется на современном обзоре отечественных и зарубежных источников, системном подходе, рассматривается через призму автоматизации проектирования и сопровождается контрольными вопросами.

Пособие предназначено для студентов и аспирантов высших учебных заведений, обучающихся по Укрупненной группе специальностей и направлений 09.00.00 «Информатика и вычислительная техника», специализирующихся в области интеллектуального анализа данных интеллектуальных систем, автоматизации проектирования сложных систем, и может быть полезным специалистам промышленных предприятий и научных организаций.

УДК 004.8 (075)
ББК 32.813 я7

ISBN 978-5-9795-1686-8

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время теория и практика проектирования накопили богатую историю как в общеметодологическом аспекте, так и в технологическом. На уровне отдельной организации это позволяет использовать исторически накопленный опыт для планирования и улучшения результатов проектной деятельности. Формализация опыта выполнения проектов и управления проектной деятельностью опирается на гранулы информации, извлеченные из исторически накопленных экспериментальных данных, экспертных знаний и кодифицированных знаний (например, в виде онтологий).

В условиях развития гибких и бережливых технологий проектирования важно иметь автоматизированные средства извлечения полезной информации для принятия рациональных проектных и управленческих решений.

Data Mining (в русскоязычной литературе по информационным технологиям этот термин переведен как интеллектуальный анализ данных) выступает одним из основных инструментов оперативного извлечения разноаспектных информационных гранул из баз данных. В этом смысле Data Mining является средством автоматизации проектной деятельности на основе развитой аналитики, включающей пространственную (или дескриптивную), темпоральную (или предикативную) аналитику.

В зарубежных источниках интеллектуальный анализ данных (ИАД) часто трактуется как более общая методология, включающая пред-обработку, анализ и пост-обработку данных и связывается с понятием Knowledge Discovery in Databases&Data Mining (KDD&DM) или с понятием Data Mining. Так, в определении C. C. Aggarwal Data Mining это научно-практическое направление, включающее изучение и создание

автоматизированных методов выбора, очистки, обработки и анализа данных для извлечения информации.[1]

В настоящем пособии мы будем придерживаться трактовки понятия ИАД как KDD&DM, отдельно выделяя процесс и задачи Data Mining.

Обычно анализ данных выполняется на основе статистических и вероятностных моделей, имеющих широкий арсенал методов и критериев качества. Однако вероятностная модель объектов и процессов не всегда может быть адекватна решаемой задаче. Тогда для решения задач анализа прибегают к помощи эвристических моделей и методов, «подсказанных» природой. К таким моделям относят искусственные нейронные сети, эволюционные алгоритмы и нечеткие системы: все они включены в направление интеллектуального анализа данных.

Дескриптивная и предикативная аналитика часто является основным формальным инструментом для разработки новых продуктов и технологий. Для разработки не только нового, но конкурентного продукта необходимо исследовать возникающие тенденции с использованием формальных методов на этапе анализа. Это исследование должно быть направлено на перспективные функции и технологии, на научные достижения и требования пользователей. Для выполнения таких исследований необходимо использовать алгоритмы извлечения полезной и новой информации. Data minig (интеллектуальный анализ данных) как научно-практическое направление, ориентировано на создание и приме-нение таких алгоритмов. Это направление активно развивается в условиях быстрого роста накопленных данных, однако недостаточно освещено в методическом аспекте применительно к проектированию систем ИАД.

Целью данного пособия является представление основ Data Mining с ориентацией на вопросы автоматизации проектирования.

1 C. C. Aggarwal. Data Mining: The Textbook, Springer International Publishing Switzerland, 2015.

В первой главе пособия освещены вопросы актуальности и потребности в развитии направления интеллектуального анализа данных и основные понятия. Рассмотрены соотношение задач Data Mining с задачами автоматизированного проектирования, Machine Learning и Knowledge Discovery in Databases и основные этапы анализа данных.

Во второй главе приведены абстрактные классы задач и методов Data Mining с дальнейшей детализацией применительно к автоматизированному проектированию.

Формализация постановки основных задач Data Mining обычно вызывает затруднения при проектировании систем анализа данных, поэтому третья глава посвящена этой проблематике. Не претендуя на полноту и строгую математическую форму, в этой главе с единых методологических позиций приведены формальные постановки задач классификации, поиска аномалий, кластеризации, прогнозирования и извлечения ассоциативных правил. Приведены постановки и примеры сравнительно новых задач Data Mining, таких как поиск концептов на основе формального концептуального анализа и генерация лингвистического резюмирования.

Четвертая глава посвящена описанию примеров систем автоматизации задач ИАД на основе Data Mining. Логика изложения реализована от вопросов в рамках концептуального проектирования абстрактной системы Data mining до реальных решений в различных прикладных областях. Приведены восемь примеров прикладных систем ИАД.

Так как в разработке систем ИАД накоплен определенный опыт возникает необходимость перехода на стандартизированный процесс их разработки. Поэтому в последней главе приведены восемь основных стандартизованных методологий в области создания KDD&DM систем.

1. ОСНОВНЫЕ ПОНЯТИЯ

Средства автоматического сбора данных, повсеместное внедрение информационных технологий хранения данных, электронный документооборот, сетевые технологии – все это ведет к росту объемов и усложнению структур хранимых данных.

В связи с этим возникает необходимость в разработке средств автоматизированного анализа данных большого объема и сложной структуры для извлечения полезной информации для дальнейшего применения. Для этих целей разработаны методы решения задач, которые относятся к различным научным направлениям, таким как статистика, теория информации. В последние десятилетия эти научные направления расширены за счет методов Data Mining, Machine Learning и Knowledge Discovery in Databases.

1.1. Data Mining, Machine Learning и Knowledge Discovery in Databases

В литературе рассматриваются три группы активно развиваемых направлений в области ИАД: Data Mining, Machine Learning и Knowledge Discovery in Databases (KDD). Иногда эти направления рассматриваются как эквивалентные, так как они нацелены на исследование и построение моделей данных, их свойств и зависимостей.

В настоящее время считается, что KDD включает Data Mining как ядро и при этом может использовать методы Machine Learning. Data Mining является частью естественного развития информационных технологий, предназначенных для анализа хранимых данных.

Термин Data Mining был введен для обозначения совокупности методов автоматизированного решения сложных задач с целью извлечения полезной информации из больших баз данных в виде *свойств*, *группировок* и *зависимостей* в данных и дальнейшего применения этой информации

для получения экономической или иной выгоды. Такую информацию часто называют *паттерном* предметной области, она характеризует текущее состояние и возможные изменения в предметной области.

Таким образом, актуальность и потребность в становлении и развитии направления ИАД обусловлены проблемой «Big Data» и необходимостью развития методов различных научных парадигм для автоматизированного извлечения полезной информации из больших данных.

На рисунке 1 представлено соотношение между этими понятиями и связь с другими классами методов и моделей. Опираясь на статистический анализ, базы данных, теорию информации, методы оптимизации и машинное обучение KDD&DM направлен на получение, возможно, менее точного результата за более короткое время за счет большей степени автоматизации решения «интеллектуальных» задач для больших данных.

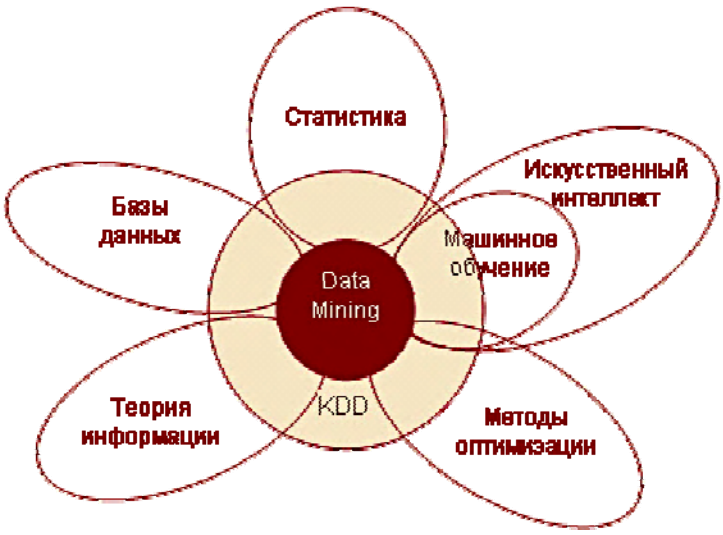

Рис. 1. Совокупность методов, образующих направление интеллектуального анализа данных и место Data Mining среди них

Поэтому необходимо представлять содержание процесса KDD&DM как системную методологию полезную для проектирования систем ИАД.

Методология KDD&DM описывает последовательность действий, которую необходимо выполнить для извлечения полезной информации. Эта последовательность этапов не зависит от предметной области, но не исключает применение экспертных знаний.

Выделяют пять концептуальных этапов проектирования систем ИАД, согласно методологии KDD&DM:

1. Постановка решаемой проблемы:
    a. Выявление и формулировка необходимых априорных знаний о предметной области, целей анализа, понимание бизнес-задач приложения, сценариев использования.
    b. Выдвижение вариантов решения проблемы и определение необходимых переменных.
    c. Выбор типа решаемой задачи Data Mining (классификация, прогнозирование, кластеризация, поиск аномалий (исключений), поиск ассоциативных правил и т. д.).
    d. Формулировка постановки решаемой проблемы. Проблема формулируется в результате тесного сотрудничества проектировщика системы ИАД и эксперта предметной области.
2. Формирование и предобработка данных для анализа:
    a. Поиск (или выбор) «сырых» данных, возможно, реализация подсистемы сбора (консолидации). В этом процессе возможны два подхода: сбор данных под управлением эксперта и случайная выборка данных.

b. Предобработка данных (нормализация, дискретизация, обработка пропущенных значений, удаление артефактов, проверка однородности). Часто на этом этапе обнаруживаются и удаляются выбросы (outliers), проводится преобразование и шкалирование данных, извлечение из данных характерных признаков или свойств (features).
c. Сокращение размерности, выбор значимых характеристик, расчет интегральных показателей и инвариантов.

3. Выбор, реализация и оценивание алгоритма Data Mining:
   a. Определение ограничений и требований к алгоритму Data Mining по точности, размеру, интерпретируемости, скорости построения и применения получаемых моделей, по типу исходных данных.
   b. Выбор одного или нескольких алгоритмов решения выбранного на первом этапе класса задачи.
   c. Выбор критериев оценивания решения задачи на основе Data Mining.
   d. Реализация, тестирование и оценивание выбранного алгоритма Data Mining (визуализация, описание, удаление избыточности, оценка точности, достоверности моделей и т. д.).
4. Применение выбранного алгоритма (алгоритмов) Data Mining для извлечения полезной информации:
   a. Оценивание «полезности» извлеченной информации.
   b. Агрегация результатов анализа по множеству переменных.
5. Интерпретация и представление результатов анализа:
   a. Лингвистическое резюмирование, как инструмент создания текстового описания результатов Data Mining.
   b. Сохранение результатов в базах данных и базах знаний.

1.2. Контрольные вопросы

1. Каковы цели и актуальность применения методов Data Mining и их приложений?
2. Охарактеризуйте три вида методов интеллектуального анализа данных и приведите графическую интерпретацию их соотношения.
3. Какие методы образуют направление интеллектуального анализа данных?
4. Охарактеризуйте и проанализируйте соотношение Data Mining, Machine Learning и Knowledge Discovery in Databases с другими классами методов.
5. Предложите дополнительные методы (свой вариант) для расширения совокупности методов направления интеллектуального анализа данных.
6. При решении каких проектных задач востребованы методы интеллектуального анализа данных?
7. В чем заключается процесс KDD&DM?
8. Перечислите и охарактеризуйте этапы концептуального проектирования систем ИАД согласно KDD. Каковы сходство и отличия от концептуального проектирования программных продуктов?
9. Какие задачи уже автоматизированы, могут быть автоматизированы и не требуют автоматизации на первом этапе проектирования систем ИАД согласно KDD&DM?
10. Сформулируйте методы предобработки и приведите не указанные в этой главе.
11. В чем заключается этап выбора и оценивания алгоритма решения задачи на основе Data Mining?
12. Постройте графическую схему процесса проектирования систем ИАД согласно методолгии KDD&DM. Это итерационная или спиральная модель процесса?

2. DATA MINING И АВТОМАТИЗИРОВАННОЕ ПРОЕКТИРОВАНИЕ

2.1. Типы задач и результатов Data Mining

В процессе решения задач Data Mining для аналитика и проектировщика извлеченная информация может быть следующих видов:

- Полезная. Такая информация ранее была неизвестна, имеет логическое объяснение, может быть использована для принятия проектных решений, приносящих выгоду.
- Тривиальная. Это информация имеет логическое объяснение, но ранее была известна и может использоваться для проверки выполнения ранее созданных проектных решений.
- Непонятная. Извлеченная информация ранее была неизвестна, но не может быть логически объяснена. Такая информация может содержать с одной стороны принципиально новые и в будущем полезные паттерны, а с другой стороны соовершенно бесполезные и противоречивые гранулы. Последнее означает, что особых закономерностей в данных не было обнаружено: либо их нет вообще, либо необходимо выбрать иной метод их обнаружения. Для уточнения требуется провести дополнительный анализ.

Термин «полезная» информация выражает степень значимости информации с точки зрения получения некоторой выгоды или выигрыша в будущем для субъекта проектной деятельности. При этом в качестве субъекта может выступать и некоторая организация (или проектная группа), и отдельный проектировщик. В этом смысле «полезность» информации является относительной характеристикой, зависящей от оценки состояния, выбора направления развития и заявленных целей субъекта, то есть от пространственно-темпоральных прогнозов и достигнутых результатов. Не вызывает сомнение, что интерпретация степени «полезности» информации выражается на основе эксперных

оценок и опирается на методы интеллектуального анализа вышеуказанных характеристик.

В направлении Data mining для извлечения полезной информации решают задачи дескриптивного и прогностического (предикативного) анализа.

К первому классу описательных задач относят:

- задачи создания группировок (сегментация, кластеризация, анализ формальных понятий);
- задачи извлечения свойств данных (классификация, лингвистическое резюмирование, поиск аномалий).

Во втором классе содержатся задачи предикативного анализа:

- задачи построения пространственных и темпоральных зависимостей (прогнозирование, корреляционный анализ, регрессионный анализ, поиск ассоциативных правил) ;
- задачи построения зависимостей в пространстве признаков (классификация).

2.2. Соотношение задач проектной деятельности с задачами Data Mining

Применительно к САПР извлечение полезной информации нацелено на повышение эффективности труда инженеров за счет сокращения трудоемкости и сроков проектирования. Обычно это обеспечивается за счет решения следующих задач проектной деятельности:

1. автоматизации оформления документации (документирование);
2. информационной поддержки и автоматизации процесса принятия решений (оценивание и принятие решений);
3. унификации проектных решений и процессов проектирования для повторного использования проектных решений, данных и наработок (применение аналогов);

4. стратегического проектирования (исследование проектной ситуации, моделирование связей и зависимостей);
5. замены натурных испытаний и макетирования математическим моделированием;
6. повышения качества управления проектированием (поиск несоответствий и новизны, поиск связей);
7. применения методов вариантного проектирования и оптимизации (поиск связей, оценивание и принятие решений).

Связь вышеперечисленных задач, решаемых в проектировании объектов с основными классами задач Data mining представлена в таблице 1.

Таблица 1

Связь задач проектирования с методами Data mining

	Сегментация/ Кластеризация / Анализ формальных понятий	Класси-фикация	Поиск анома-лий	Поиск ассоциаций/ корреляций	Прогнози-рование	Резюмиро-вание/ Агрегация
Исследование проектной ситуации	+	+	+			
Поиск аналогов	+	+		+		
Поиск несоответствий и новизны	+		+	+		
Моделирование связей		+			+	
Поиск связей	+			+	+	
Оценивание вариантов для принятия решений	+	+				+

2.3. Контрольные вопросы

1. Какие виды информации извлекаются при решении задач Data Mining?
2. Приведите интерпретацию термина «полезные знания» для проектировщика, извлеченные методами Data Mining.
3. Какие типы задач рассматриваются в Data Mining? Изобразите графическую классификацию задач Data Mining.
4. В чем сущность проектной деятельности и каковы основные этапы проектирования?
5. За счет каких технологий возможно повысить эффективность труда проектировщиков?
6. Приведите определение системы автоматизации проектирования (САПР).
7. Дайте краткую характеристику процессу автоматизации проектирования.
8. Приведите перечень задач автоматизации проектирования и их связь с задачами Data Mining.
9. Создайте схему, связывающую этапы проектирования программных продуктов в гибкой методологии проектирования с задачами Data Mining.
10. Приведите примеры возможного применения решения задач Data Mining в области программной инженерии. Какую информацию можно извлечь при разработке программных продуктов?

3. ФОРМАЛЬНАЯ ПОСТАНОВКА ОСНОВНЫХ ЗАДАЧ DATA MINING

Согласно выделенным пяти этапам концептуального проектирования систем ИАД, приведенным в главе 1, первым этапом является определение формулировки решаемой задачи. Этот этап является наиболее важным, так как от корректности и адекватности формулировки проблемы зависит эффективность всего проекта, и неточности на этом этапе могут привести к значительным экономическим потерям.

3.1. Введение в постановку задач Data Mining

В общем виде постановку любой задачи можно представить в виде самой общей системной модели «вход-выход»:

$$Y = F(X), \tag{1}$$

где Y это требуемый результат, X это набор исходных данных (база данных), F это некоторая система, преобразующая исходные данные в требуемый результат.

Обычно исходные данные и требуемый результат задаются экспертом предметной области, а задачей разработчика системы Data Mining является спроектировать и реализовать ее. Для этого проектировщику необходимо определить подходящий класс задач Data Mining, выбрать модели представления исходных данных, методы обработки и хранения данных, обосновать архитектуру системы F. Отметим, что для выбора метода (алгоритма) решения задачи Data Mining определяющими являются представление исходных и выходных данных, точность и время. Часто выбираются несколько конкурирующих методов и среди них происходит поиск наиболее эффективного (по критерию точность/время).

Рассмотрим назначение и формальную постановку основных задач Data mining, используемых для извлечения информации из баз данных, хранящих множество объектов в виде совокупности их атрибутов.

Исходная база данных в выражении (1) может содержать:

1. «Сырые данные» (raw data). Это данные наблюдений или измерений.
2. Преобразованные данные, полученные в результате некоторого алгоритма.
3. Свойства (feature). Это преобразованные «сырые» данные, характеризующие некоторые аспекты данных.
4. Параметры моделей, описывающих свойства «сырых» данных.
5. Дополнительные данные, необходимые для решения задачи.

При рассмотрении формальных постановок задач Data Mining используем подход, ориентированый на двумерное представление базы данных (исходных данных для (1)), предложенный C. C. Aggarwal [1] без рассмотрения вариантов представлений данных.

Определение 1.1. База многомерных данных MD(n × d) это множество из n записей X_1 , …, X_n, таких, что каждая запись X_i ($i = 1,2,\ldots,n$) состоит из значений ($x^1{}_i$,..., $x^d{}_i$), где d количество атрибутов.

Используя теорию отношений, приведем определение исходной базы данных для выражения (1) в виде многозначного контекста.

Определение 1.2. Многозначный контекст это четверка [2]:

$$K = (G, M, W, J),$$

где G представляет собой совокупность объектов (строки базы данных MD), M это набор многозначных атрибутов (столбцы базы данных MD), W это набор значений атрибутов и J это отношение, $J \subseteq G \times M \times W$,

такое, что для любого $g \in G$, $m \in M$ существует по крайней мере одно значение $w \in W$, удовлетворяющее $(g,m,w) \in J$. Выражение $(g,m,w) \in J$ обозначает, что объект g имеет атрибут m со значением w.

Рассмотрим в следующих разделах постановку основных задач Data Mining.

3.2. Постановка задач кластеризации и сегментации

Кластеризация и сегментация предназначениы для поиска и группировки сходных, похожих, аналогичных объектов, имеющих близкие по некоторой метрике значения.

Сегментация применяется в том случае, если исходные данные однородные и представлены вектором. В более сложном случае, когда данные однородны и представлены в многомерном виде, применяют кластерный анализ.

Определение 2.1. (Data Clustering) Для базы данных MD, определить ее разбиение по строкам на множество кластеров (групп) $C_1,\ldots, C_k$, так чтобы в каждом кластере содержались похожие («similar») строки, а в разных – непохожие.

Рассмотрим две записи $X = (x_1 , \ldots , x_d)$ и $Y = (y_1 , \ldots , y_d)$. Естественно предложить способ вычисления сходства между X и Y в виде суммы мер сходства $S(x_i, y_i)$, вычисленных для каждого атрибута.

Простейший вариант выбора меры $S(x_i, y_i) = 1$, если $x_i = y_i$, иначе $S(x_i, y_i) = 0$. Однако сходство или различие, которые редко появляются, статистически более значимы. В контексте категориальных данных статистические свойства данных важно использовать при вычислении сходства. Такие глобальные статистические свойства данных применены в

метрике *Mahalanobis* для вычисления более точной меры сходства. Идея этой метрики в том, чтобы нетипичные значения категориальных атрибутов имели большие значения весов, чем значения которые встречаются часто.

3.3. Постановка задачи классификации

Целью классификации является распределение множества объектов по заранее заданным классам (группам), так чтобы в каждом классе находились похожие объекты. В отличие от кластеризации задача классификации формулируется в условии, когда имеется в наличии обучающее множество: база данных объектов MD(n × d), и для каждого ее объекта определена метка класса из некоторого множества.

Классификация относится к задачам, требующим обучения с учителем по известным примерам. При обучении с учителем база данных MD(n × d) и ассоциированный с ней вектор меток классов Cl(n) разбивают на два множества:

- Обучающее множество (training set), множество примеров с заданными классами, использующееся для обучения (конструирования) модели (классификатора).
- Тестовое (test set) множество, это множество также содержит входные и выходные значения примеров. Здесь выходные значения (метки классов) используются для проверки обученности классификатора.

Точность (и ошибка) классификации на обучающем и тестовом множестве оценивается по критерию совпадения (и несовпадения) полученных меток классов с известными. Для этих целей применяют также и *кросс-проверку* (cross-validation): (a) точность классификации тестового множества сравнивается с точностью классификации

обучающего множества; (б) если классификация тестового множества дает приблизительно такие же результаты по точности, как и классификация обучающего множества, считается, что данная классифицирующая модель прошла кросс-проверку.

Для оценивания точности классификатора разделение на обучающее и тестовое множества рекомендуется осуществлять *k*-раз (*k-cross-validation*) с последующим усреднением методом случайного деления выборки в определенной пропорции, например, обучающее множество включает две трети данных, а в тестовом наборе данных содержится одна треть данных.

Определение 3.1. (Data Classification) Для исходной базы данных MD(n × d) и для ассоциированного с ней вектора меток классов Cl(n), значения которого ассоциированы с каждой из *n* строк (записью в MD), требуется создать модель (классификатор) MA для определения (предсказания) метки класса Cl(n+1) для (n+1)-й d-размерной записи.

Классификация может быть использована также и в качестве средства прогнозирования значения, представленного в виде метки класса.

3.4. Постановка задачи прогнозирования

В отличие от вышерассмотренной формулировки задачи классификации задача прогнозирования в более общем виде формулируется как задача идентификации зависимости данных или от моментов времени, или от предыдущих значений. В этом случае записи базы данных рассматриваются в виде упорядоченной по моментам времени последовательности значений, используемой в качестве обучающих примеров. Результатом решения задачи прогнозирования являются новые записи в базе данных для очередных моментов времени.

Определение 4.1. (Data Forecasting) Для базы данных MD(n×d) и ассоциированного с ней вектора моментов времени Y(n) требуется создать модель (алгоритм) MF, которая для значений Y(n+1), Y(n+2),…, Y(n+p) может вычислять значения строк мартицы MD((n+1)×d), MD((n+2)×d),…, MD((n+p)×d). Переменная *p* называется горизонтом прогноза. В этом случае матрица MD(n×d) вместе с вектором Y(n) образуют обучающее множество для построения модели MF.

Для оценивания точности модели прогнозирования применяют меры отклонения полученного значения от реального. Так же, как и в задаче классификации, в задаче прогнозирования может быть применена и *k*-значная кроссвалидация.

Таким образом, прогнозирование решает задачу вычисления нового, ранее неизвестного значения по исторически накопленным данным на основе некоторой модели зависимости. Эта модель может быть представлена в виде параметрической функции, параметры которой необходимо оценить или на основе некоторых правил следования, извлекаемых из данных.

3.5. Постановка задачи поиска ассоциативных правил

Поиск ассоциативных правил сфокусирован на обнаружении часто и совместно встречающихся наборов значений в строках базы данных MD(n × d). Введем определение часто встречающихся наборов значений (паттернов).

Определение 5.1. (Frequent Pattern Mining) Для бинарной матрицы MD(n × d), требуется определить все подмножества столбцов, такие что все значения в этих столбцах принимают значения 1 не менее чем для *s*-той доли строк этой матрицы. Значение *s* называют минимальной *поддержкой* (support) этого паттерна.

Задача поиска ассоциации в базе данных содержательно формулируется в виде извлечения ассоциативных правил (Association Rules), описывающих часто встречающиеся паттерны в значениях строк матрицы MD(n × d).

Наборы значений (или паттерны), которые удовлетворяют требованиям минимальной *поддержки* называют частыми. Частые наборы могут использоваться для создания правил ассоциации с использованием меры, известной как *достоверность* (confidence).

Определение 5.2. *Достоверность правила* X ⇒ Y является условной вероятностью того, что строка матрицы MD(n × d) содержит набор *Y*, при условии, что она содержит и набор значений *X*. Эта вероятность оценивается величиной, равной результату деления *поддержки* множества $X \cup Y$ на *поддержку* множества *X*.

Для оценки полезности извлеченного правила вычисляют показатель, называемый *улучшение* (improvement). Он позволяет судить, полезнее ли это правило случайного угадывания.

Определение 5.3. *Улучшением правила* X ⇒ Y является отношение числа строк матрицы MD(n × d), содержащих наборы $X \cup Y$, к произведению количества строк, содержащих набор *X*, и количества строк, содержащих набор *Y*.

Если улучшение правила больше единицы, то это значит, что с помощью ассоциативного правила X ⇒ Y предсказать наличие набора *Y* вероятнее, чем случайное угадывание,□□а если меньше единицы, то наоборот.

Определение 5.4. (Association Rules) Пусть *A* и *B* два набора значений в строках бинарной матрицы MD(n × d). Правило A ⇒ B является валидным на уровне поддержки *s* и на уровне достоверности *c*, если удовлетворены условия:

1. Уровень поддержки множества *A* не меньше *s*.

2. Уровень достоверности правила A ⇒ B не меньше *c*.

3. Улучшение правила A ⇒ B больше 1.

Правила ассоциации могут рассматриваться как модель зависимости типа «Если-То» с некоторой степенью достоверности. В этом случае извлеченные ассоциативные правила полезно использовать в качестве предикативной модели, позволяющей по набору *A* спрогнозировать набор *B*.

3.6. Постановка задачи поиска и обнаружения аномалий

Рассмотрим задачу поиска и обнаружения аномалий (или исключений). В общем случае эта задача может быть сформулирована в следующих вариантах:

1. Задача поиска и обнаружения аномалий, вытекающих из контекста наблюдаемых данных. Эта задача формулируется при условии, что известна информация об ограничениях, о допустимых значениях, свойствах или поведении данных.

2. Задача поиска и обнаружения аномалий при условии наличия априорной информации о данных с имеющимися аномалиями и данных без аномалий (задача классификации данных).

3. Задача поиска и обнаружения аномалий в условиях неопределенности.

Последний вариант задачи поиска и обнаружения аномалий может быть сформулирован в предположении, что аномальный объект характеризуется наличием редких и нетипичных значений (или некоторых свойств).

Аномалии классифируются на одиночные (выбросы) и на множественные. Ниже приведена формулировка постановки задачи поиска и обнаружения аномалий типа одиночный выброс в условиях неопределенности.

Определение 6.1. (Outlier Detection) Для базы данных MD(n × d), определить строки, такие, что их значения значительно отличаются от значений в остальных строках. Такие значения называются аномальными выбросами или аномалиями.

Как следует из данного определения, поиск аномальных выбросов можно рассматривать с точки зрения обнаружения значительного несходства между строками, а значит, на основе кластеризации данных.

С другой стороны, задача поиска редкости значений (паттернов) обратна задаче поиска частых паттернов, рассмотренных в задаче поиска ассоциацивных правил.

3.7. Постановка задачи поиска концептов на основе формального концептуального анализа

Формальный концептуальный анализ (Formal Concept Analysis (FCA)) предназначен для построения отношений между концептами. В более частном случае – для создания группировок данных, обладающих общими свойствами, но возможно, не сходными значениями. Такие группы называют концептами. Этот класс задач анализа основан на математической теории отношений подобия.

Отличительной особенностью формального концептуального анализа является то, что с его помощью можно в автоматическом режиме выделять схожие по признакам группы объектов и строить отношения между ними. Теория концептуального подобия разработана Гантером (Ganter) и Вилле (Wille) (1989) [3]. Общий процесс в концептуальном

подобии начинается с представления знаний о предметной области в виде таблиц данных с произвольными значениями и, возможно, пропущенными значениями. Эти таблицы данных, являясь по сути базой данных MD(n × d), в формальном концептуальном анализе получили название *многозначного контекста (*A many-valued context). Напомним определение многозначного контекста, приведенного в начале главы 3.

Определение 7.1. Многозначный контекст это четверка [2]

$$K = (G, M, W, J), \tag{2}$$

где G представляет собой совокупность объектов (строки базы данных MD), M это набор многозначных атрибутов (столбцы базы данных MD), W это набор значений атрибутов и J это отношение, $J \subseteq G \times M \times W$, такое, что для любого $g \in G$, $m \in M$ существует по крайней мере одно значение $w \in W$, удовлетворяющее $(g,m,w) \in J$. Выражение $(g,m,w) \in J$ обозначает, что объект g имеет атрибут m со значением w.

В этом случае многозначные атрибуты m обычно интерпретируется как функция (частичного) шкалирования и обозначается в виде $m(g) = w$ при $(g,m,w) \in J$.

Определение 7.2. Формальный контекст это тройка

$$C = \langle G, Y, I \rangle, \tag{3}$$

где G представляет собой совокупность объектов, Y это множество атрибутов, $I \subseteq G \times Y$ это бинарное отношение между G и Y. При этом семантика выражения $\langle g, y \rangle \in I$ такова: «Объект g имеет атрибут y».

То есть каждое $y \in Y$ является свойством на W (некоторым подмножеством значений $w \in W$). Это означает, что для каждого атрибута $m \in M$ в многозначном контексте на множестве его возможных значений

$w \in W$ необходимо построить некоторое шкалирование для преобразования каждого значения $w \in W$ в некоторое свойство $y \in Y$, которое может быть выражено числом, интервалом, лингвистическим термом или нечетким интервалом с функцией принадлежности.

Определение 7.3. *Концепт* (понятие) это максимальная коллекция объектов (подмножество объектов *G*), которая характеризуется сходными свойствами (**атрибутами подобия)** $y \in Y$, то есть это группировка объектов, обладающих подобными свойствами.

Задача формального анализа заключается в определении концептов и отношений между ними, что можно выполнить на основе предварительного преобразования многозначного контекста в формальный контекст: $\boldsymbol{K \Rightarrow C}$. В результате анализа можно по общим свойствам сгруппировать объекты и построить концептуальную решетку в виде графа, отображающего объекты, имеющие общие свойства и отношения частичного порядка. На основе концептуальных решеток возникает возможность находить знания о группах, схожих по имеющимся свойствам в гетерогенных атрибутах, и находить различные зависимости, строить иерархии и онтологии, в том числе в области программной инженерии [4, 5].

Так как формальный концептуальный анализ не достаточно широко представлен в литературе приведем краткий пример, поясняющий многозначный и формальные контексты.

Пример. На рисунке 2 приведен многозначный контекст, а на рисунке 3 показан построенный формальный контекст на основе некоторой шкалы и его концептуальная решетка.

K_0	sex	age
ADAM	m	21
BETTY	f	50
CHRIS	/	66
DORA	f	88
EVA	f	17
FRED	m	/
GEORGE	m	90

Рис. 2. Пример многозначного контекста

	sex		age				
K	m	f	<18	<40	≤65	>65	≥80
ADAM	×			×	×		
BETTY		×			×		
CHRIS						×	
DORA		×				×	×
EVA		×	×	×	×		
FRED	×						
GEORGE	×					×	×

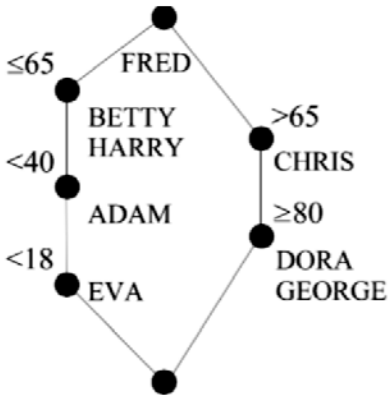

Рис. 3. Формальный контекст и его концептуальная решетка, отображающая объекты, имеющие общие свойства и отношения частичного порядка

3.8. Постановка задачи резюмирования и агрегации

Резюмирование определяется как некоторая общая характеристика множества объектов, представленных в виде многозначного контекста или в виде формального контекста (например, в виде (2) или (3)), по отношению к отдельному атрибуту или отдельному свойству этого атрибута. Результат резюмирования может быть выражен в виде количественной характеристики операции агрегации: например, суммирования объектов в каждом кластере, вычисления среднего,

минимального, максимального значения выборки или процента данных относящихся к искомому кластеру.

Лингвистическое резюмирование позволяет перейти от количественных характеристик к лингвистическим характеристикам (лингвистическим оценкам), понятным человеку и выражающим полезную информацию.

Для лингвистического резюмирования объектов исследования по отдельному атрибуту (свойству) рекомендуется использовать базовый подход, предложенный Yager [6].

В рамках базового подхода рассматривается множество из n объектов в виде записей $G = \{g_1,\ldots, g_n\}$ в базе данных D $\{y(g_1),\ldots,y(g_n)\}$ с одним атрибутом (свойством) y, при этом $y(g_i)$ – это значение свойства для работника g_i .

Результат резюмирования с учетом некоторой неопределенности, моделируемой в виде степени истинности, состоит из следующих компонент:

g`s – обозначение объекта исследования (например, проектировщик) по отношению к атрибуту y , например, y = «зарплата»;
P – лингвистическое значение результата резюмирования (например, «низкий» для атрибута y = «зарплата»);
Q – лингвистический квантификатор (например, «большинство», «в целом», «меньшинство», «все»), основанный на количественной оценке исследуемых объектов по отношению к P;
T – степень истинности резюмирования/высказывания P.

Степень истинности необходима для выбора наиболее значимой характеристики при автоматическом резюмировании объектов исследования по множеству его атрибутов.

Таким образом, ядром лингвистического резюмирования является набор предложений, которые могут быть записаны в абстрактном виде:

Q g`s имеет значение P.

Пример такого предложения для базы данных *D* с одним атрибутом *Y* = «зарплата»:

«Меньшинство (*Q*) работников (*g`s*) имеют высокую (*P*) зарплату (*Y*)».

При реализации лингвистического резюмирования в таком формате предварительно требуется разработать лингвистическую шкалу для генерации нечетких лингвистических оценок с функциями принадлежности для атрибута *Y* и для получения качественной оценки *Q* на основе некоторого количественного распределения множества объектов.

Определение 8.1. Для базы данных *D* предполагается, что мера истинности результата *T* лингвистического резюмирования *P* по отношению к атрибуту (свойству) *Y* и характеристике *Q* выражает степень истинности суждения, что *Q* наборов объектов имеет значение *P* в атрибуте *Y* .

Фактически лингвистическое резюмирование может быть принято к рассмотрению в качестве метода получения множества агрегаций наряду с известными количественными методами агрегации (минимаксной, максиминной, Вальда и т. д.).

3.9. Контрольные вопросы

1. Приведите формальную постановку основных задач Data Mining.
2. Перечислите и охарактеризуйте основные задачи Data Mining.
3. Чем отличается классификация от кластеризации?
4. Приведите варианты, сходство и отличия задач группировки и их графическую иллюстрацию.
5. Приведите отличия и сходство задачи прогнозирования от задачи поиска ассоциативных правил с привлечением общей системной модели решения задач и на основе их формальной постановки.
6. Опишите сущность формального концептуального анализа и его применение.
7. Сформулируйте постановку и приведите примеры задачи лингвистического резюмирования.
8. Охарактеризуйте задачу поиска аномалий и сопоставьте ее с кластеризацией данных на разных уровнях представления.
9. Подберите адекватные формальные постановки задач Data Mining для решения следующих проблем:
   a. Разделить проекты на проекты, имеющие высокую степень успешности, среднюю и неуспешные, и определить их метрики.
   b. Найти аналогичные проекты по набору требований в виде гетерогенных параметров.
   c. Определить риски проектирования и реализации программного продукта по диаграмме выгорания.
   d. Спрогнозировать время выполнения проекта исходя из набора требований, KPI персонала и объема финансирования.
   e. Определить, какую квалификацию имеет «Большинство разработчиков».

4. ПРИМЕРЫ СИСТЕМ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Системы интеллектуального анализа данных (ИАД) – это класс программных систем поддержки принятия решений, в том числе проектных решений, задачей которых является автоматизация поиска скрытых, ранее неизвестных, содержательных и потенциально полезных закономерностей в больших объемах разнородных, сложно структурированных данных [7].

4.1. Основы разработки систем Data Mining

Целью создания интеллектуальных систем анализа и систем реализующих задачи Data Mining, в том числе является обработка и управление типами данных, образованными сложными слабоформализованными структурами для извлечения новых знаний о состоянии и тенденциях развития объектов предметной области.

Выделяют следующие базовые модули в системах, ориентированных на автоматизированный анализ данных на основе Data Mining: модуль предобработки данных, модуль идентификации модели, модуль настройки и оценивания модели, модуль применения методов Data Mining для выполнения анализа и получения результатов, хранилище, содержащее исходные данные и полученные результаты, база знаний, представляющая организованную совокупность знаний системы и проблемной области.

Проектировщику систем Data mining необходимо определить в первую очередь семантику системы, выраженную следующими стратами знаний [8]:

1. *Где-знания*. Область применения системы, пространственные ограничения, условия, соотношение с другими системами.
2. *Кто-знания*. Пользователи системы, компетенции и требования пользователей.
3. *Зачем-знания*. Цель применения системы и совокупность решаемых задач.
4. *Что-знания*. Входные и выходные данные системы. Их типы, структуры, модели.
5. *Как-знания*. Функции, операции, алгоритмы, определяющие функционирование системы.
6. *Почему-знания*. Значимость, актуальность применения системы.
7. *Когда-знания*. Временные ограничения.
8. *Сколько-знания*. Количественные показатели, постоянные, затраты, показатели эффективности.
9. *Какие-знания*. Качественные оценки системы.

Заметим, что страты знаний могут быть укрупнены в такие объекты системы, как вход-выход с возможными ограничениями (страты 4, 1, 2, 7), функции (страты 3, 5), критерии (страты 6, 8, 9). Эти укрупненные страты знаний могут быть взяты за основу при концептуальном проектировании.

Рассмотрим в рамках концептуального проектирования абстрактной системы Data mining *S* определение ее модели, ориентированной на лингвистическую трихотомию «ЗНАК-ЗНАЧЕНИЕ-ОБОЗНАЧЕНИЕ».

Обобщенная компонентная модель системы *S*, включающая вход *X*, выход *Y*, множество моделей преобразования входа в выход *F*, множество критериев *K*, используемых для идентификации модели, оценки ее

параметров и критериев, определяющих качество моделей F, представима в виде $Cs = <Y, X, F, K>$. Эта традиционная модель системы может иметь, по крайней мере, двухуровневое представление для задач в конкретной предметной области: на одном уровне – в терминах конечного пользователя, на другом – в терминах математических моделей и методов.

Каждая компонента модели Cs характеризуется своей семантикой, имеющей внутреннюю и внешнюю интерпретации. Внешняя интерпретация семантики задает интенсиал модели S в форме лингвистического представления каждой компоненты, выраженной в терминах конечного пользователя, а внутренняя – определяет экстенсиал модели S в виде правил выполнения допустимых операций для каждой компоненты, совокупности отношений и математических зависимостей между компонентами модели. Таким образом, внутренняя семантика определяется моделями решения задач анализа и используется для генерации «ЗНАЧЕНИЙ», а внешняя семантика задает субъектно-объектное отображение предметной области и объектов известных исследователю формальных систем в компоненты модели S и определяет «ОБОЗНАЧЕНИЕ».

Определим для модели Cs системы S функционал $Rs = R(Y, X, F, K)$, задающий внутреннюю семантику, и функционал $Ps = P(Y, X, F, K)$ для представления внешней семантики.

Тогда концептуальная модель абстрактной системы Data mining S может быть определена в виде $Ms = < Cs, Rs, Ps>$, и в этой модели представлена лингвистическая трихотомия «ЗНАК-ЗНАЧЕНИЕ-ОБОЗНАЧЕНИЕ». Сопоставление компонент и семантических моделей системы S при анализе конкретного класса задачи Data mining обычно осуществляет проектировщик системы, который принимает наиболее адекватное решение исходя из контекста среды и своего опыта.

Классификация систем ИАД на системы общего и специализированных классов приведена в [10], там же представлены примеры приложений, в которых реализованы задачи ИАД. Ниже рассмотрим некоторые примеры систем ИАД.

4.2. Система экспресс-анализа экономической эффективности предприятий

Internet-сервис экспресс-анализа предприятий [tsas.ulstu.ru] на основе прогнозирования временных рядов и анализа нечетких тенденций [9] ориентирован на сегмент рынка электронных сервисов. Моделирование и прогнозирование технико-экономических показателей для экспресс-анализа предприятий реализовано на основе интегрального метода, включающего несколько нечетких моделей и моделей на основе нейронных сетей.

Основными решаемыми задачами экспресс-анализа предприятий в Internet-сервисе являются следующие.

- Расчет и прогнозирование значений технико-экономических показателей по данным открытой бухгалтерской отчетности.
- Моделирование, прогнозирование и лингвистическое резюмирование тенденций развития по расчетным и прогнозным технико-экономическим показателям.
- Экономический анализ и интерпретация результатов экспресс-анализа.

На рисунке 4 представлен отчет, формируемый в разработанной системе в результате решения задачи прогнозирования, состоящий из текстовой части и графической интерпретации.

ОТЧЕТ

ПО АНАЛИЗУ ДЕЯТЕЛЬНОСТИ

ЗАО «Электроприбор»

Раздел 1. Показатели финансовой независимости (финансовой устойчивости)

Коэффициент финансовой независимости в части оборотных средств отражает долю участия собственного капитала предприятия в формировании его активов. Рекомендуемое значение: от 0,2 до 0,5.

По представленным Вами данным и проведенному экономическому анализу, система дает прогноз на следующий период: Предприятие имеет возможности для проведения независимой финансовой политики, рост показателя свидетельствует об укреплении финансовой независимости организации в части оборотных активов.

Прогноз

Дата	**Предполагаемое изменение**	**Предполагаемая интенсивность изменения**
2008-12-31	стабильность	малый

Рис. 4. Результат работы системы экспресс-анализа экономического состояния предприятия

Методика экспресс-анализа предприятия, реализованная в Internet-сервисе, предусматривает расчет и анализ четырех групп показателей, наиболее часто используемых в экономическом анализе деятельности предприятия:

1. показатели ликвидности и платежеспособности;
2. показатели финансовой независимости;
3. показатели рентабельности;
4. показатели деловой активности.

4.3. Система Data mining в задачах мониторинга поведения пользователей

Рассмотрим в сокращенном варианте пример из работы [10]. Задача заключается в проектировании системы мониторинга поведения пользователей некоторой системы для выявления возможных с их стороны угроз. Исходное предположение по результатам исследования заключается в том, что активность пользователей и программ можно полностью отследить и построить ее адекватную модель, чтобы в дальнейшем применять ее для классификации угроз.

Особенности проектируемой системы, решающей задачи выявления вторжений включают:

1. накопление исторической информации;
2. моделирование нормального поведение или вторжения;
3. поиск аномалий в поведении пользователей;
4. классификацию и анализ событий, которые описывают текущую активность в системе на соответствие построенным моделям.

Архитектура системы мониторинга поведения пользователей представлена на рисунке 5.

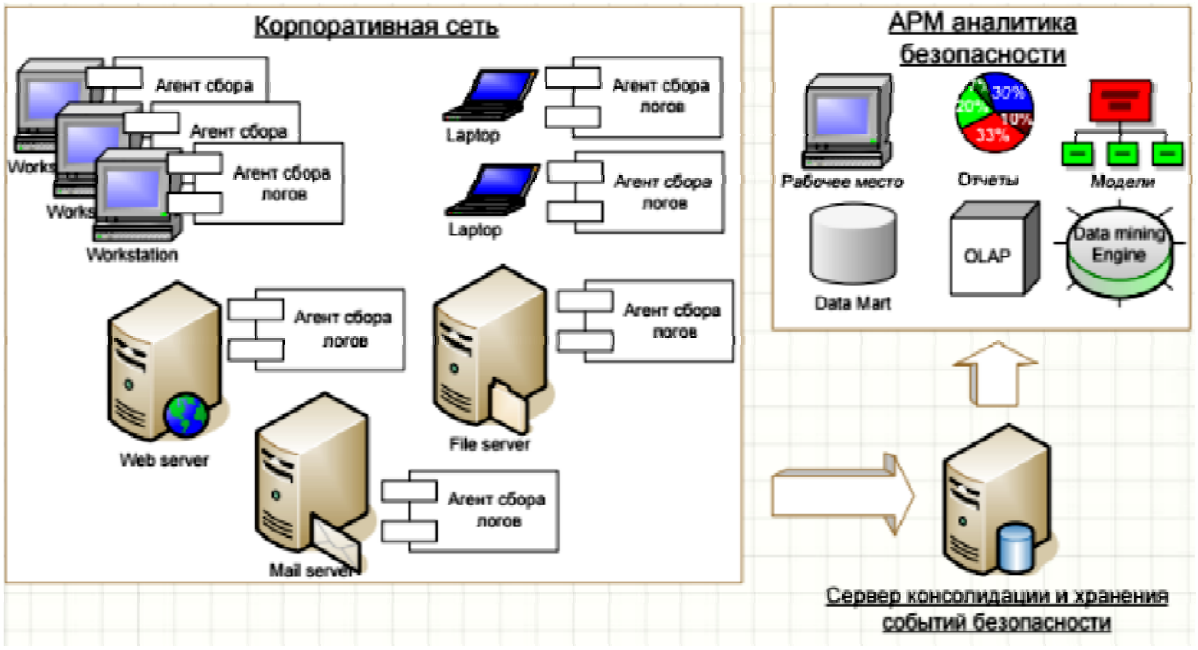

Рис. 5. Архитектура системы мониторинга поведения пользователей

4.4. Анализ и прогнозирование качества технологического процесса

При анализе и прогнозировании качества произведенной продукции возникает вопрос, какие параметры производственного процесса влияют на качество продукции. Для решения указанной проблемы была разработана система с алгоритмом на основе нечетких деревьев решений с поддержкой эволюционных методов оптимизации нечетких переменных и структуры правил [10], которая:

1. строит модель зависимости качества продукции от характеристик производственного процесса, представимую в виде системы нечетких правил «если ... то ... иначе»;
2. прогнозирует показатели качества изделия по характеристикам производственного процесса;
3. упорядочивает характеристики технологического процесса по степени влияния на качество.

4.5. Применение Data mining в образовательном процессе

Рассмотрим систему поддержки принятия решений по улучшению образовательного процесса [11]. Рассматриваются два класса объектов: O_1 – это множество студентов и O_2 – это множество дисциплин. Каждый студент O_{1j} характеризуется набором атрибут двух видов. Первый вид связан с его баллами по дисциплинам, второй вид атрибута описывает процент его контактной работы с образовательными ресурсами в процессе изучения дисциплины. Каждая дисциплина O_2 описывается следующими атрибутами: количество «5», количество «4», количество «3» и количество «2».

Постановка задачи включает решение двух частных задач [11]. Первая задача заключается в поиске разбиения студентов на группы на основе их атрибутов так, чтобы можно было определить успешных, стабильных и неуспешных студентов для дальнейшей корректировки образовательных технологий. Вторая задача формулируется как задача выделения групп дисциплин по усредненным характеристикам их освоения студентами. На рисунке 6 представлена модель архитектуры системы второго уровня для анализа образовательного процесса в высшем учебном заведении.

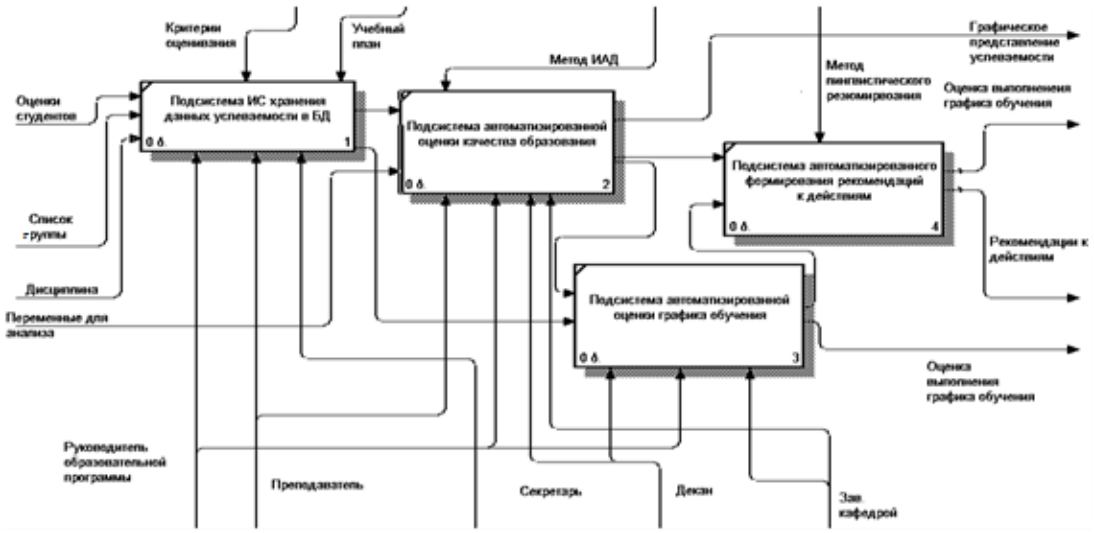

Рис. 6. Архитектура системы анализа образовательного процесса

На рисунке 7 изображена модель подсистемы, реализующей метод кластеризации.

Рис. 7. Подсистема реализующая метод Data Mining

Реализация данной системы позволяет удобно хранить данные об успеваемости студентов, быстро получать графическое представление, является инструментом поддержки принятия решений, направленных на улучшение образовательного процесса, основываясь на результатах интеллектуального анализа.

4.6. Применение Data mining в компьютерных играх

В большинстве компьютерных и мобильных игр в конце каждой игровой сессии игроку предоставляется информация о его результатах. Перспективным является, помимо рейтингов в компьютерных играх, предоставлять более полную информацию игроку на основе персонализированной обратной связи. На рисунке 8 изображена модель взаимодействия модулей в системе обеспечения обратной связи, ориентированная на проведение предварительной классификации игрока по набору игровых метрик, а затем выдачи персонализированной рекомендации [12].

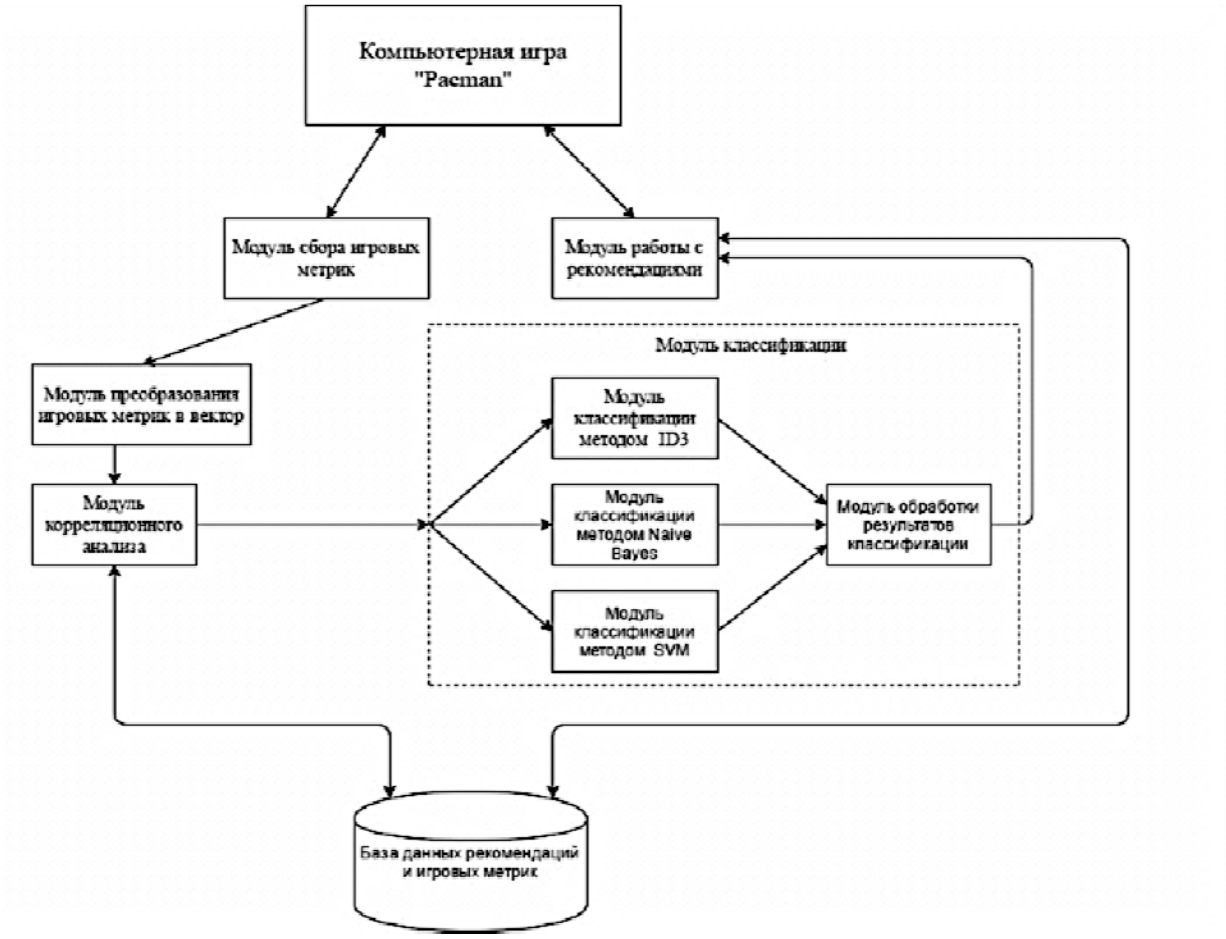

Рис. 8. Архитектура системы обеспечения обратной связи в компьютерных играх

Применение разработанной системы позволит увеличить результативность игровых сессий за счет оценки результативности игрока с использованием его класса.

4.7. Система прогнозирования процессов по временным рядам

Прогнозирование временных рядов предназначено для получения прогноза отдельного процесса. Решение таких задач востребовано в экономике, электроэнергетике, медицине, технике и управлении. На рисунке 9 изображена абстрактная модель разработанной системы [13] для прогнозирования временных рядов, использующая методы интеллектуального анализа. Актуальность подхода к анализу ВР, который

лежит в основе данного приложения, заключается в сочетании методов классического статистического анализа и теории нечетких временных рядов для получения максимально эффективного прогноза.

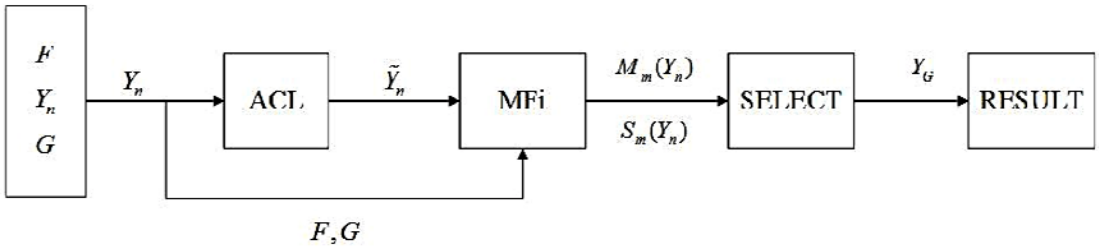

Рис. 9. Схема движения данных сервиса для прогнозирования ВР (где Y_n – числовой ВР; F – функция сглаживания ВР; G – горизонт прогноза ВР; $M_m(Y_n)$ – множество моделей ВР полученных в результате прогнозирования ВР; $S_m(Y_n)$ – моделей остатков ВР полученных в результате прогнозирования остатков ВР 4; Y_G – выбранный прогноз ВР

Стадии обработки данных в форме временного ряда формируют следующий список блоков-модулей:

1. Модуль ACL-шкалы (ACL), предназначенный для выполнения предобработки в виде преобразования исходных данных в нечеткое представление.

2. Модуль математических моделей прогнозирования (MFi). Здесь хранятся, оцениваются модели временных рядов различных научных подходов: нечеткие модели и статистические модели.

3. Модуль выбора наилучшей модели по результатам прогнозирования с использованием множества критериев (SELECT).

4. Модуль визуализации и резюмирования результатов (RESULT).

На первом этапе выполняется фаззификация временного ряда.

На втором этапе строятся нечеткие модели тренда и вычисляются прогнозные значения согласно горизонту прогноза G. Параллельно с этим выполняется прогнозирование остатков временного ряда в блоке с помощью стохастической модели ARIMA. На выходе блока формируется

множество моделей прогноза основного ВР, а также множества моделей остатков.

На третьем этапе выполняется выбор наилучшей модели прогнозирования на основании анализа критериев качества полученных прогнозов основного временного ряда и остатков.

4.8. Система оценивания стоимости нового объекта на рынке

Вычисление цены объекта вступающего на рынок является актуальной проблемой. Рассмотрим приложение (рисунок 10), реализующее методы интеллектуального анализа данных в задачах построения зависимости стоимости объектов по набору их параметров [14]. Исходные данные для построения зависимости содержат обучающее множество объектов с их параметрами и известной стоимостью. Задача определения стоимости нового решалась двумя методами построения зависимостей: на основе линейной множественной регрессии и с использованием нейронной сети прямого распространения, обучаемой методом обратного распространения ошибки.

Обученная нейронная сеть сохраняется в файл, затем может быть использована для оценивания стоимости нового объекта.

4.9. Прогнозирование потребления электроэнергии

Накопленные данные о потреблении электроэнергии конечными потребителями позволяет на этой основе делать прогнозы, полезные компаниям-поставщикам энергоресурсов. Система, реализующая такие прогнозы, рассмотрена в [15], и ее архитектура представлена на рисунке 11.

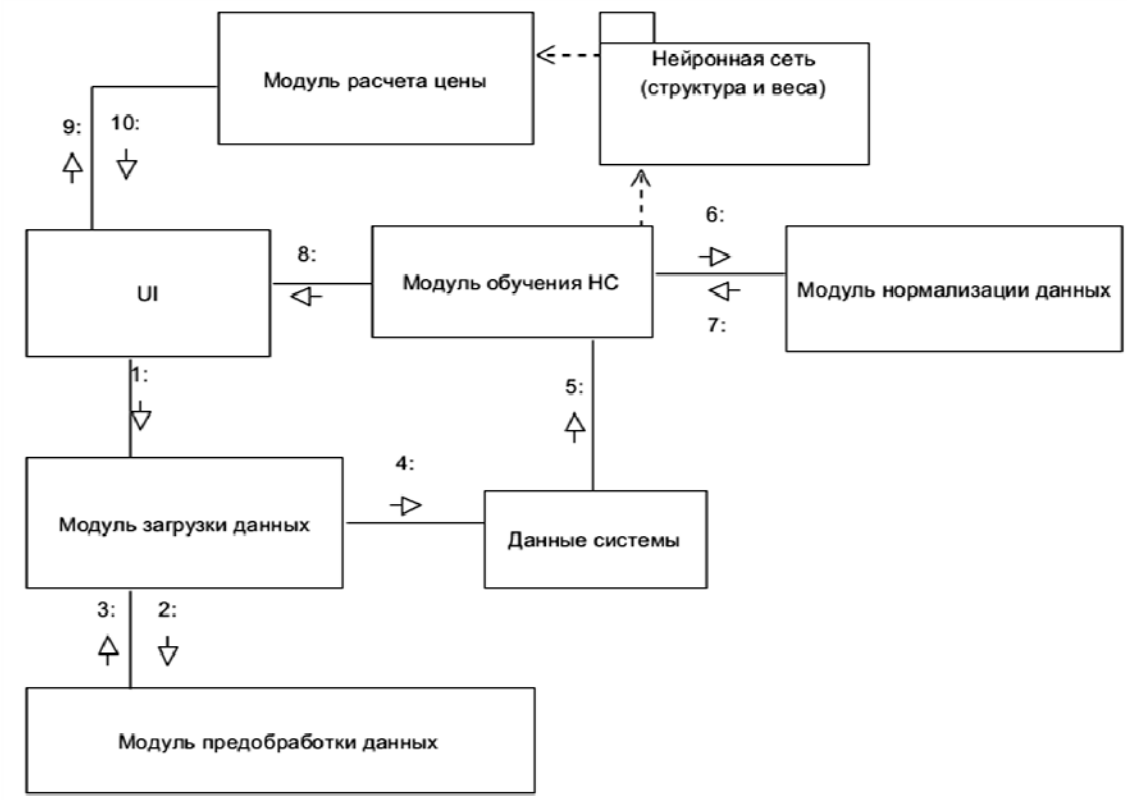

Рис. 10. Архитектура системы оценивания стоимости нового объекта рынка

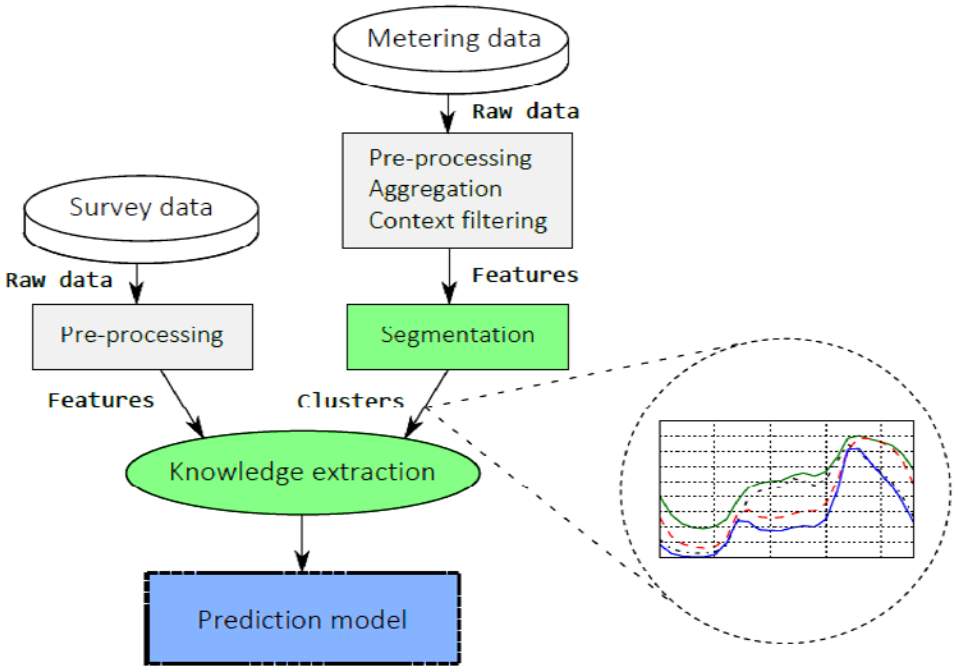

Рис. 11. Архитектура системы прогнозирования объемов потребления электроэнергии

Исходные данные представлены двумя видами:

- данные опроса семей-потребителей (survey data). Это информация о потребителях, проживающих в одной квартире (возраст, количество, семейный доход), и перечень используемой бытовой техники (холодильники, компьютеры, кондиционеры и т. д.);
- показания приборов энергетического учета (metering data).

В системе реализуются следующие методы обработки данных:

1. Предобработка данных (pre-processing) для восстановления пропущенных значений, для выбора свойств (features) групп потребителей (по уровню дохода, количеству членов семьи и т. д.).
2. Фильтрация данных (context filtering) для выбора показаний за сезон (например, лето или зима), для выбора дневных/вечерних показаний.
3. Формирование (aggregation) 24-часового трафика потребления электроэнергии.
4. Сегментация 24-часового трафика потребления электроэнергии (segmentation) для получения кластеров (см. рисунок 11). В качестве алгоритма сегментации используется кластеризация методом fuzzy c-means.
5. Извлечение моделей (паттернов) зависимостей объемов потребления от свойств семей-потребителей и кластеров трафика электро-потребления (knowledge extraction).
6. Прогнозирование паттернов потребления электроэнергии в каждом полученном кластере на основе извлеченной информации. Для прогнозирования использованы три различных метода: регрессия, нечеткая модель прогнозирования Такаги-Сугено и классификатор на основе метода опорных векторов.

4.10. Контрольные вопросы

1. Опишите типовую архитектуру системы анализа данных Data Mining.
2. Охарактеризуйте концептуальное проектирование систем Data Mining.
3. В чем выражается семантика при проектировании систем Data Mining?
4. Какие страты знаний наиболее значимы для проектирования?
5. Приведите примеры систем Data Mining и укажите, какие задачи ИАД в них решаются и какие задачи могут быть решены дополнительно.
6. Сформулируйте формальную постановку задач для приведеных в этой главе примеров.
7. Сопоставьте примеры систем Data Mining, в которых решаются задачи прогностической аналитики, какие подходы в них используются?
8. В каких примерах систем Data Mining решаются задачи дескриптивного анализа?

5. ОСНОВНЫЕ СТАНДАРТЫ ПРОЦЕССА KDD&DM

В настоящем разделе рассмотрим стандартизованные модели (методологии), задающие этапы процесса интеллектуального анализа данных (KDD&DM), приведенные в работе [16]. Представленные методологии по сути описывают процесс KDD, который помимо основного процесса Data Mining (DM) включает процессы предобработки и постобработки. Этапы процесса KDD&DM, рассматриваемые с точки зрения реализации как отдельные компоненты, позволяют проектировщику систем интеллектуального анализа данных определять степень их автоматизированности и конструировать различные архитектуры.

На рисунке 12 показана эволюция стандартизованных методологий, полезных в разработке систем Data Mining. На рисунке 12 показано, что KDD выступает отправной точкой в методологии анализа, и методология CRISP-DM представлена как базовая методология, на которой основаны большинство последующих подходов. По результатам опросов KDnuggets, приведенных в [17], 42% опрошенных компаний использует методологию *CRISP-DM*, 10% – методологию *SEMMA*, 6% – собственную методологию организации, 28% – свою собственную методологию, другими методологиями пользуется 6% опрошенных. Не пользуются никакой методологией 7% опрошенных.

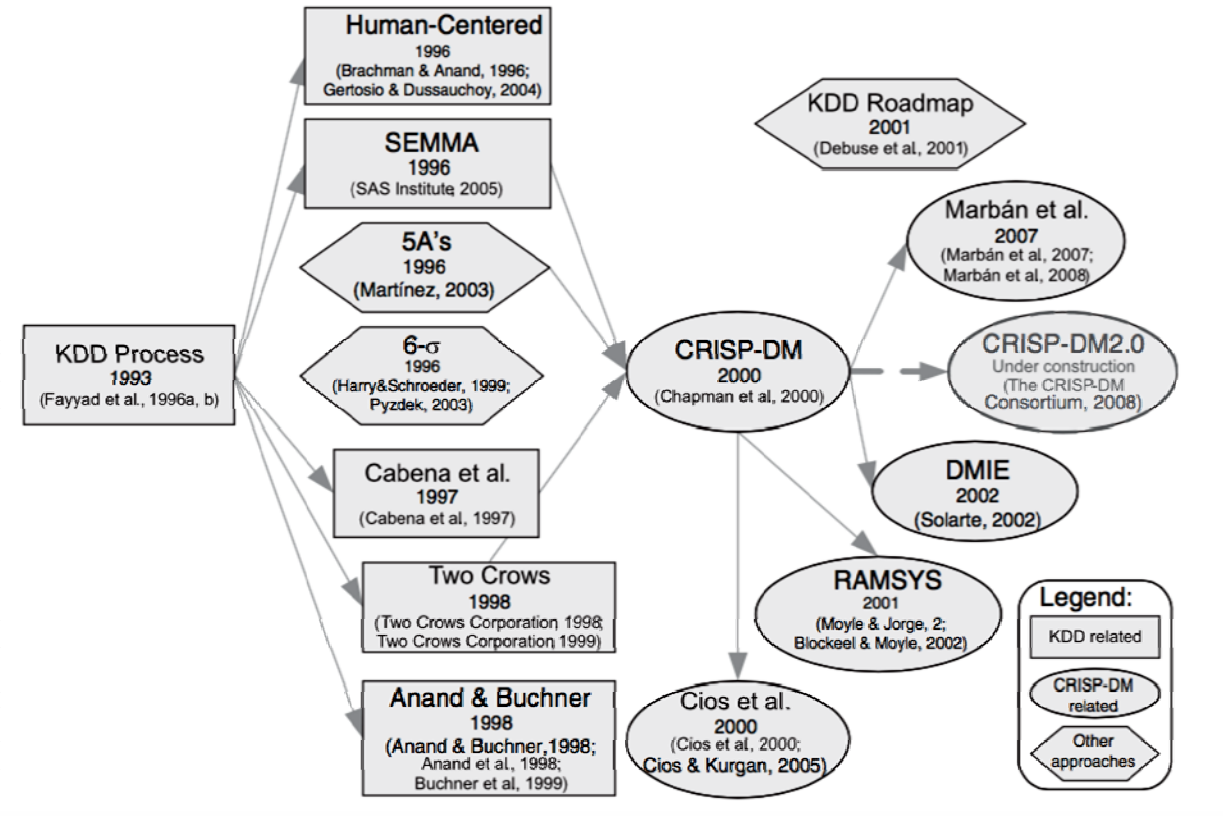

Рис. 12. Эволюция стандартизованных методологий в разработке систем Data Mining

5.1. Методология SEMMA

методология SEMMA реализована в среде SAS Data Mining Solution (1996 г.). Ее аббревиатура образована от слов:

S – Sample («Отбор данных», т. е. создание выборки),

E – Explore («Исследование отношений в данных»),

M – Modify («Модификация данных»),

M – Model («Моделирование зависимостей»),

A – Assess («Оценка полученных моделей и результатов»).

Процесс реализации анализа данных KDD&DM в соответствии с методологией SEMMA изображен на рисунке 13.

Рис. 13. Этапы методологии SEMMA [18]

Методология SEMMA подразумевает, что все процессы создания систем Data Mining выполняются для всех необходимых работ по обработке и анализу данных. Подход SEMMA сочетает структурированность процесса и логическую организацию инструментальных средств, поддерживающих выполнение каждого из шагов. Методология SEMMA упрощает применение методов статистического исследования и визуализации, позволяет выбирать и преобразовывать наиболее значимые переменные, создавать модели с этими переменными, чтобы предсказать результаты, подтвердить точность модели и подготовить модель к развертыванию.

Эта методология не навязывает каких-либо жестких правил. В результате использования методологии SEMMA разработчик может располагать научными методами построения концепции проекта, его реализации, а также оценки результатов проектирования системы KDD&DM.

5.2. Методология CRISP-DM

Методология CRISP-DM (The Cross Industrial Standard Process for Data Mining) (Chapman et al., 2000 г.) представляет документированную и свободно распространяемую модель, описывающую основные фазы, выполнение которых позволяет организациям получать максимальную выгоду от использования методов Data Mining. Эта методология является наиболее популярной и распространенной методологией. К наиболее важным факторам успеха CRISP-DM относятся платформо-независимость и возможность адаптации к различным прикладным областям.

В соответствии со стандартом CRISP-DM, KDD&DM является непрерывным процессом со многими циклами и обратными связями.

Жизненный цикл проекта системы интеллектуального анализа данных (рисунок 14) состоит из шести этапов. При этом последовательность этапов не является строгой. Стрелки указывают наиболее важные и частые зависимости между фазами. Внешний круг на рисунке указывает на цикличность интеллектуального анализа данных.

Уточнения, полученные в ходе процесса, могут породить другие более конкретные вопросы. Последующие этапы интеллектуального анализа данных извлекают выгоду из предыдущих.

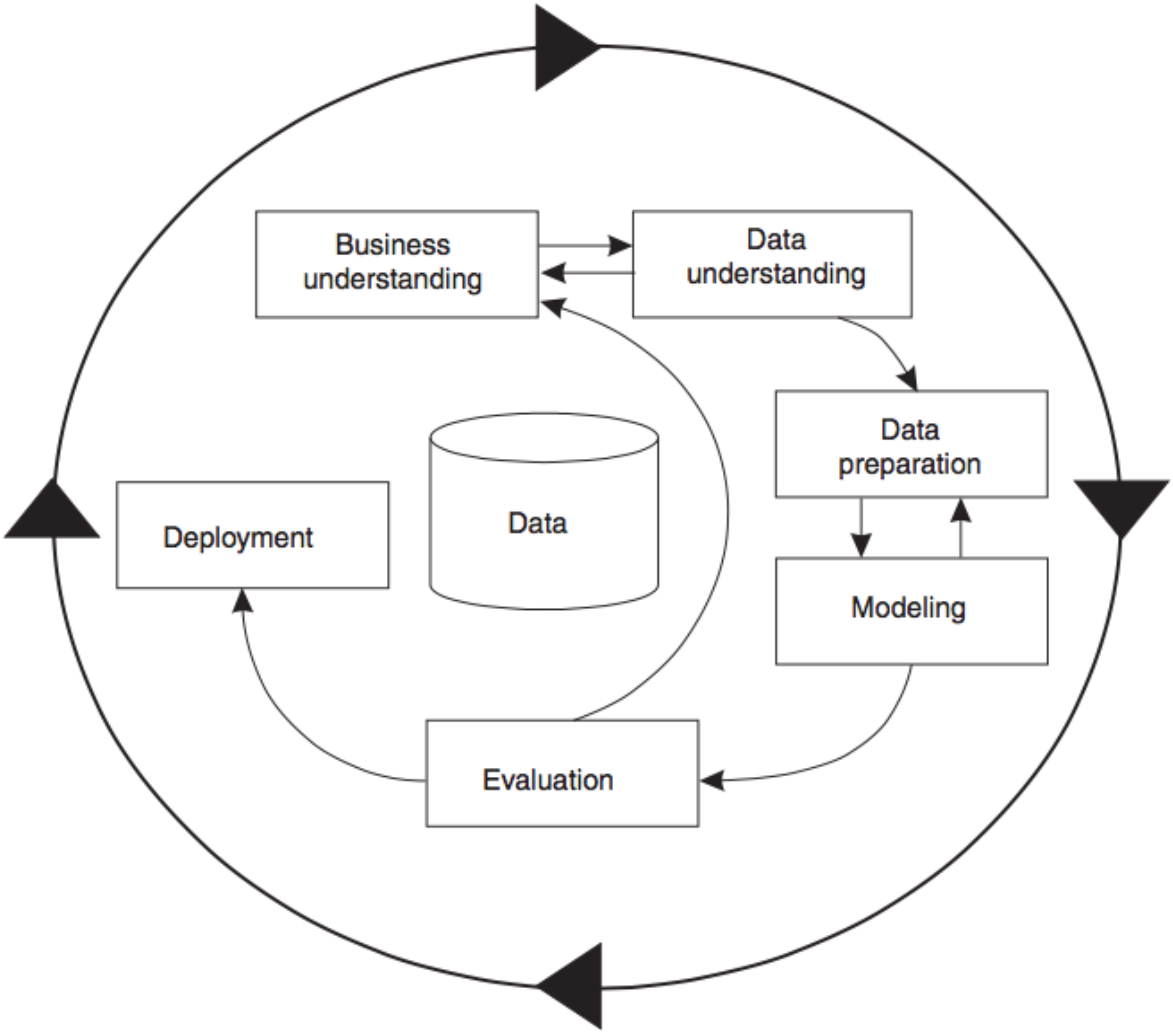

Рис. 14. Этапы KDD&DM по стандарту CRISP-DM

Рассмотрим подробнее этапы KDD&DM по стандарту CRISP-DM:

1. Осмысление бизнеса (Business understanding). Этот начальный этап посвящен исследованию цели проекта и требований с точки зрения бизнеса, а затем преобразованию этих знаний в *формальную постановку задачи* Data Mining, а также разработке предварительного плана, направленного на достижение целей.

2. Осмысление данных (Data understanding). Понимание данных начинается с первоначального сбора данных и переходу к ознакомлению с данными, выявлению проблем качества данных. Необходимо понять

структуру данных, обнаружить интересные подмножества для формирования гипотез с целью последующего анализа скрытых закономерностей.

3. Подготовка данных (Data preparation). Фаза подготовки данных охватывает все виды деятельности, чтобы определить окончательный набор данных из исходного набора данных. Задачи подготовки данных с большой вероятностью будут выполняться не один раз и могут выполняться также на последующих этапах. На данном этапе формируются таблицы с набором записей и атрибутов (база данных MD(nxd)), а также необходимые преобразования и очистка данных для моделирования.

4. Моделирование (Modeling). В этой фазе идет выбор методов Data Mining согласно определенной на первом этапе формальной постановки задачи и их применение. Кроме того, на этом же этапе идет подгонка параметров моделей Data Mining для извлечения полезной информации на обучающих и тестовых примерах.

5. Оценка результатов (Evaluation). Прежде чем приступить к окончательному развертыванию модели Data Mining важно более тщательно оценить эту модель и оценить все шаги построения модели, а также понять решает ли она задачу извлечения полезной информации для бизнес-задачи. В конце этой фазы принимается решение по использованию результатов интеллектуального анализа данных.

6. Внедрение (Deployment). Если модель Data Mining сформирована, это не означает что проект закончен. Полученная информация должна быть представлена таким образом, чтобы заказчик мог ее интерпретировать и использовать в своей работе. В зависимости от требований этап развертывания может быть как простым (простая генерация отчетов), так и более сложным (при котором может потребоваться повтор интеллектуального анализа данных).

5.3. Методология Cabena

В методологии Cabena (Cabena et al., 1997 г.) интеллектуальный анализ данных определяется как процесс извлечения ранее неизвестной, допустимой и полезной информации из больших баз данных для использования в принятии важных бизнес-решениях.

Этапы методологии Cabena представлены на рисунке 15. В этой методологии выделены следующие пять этапов: выборка, предобработка, трансформация, извлечение полезной информации. Заключительным этапом выступает анализ (интеграция) и интерпретация извлеченной информации с позиции полезных знаний для бизнеса. Полученные знания сохраняются и в дальнейшем используются для анализа. Достоинством рассматриваемой методологии является простая и понятная графическая интерпретация, удобная для проектирования систем KDD&DM. Согласно этой методологии был реализован интернет-сервис экспресс-анализа экономической деятельности предприятия, описанный в главе 4.

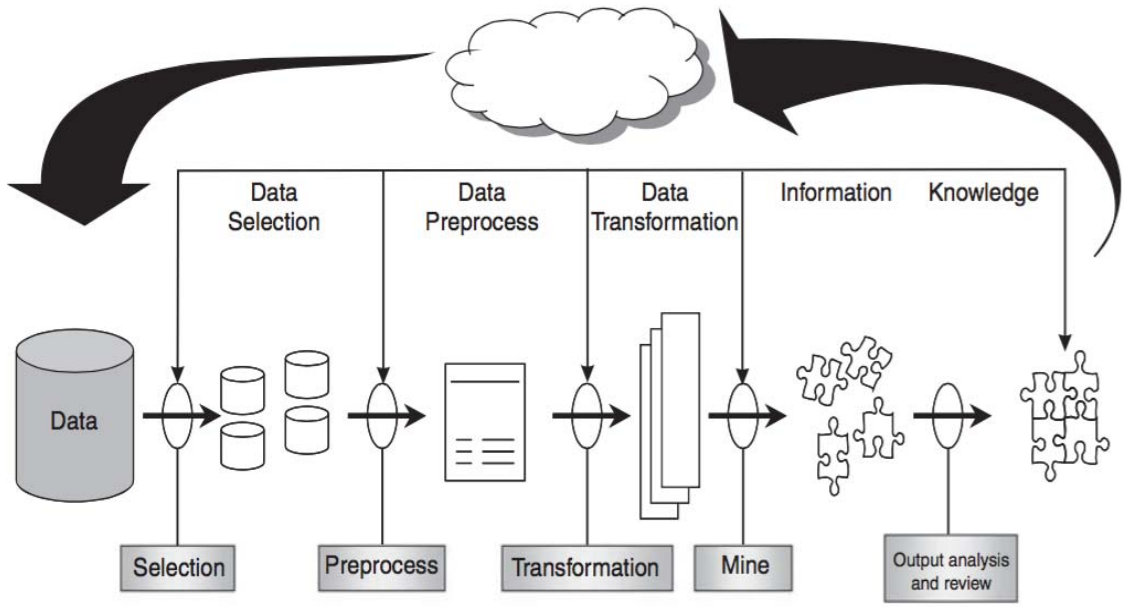

Рис. 15. Этапы методологии Cabena

5.4. Методология Two Crows

Методология Two Crows (1998 г.) очень близка к описанию базового процесса KDD&DM, хотя используются разные названия для аналогичных этапов.

Этапы методологии Two Crows представлены на рисунке 16. В отличие от методологии Cabena в методологии Two Crows выделяют семь этапов и показаны обратные связи не от последнего этапа к первому, а между промежуточными этапами, что является более информативным. Приведем основные этапы методологии Two Crows: определение бизнес-задачи, создание базы данных для Data Mining, исследование данных, подготовка данных для моделирования, создание модели для Data Mining, оценивание модели, применение модели и результатов.

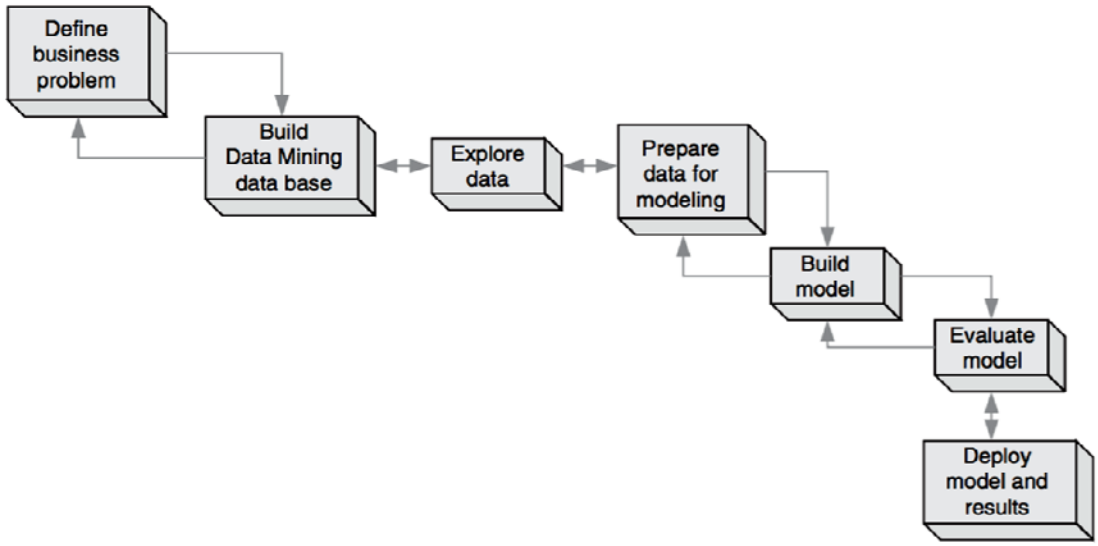

Рис. 16. Этапы методологии Two Crows

С другой стороны в графическом представлении этой методологии не приводится связь этапов с типами исходных и результирующих данных, а также не указана возможность хранения результатов анализа.

5.5. Методолгия RAMSYS

Методология RAMSYS (Rapid collaborative data mining system, 2001 г.) ориентирована на сочетание методологий решения проблем, обмена знаниями и легкости коммуникаций. В методологии RAMSYS включена новая задача, заключающаяся в представлении итоговой модели Data Mining в виде комбинации из набора лучших моделей, полученных на этапе оценивания. Этапы методологии RAMSYS представлены на рисунке 17.

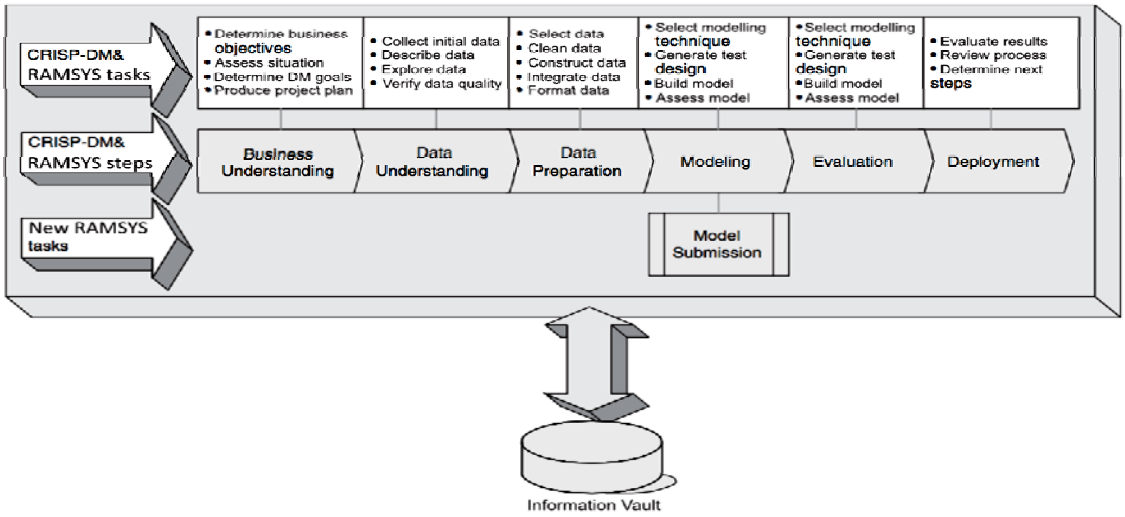

Рис. 17. Этапы методологии RAMSYS

Отличием графического представления этой методологии является формулировка задач для каждого этапа процесса KDD&DM, которые по сути являются названием микро-этапов.

5.6. Методология Five A's

Методология Five A's предложена корпорацией SPSS (2007 г.). Название этой методологии (см. рисунок 18) образовано по первым буквам этапов циклического процесса Data Mining: assess (оценивать), access (выбирать), analyze (анализировать), act (моделировать), automate (автоматизировать). Положительный аспект этой методологии в идее

автоматизации процесса интеллектуального анализа данных для того, чтобы пользователи, не являющиеся экспертами в области анализа данных, могли применить ранее полученные модели к новым данным.

Негативный аспект методологии Five A’s заключается в том, что не устанавливается иных альтернативных способов применения построенной модели или обнаружения знания. Еще один важным минусом является то, что методология не включает этап понимания данных, которые считается важным в CRISP-DM, чтобы понять и проверить качество модели, а также предотвратить возможные проблемы в ходе развития проекта.

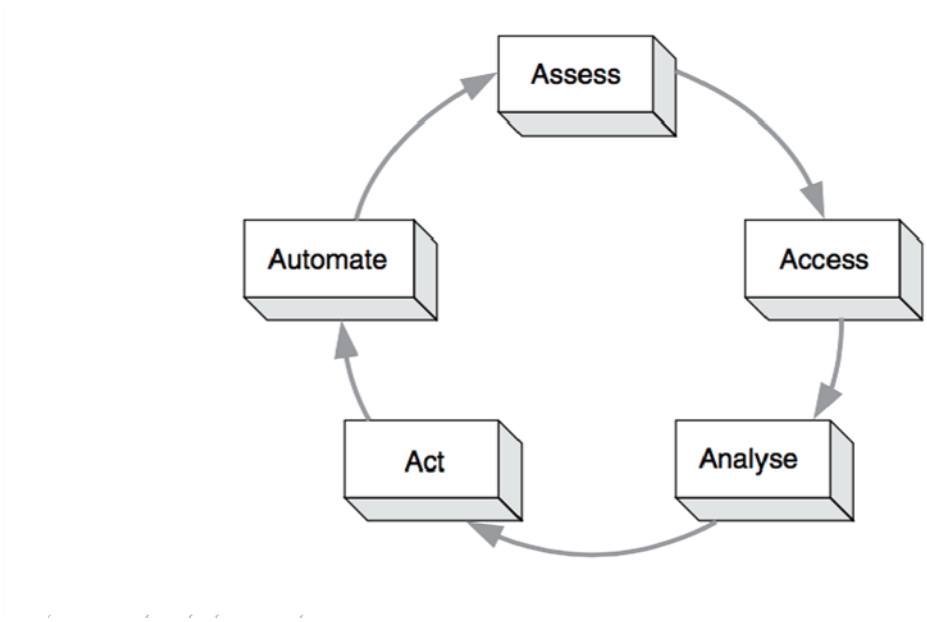

Рис. 18. Этапы методологии **Five A’s**

5.7. Методология Marba´n

Эта методология (Marba ´n et al., 2008 г.) основана на идее рассмотрения процесса Data Mining с позиции процесса проектирования и решения инженерных задач. Поэтому процессы разработки систем Data Mining содержат задачи и действия, необходимые в инженерной

деятельности, не включенные в стандарт CRISP-DM. Для этих целей предлагается основываться на двух апробированных стандартах в области разработки ПО: IEEE 1074 (IEEE, 1991) and ISO 12207 (ISO, 1995). Процессная модель управления разработкой системы Data Mining с позиции инженерных задач изображена на рис. 19. Она включает процессы улучшения, обучения и разработки. Процессы разработки основаны на:

- предобработке данных;
- обработке данных, ядром которой является KDD&DM процесс (выборка, трансформация, преобразование, Data Mining, интерпретация/оценка);
- постобработке результатов. В пост-обработке используются традиционные этапы в программной инженерии.

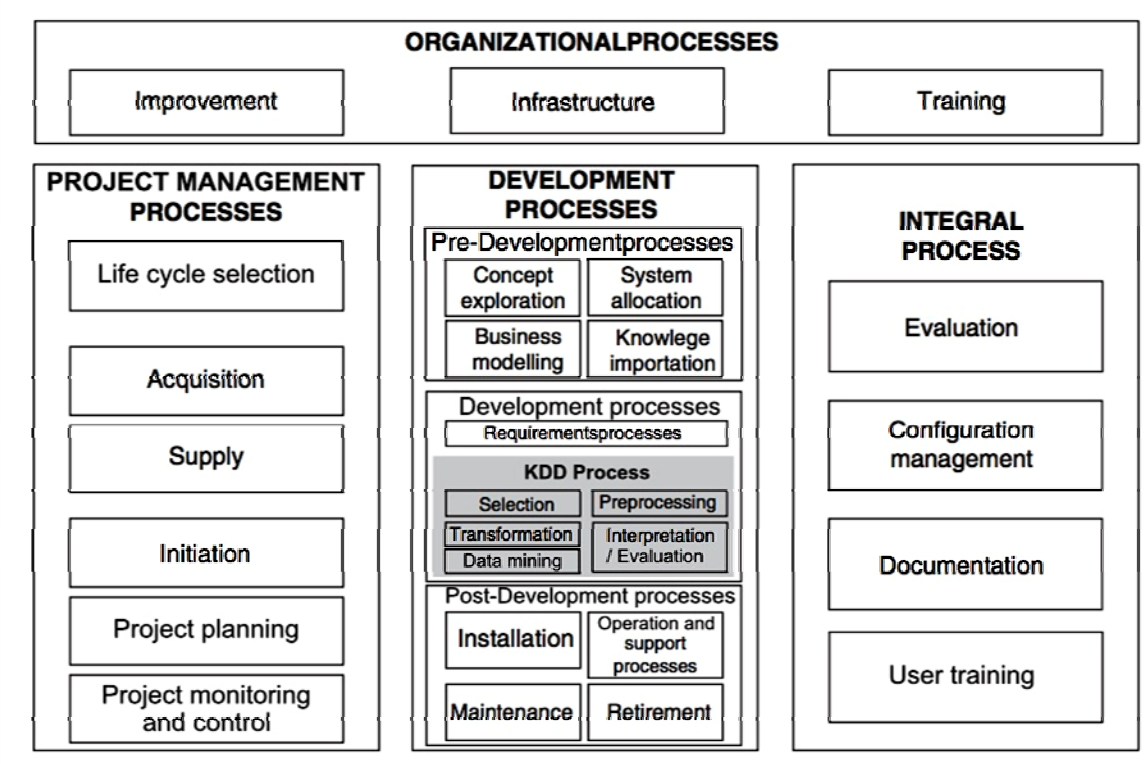

Рис. 19. Процессная модель системы data mining с позиции инженерных задач

5.8. Методология KDD Roadmap

KDD Roadmap (Debuse et al., 2001 г.) это методология интеллектуального анализа данных, используемая в Witness Miner toolkit.

Как показано на рисунке 20, KDD Roadmap является итеративной методологией и состоит из восьми шагов:

1. формулировка проблемы;
2. выделение ресурсов;
3. очистка данных;
4. предварительная обработка данных;
5. Data Mining;
6. оценка;
7. интерпретация;
8. использование.

Основной вклад KDD Roadmap – это включение в рассмотрение задачи ресурсообеспечения (resourcing) для повышения эффективности процесса обнаружения знаний.

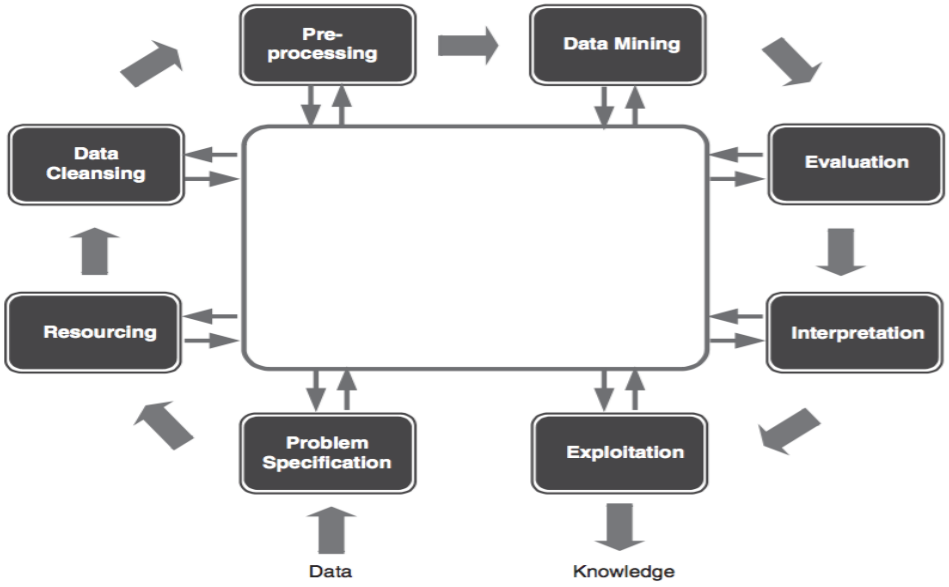

Рис. 20. Этапы методологии KDD Roadmap

5.9. Сравнение методологий

В таблице 2 на основе работы [16] представлено сравнение этапов рассмотренных методологий KDD&DM относительно наиболее распространенной методологии CRISP-DM.

Стандарты процессов KDD&DM развиваются, появляются также новые, дополняющие современные стандарты. Это свидетельствует о достаточной «зрелости» методологий KDD&DM, позволяющей применять их для проектирования систем интеллектуального анализа данных.

5.10. Контрольные вопросы

1. В чем заключается методология CRISP-DM?
2. В чем заключается методология RAMSYS?
3. В чем заключается методология SPSS?
4. В чем заключается методология KDD Roadmap?
5. В чем отличие вышеперечисленных методологий? Дайте свою оценку каждому методу. Сделайте вывод о том, какая методология для каких предметных областей будет лучшей.

Таблица 2

Сравнение методологий KDD&DM

CRISP-DM / RAMSYS	KDD-Outlined	KDD-Detailed	Human-Cent Approach	SEMMA	5A's	6-sigma	Cabena et al.	Two Crows	Anand & Buchner	Cios et al.	KDD Roadmap	DMIE	Marbán et al.
6	5	9	6	5	5	5	5	7	8	6	8	5	6
													Life Cycle Selection Processes
Business Understanding		Learning the Application Domain	Task Discovery		Assess	Define	Select	Define Business Problem	Domain Knowledge Elicitation	Understanding the Problem Domain	Resourcing	Analyse the Organization	Project Management Processes
									Human resource Identification				
			Data Discovery						Problem Specification		Problem Specification	Structure the Work	
Data Understanding	Selection	Creating a Target Data Set		Sample		Measure	Pre-process	Build DM Data Base	Data Prospecting	Understanding the Data		Develop Data Model	Pre-Development Processes
			Data Cleaning	Explore				Explore Data	Methodology Identification		Data Cleansing		
Data Preparation	Pre-processing	Data Cleaning and Pre-processing			Access					Preparation of the Data			Development Processes
	Transformation	Data Reduction and Projection	Model Development	Modify			Transform	Prepare Data for Modeling	Data Pre-processing		Pre-processing		
Modeling	Data Mining	Choosing the Function of DM		Model	Analyse	Analyse	Mine	Buid Model	Pattern Discovery	Build model	Data Mining	Implement Model	
		Choosing the DM Algorithm											
		Data Mining				Improve							
Evaluation	Interpretation / Evaluation	Interpretation	Data Analysis	Assess	Act	Control	Analyse and Assimilate	Evaluate Model	Knowledge Post-processing	Evaluation of the Discovered Knowledge	Evaluation		Integral Processes
											Interpretation		
Deployment		Using Discovered Knowledge	Output Generation					Deploy Model and Results		Using the Discovered Knowledge	Exploitation		Post-development Processes
					Automate								
												Establish On-going Support	Post-development Processes

ЗАКЛЮЧЕНИЕ

Проектирование и разработка систем интеллектуального анализа данных (Knowledge Discovery in Databases&Data Mining) является актуальным направлением вследствие необходимости решать задачи извлечения полезной информации из больших данных, накопленных в базах данных и хранилищах.

В настоящем пособии рассмотрены некоторые вопросы концептуального проектирования, которые необходимо решать при создании систем интеллектуального анализа данных. Впервые дана связь задач автоматизации проектирования с задачами интеллектуального анализа данных и приведены с единых методологических позиций формальные постановки задач процесса Data Mining, который составляет ядро интеллектуального анализа данных. Новым также является рассмотрение и сравнение стандартизованных методологий процесса разработки систем интеллектуального анализа данных (Knowledge Discovery in Databases&Data Mining). Авторы надеются, что содержание пособия, в особенности множество примеров, позволят исследователям и специалистам повысить эффективность при разработке автоматизированных систем.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Aggarwal, C. C. Data Mining: The Textbook, Springer International Publishing Switzerland, 2015.
2. Gugisch, R. Many-valued Context Analysis using Descriptions. In ICCS 2001, LNAI 2120.
3. Ganter, B., Wille, R. Formal Concept Analysis. In Mathematical Foundations. Springer Verlag, Berlin, 1999.
4. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллекту альному анализу данных [http://www.machinelearning.ru/].
5. Buchli, F. Detecting Software Patterns using Formal Concept Analysis, 2003.
6. Yager, R. R. A new approach to the summarization of data // Information Sciences, vol. 28, 1982.
7. Han, J., Kamber, M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2000.
8. Ярушкина, Н. Г. Интеллектуальный анализ временных рядов : учебное пособие / Н. Г. Ярушкина, Т. В. Афанасьева, И. Г. Перфильева. – Ульяновск : УлГТУ, 2010.
9. Афанасьева, Т. В. Internet-сервис экспресс-анализа экономического состояния предприятия / Т. В. Афанасьева и др. // Двенадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2010 (20 сентября – 24 сентября, 2010 г., Тверь, Россия) : труды конференции. – В 4-х томах. – Том 4.
10. Машечкин, И. В. Методы интеллектуального анализа и некоторые их приложения / И. В. Машечкин, М. И. Петровский. Доступно по адресу [citforum.ru/seminars/cbd2009/2_7.ppt].

11. Афанасьева, Т. В. Применение кластеризации в электронном образовании / Т. В. Афанасьева, Н. Д. Жучков, А. Н. Солдаткин // Электронное обучение в непрерывном образовании : III Международная научно-практическая конференция : сборник научных трудов. – Ульяновск : УлГТУ, 2016.
12. Прохоров, Е. Э. Сегментация игроков в компьютерные игры по типу поведения // Прикладные информационные системы : Третья Всероссийская НПК : сборник научных трудов / Е. Э. Прохоров ; под ред. Е. Н. Эгова. – Ульяновск : УлГТУ, 2016.
13. Афанасьева, Т. В. Сервис прогнозирования на основе комбинирования моделей нечетких временных рядов и ARIMA / Т. В. Афанасьева и др. // Пятнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2016 : труды конференции. В 3-х томах. Т. 3 : Смоленск: Универсум, 2016.
14. Афанасьева, Т. В. Сравнение результатов регрессионной модели и модели нейронной сети при оценке стоимости нового апартамента / Т. В. Афанасьева, К. П. Золотова, Е. А. Савенкова, М. М. Фирулина // Прикладные информационные системы : Четвертая Всероссийская НПК : сборник научных трудов. – Ульяновск : УлГТУ, 2017.
15. Joaquim L. Viegas, Susana M. Vieira, Joao M.C. Sousa. Fuzzy clustering and prediction of electricity demand based on household characteristics // International Joint Conferece IFSA-EUSFLAT (16th World Congress of the International Fuzzy Systems Association (IFSA), 9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)), Gijon (Asturias) Spain, 2015.
16. G. MARISCAL, O´. MARBA´ N AND C. FERNA´ NDEZ . A Survey of KD & DM process models and methodologies // The Knowledge Engineering Review, Vol. 25:2, & Cambridge University Press, 2010.

16. G. MARISCAL, O´. MARBA´ N AND C. FERNA´ NDEZ . A Survey of KD & DM process models and methodologies // The Knowledge Engineering Review, Vol. 25:2, & Cambridge University Press, 2010.
17. Солонин, Е. Б. Интеллектуальные технологии поиска и анализа данных / Е.Б. Солонин // Электронное текстовое издание, УрФУ, Екатеринбург, 2015.
Доступно по адресу
[http://www.study.urfu.ru/Aid/Publication/13334/1/Solonin.pdf]
18. Интуит – Национальный открытый университет. –
Доступно по адресу
http://www.intuit.ru/EDI/14_02_16_4/1455402139-23616/tutorial/25/objects/21/files/21_3.gif.

Учебное электронное издание

АФАНАСЬЕВА Татьяна Васильевна
АФАНАСЬЕВ Александр Николаевич

ВВЕДЕНИЕ В ПРОЕКТИРОВАНИЕ СИСТЕМ
ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Учебное пособие

Редактор Ю. С. Лесняк

ЛР № 020640 от 22.10.97.

ЭИ 1043. Объем 2,6 Мб.

Печатное издание
Подписано в печать 21.08.2017. Формат 60×84/16.
Усл. печ. л. 3,72. Тираж 75 экз. Заказ 820.

Ульяновский государственный технический университет
432027, г. Ульяновск, Сев. Венец, 32.
ИПК «Венец» УлГТУ, 432027, г. Ульяновск, Сев. Венец, 32.
Тел.: (8422) 778-113
E-mail: venec@ulstu.ru
venec.ulstu.ru