

## ОГЛАВЛЕНИЕ

<b>ВВЕДЕНИЕ</b>	<b>3</b>
<b>1 АНАЛИЗ ДАННЫХ В ДЕЯТЕЛЬНОСТИ КОММЕРЧЕСКОГО ПРЕДПРИЯТИЯ</b>	<b>7</b>
§1.1 До века вычислительной техники	7
§1.3 Знания как актив	10
§1.4 Нечеткие множества	11
§1.5 Нейронные сети	13
§1.6 Генетические алгоритмы	16
§1.7 Data Mining	19
§1.8 Будущее: децентрализованные базы данных	26
<b>2 ДАННЫЕ В КОММЕРЧЕСКОЙ МЕДИЦИНЕ</b>	<b>29</b>
§2.1 Здоровоохранение как коммерческая деятельность	29
§2.2 Процессный подход в медицине	30
§2.3 Анализ данных и медицинская тайна	33
§2.4 Потенциал новых методов анализа и обработки информации	34
<b>3 ВЫДЕЛЕНИЕ ЗНАНИЙ В БАЗЕ ДАННЫХ КОММЕРЧЕСКОЙ КЛИНИКИ</b>	<b>38</b>
§3.1 Объект исследования	38
§3.2 Обзор решений и выбор инструментария для эксперимента	40
§3.3 Описание эксперимента	47
§3.4 Ценность полученного результата	53
<b>4 ЗАКЛЮЧЕНИЕ</b>	<b>55</b>
<b>5 СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ.</b>	<b>58</b>

## ВВЕДЕНИЕ

Умение учиться на ошибках, как своих, так и чужих, издревле отличало человека разумного от других биологических видов. Переосмысление своего, а еще лучше доступного чужого опыта дает возможность в своих решениях опереться на накопленные знания, избежать уже кем-то пройденных ошибок. В человеке самой природой заложены механизмы обучения, переосмысления (рефлексии), абстрактного мышления, которые подарили нам бесценную возможность осознать и предвидеть ход событий даже в принципиально новых ситуациях. Благодаря этим свойствам нашего сознания мы способны принимать решения и действовать при отсутствии релевантного опыта. Развитие этих возможностей привело человечество к постулированию научного метода познания, когда эмпирический опыт человека или группы стало возможно последовательно встраивать в совершенно новую картину мира, что ознаменовало эру царствования науки [1]. В этой картине диалектически уживаются стремление к предсказуемости, изученности окружающего мира и неведомая до того новизна вкупе с бурным ростом сложности во всех сферах человеческой деятельности. Наши современники в своем бытовом и профессиональном обиходе оперируют таким количеством объектов и понятий, которое век назад подвергал изучению практикующий ученый.

Бурный рост сложности коснулся экономики, как в хозяйственной деятельности непосредственно, так и в организационного управления. При том, что сложность вынуждает субъекты хозяйственной деятельности все больше специализироваться, дробиться, локализуя тематический участок приложения деловых усилий, взрывной рост разнообразия и количества хозяйственных операций, объема номенклатуры, методологий, процессов и инструментов внутри организации не останавливается. Сложность сама по себе требует учитывать и планировать все материальные и нематериальные объекты, значимые для ведения деятельности предприятия. Но помимо этого, в условиях конкурентного рынка, очень важным становится качество и скорость этого учета и обработки. Достаточно заметить, что сами данные, а также мощность и надежность систем обработки данных с 80х годов становятся ключевыми активами компаний, сравнимыми по важности с оборудованием, людьми, материалами и т.д.

Помимо собственно данных для планирования и прогнозирования еще более значимыми становятся зависимости, особенно определяющие закономерности между данными, имеющими принципиально различные области возникновения и применения. Теперь уже человек в силу особенностей мышления не имеет объективной возможности многие взаимосвязи даже выявить. Классический подход «наблюдение – гипотеза – эксперимент – подтверждение» не работает там, где взаимосвязанные факты имеют столь

различное происхождение, что даже не рождает гипотезу. В начале 21 столетия к этому добавляется тот факт, что скорости процессов, обилие изменений и большой объем генерируемых фактов и событий выдвигают принципиально новые требования к анализу накопленной информации.

Даже небольшое предприятие за короткий срок своей работы способно накопить значительные массивы данных. Как мы уже замечали, в диалектическом единстве пребывают нарастающая скорость изменений в окружающей компании среде и стремление к упорядочиванию событий, выстраиванию устойчивых правил их взаимосвязей. Маркетологи взялись за поиск таких правил в поведении потребителей, технологи – в технических процессах. Сталкиваются они с озвученной выше проблемой: будучи людьми, они в силах работать с только с теми зависимостями, которые можно заметить и выделить там, где что-то – опыт, интуиция, другая методология – подсказывает искать. Закон Парето [2] подсказал маркетологам способы анализа клиентской базы, а идея спросить потребителя привела их к фокус группам и иным методам обратной связи. Теория вероятности стала основой методологий, работающей с прогнозами и рисками. Вычислительная техника являлась лишь средством быстрых расчетов больших массивов данных. Однако вперед всегда вырываются те, кто способен уловить нечто неочевидное, совершенно новое.

В современном бизнесе с учетом глобальной открытости мира особенно актуальным становится поиск новизны, которую оценит потребитель и которая позволит отличаться от других. Источником таких знаний являются именно накопленные данные, то есть пережитый и зафиксированный компанией опыт. Примером такого опыта является набор сведений о потребителях, их поведении в условиях номенклатурного разнообразия продукции и услуг, о времени, месте и иных показателях, сопровождающих такое поведение. Современные информационные системы для разных задач учета и управления фиксируют значительный объём данных, в котором вышеупомянутые технологи и маркетологи выделяют понятные им подмножества сведений. Точками же взаимосвязей этих подмножеств в лучшем случае является идентифицирующий номенклатуру или потребителя набор. Маркетолог знает, кто какой товар купил, кладовщик – сколько его в запасах. Производственник или закупщик берет набор информации, который помогает определить снабжение. И «вдруг» выясняется, что теплым летом надо бы лучше снабжать потребителя помимо холодной воды какими-нибудь определенными сортами колбасы. Маркетологи потом расскажут, как выяснили про то, что ее любят есть на пикнике. А снабженцу бы это знать весной, у него были бы меньше издержки. Закономерность-то есть, но вот лето до того было дождливым всегда, однако в теплые выходные становится заметно. При анализе валовой потребности в снабжении оценивалась оборачиваемость товарных

позиций в срезе в неделях и месяцах, что не позволило зависимость спроса от сезона обнаружить.

Коммерческая медицина сталкивается в своей деятельности ровно с теми же проблемами, что и разобранный выше условный ритейл. Номенклатурное разнообразие, конкурентная среда, обилие околomedicalных услуг, большие издержки на реагентах и расходных материалах, капиталоемкие основные фонды – вполне стандартный набор с учетом отрасли. Подробнее задачи и проблемы медицинского центра мы рассмотрим в главе 2. Сейчас же обратим внимание на то, что показатели эффективности медицинского учреждения включают такие метрики, как точность первичной диагностики, сокращение периода восстановления пациентов и другие факторы, имеющие, помимо экономического смысла, высокую социальную значимость, поскольку имеют непосредственное отношение к здоровью человека.

Выделение знаний в медицинской деятельности в связи с этим имеет большое значение. Эти знания способны помочь лучше лечить людей, и, что еще более важно, точнее и раньше диагностировать заболевания по косвенным, слабоуловимым признакам.

### **Цель работы.**

В качестве целевого состояния, к которому мы намереваемся прийти в ходе выполнения настоящей работы, положим следующее: *«Используя электронную базу данных коммерческого медицинского центра, выделить слабые, неочевидные закономерности, представляющие ценность для бизнеса».*

Чтобы цель больше соответствовала модели S.M.A.R.T [3, с. 35-36], необходимо формализовать два нечетких положения:

1. Неочевидными мы будем считать такие закономерности, поддержка которых составит от 0.5% до 5%. Далее, при анализе методов, а также в описании эксперимента понятие поддержки буде раскрыто.

2. Представляющие ценность – такие закономерности, интерпретация и использование которых бизнесу нарастить положительные объективные показатели, например, конверсию рекламных компаний, средний чек и т.д.

Обозначенная цель повлекла постановку следующих задач:

- Выбор методологии анализа данных.
- Подбор инструментария.
- Выбор среза данных для анализа.
- Проведение экспериментального анализа на данных, предоставление результата эксперимента бизнесу.

Вышеуказанные задачи последовательно решались в ходе работы с применением всего опыта, приобретенного автором за время обучения на курсе МВА.

### **Объект изучения (эксперимента).**

На момент написания настоящей работы автор работал в должности технического директора в организации, разрабатывающей программное обеспечение, автоматизирующее деятельность медицинской организации – медицинскую информационную систему (далее МИС). В рамках своей деятельности удалось наблюдать работу нескольких медицинских центров разного масштаба. В одном из них предложение продемонстрировать тот факт, что в данных методами анализа можно выделить неочевидные закономерности, встретило понимание. Сеть клиник предоставила доступ к своей базе на условиях обезличенности пациентов. Для эксперимента были выбраны именно пациенты для их профилирования по признакам. Подробнее эксперимент изложен в главе 3.

# 1 АНАЛИЗ ДАННЫХ В ДЕЯТЕЛЬНОСТИ КОММЕРЧЕСКОГО ПРЕДПРИЯТИЯ

## §1.1 До века вычислительной техники

Как мы уже указывали в введении, осмысление опыта является инструментом для выработки решений человеком с древних времен. Человек как биологический вид в ходе эволюции выработал такой инструмент как внегеномное наследование: дети впитывают социальные навыки окружения также жёстко, как и геномные признаки (Савельев, 2010). С развитием цивилизации, усложнялись социальные и хозяйственные отношения в обществе. Используя вышеуказанный инструмент внегеномного наследования, ранние общества выработали кодексы традиционных правил и табу, позволяющие социуму переносить опыт через поколения.

Одновременно с ростом сложности человеческой деятельности возникала необходимость в накоплении и передачи сведений, имеющих некоторые новые свойства - знаний. Тут и далее мы будем придерживаться модели DIKW (David Weinberger, 2010.), отображение которой приведено на Рисунке 1.



Рисунок 1. Иерархическая модель DIKW.

Идея модели состоит в том, что на каждом новом уровне сведения приобретают некоторое новое качество, увеличивают ценность, и, обычно, теряют в объеме. К сожалению, модель имеет семантические проблемы с однозначностью перевода на русский язык. Нам же важно именно то, что знания – это такие сведения, обладание и применение которых принципиально отличает деятельность с ними от таковой без них. Скажем, наличие именно знаний, необходимых для оказания первой помощи человеку, может являться определяющим фактором для чьей-то жизни.

Для передачи знаний социум придумал образование. Получение образования в человеческом обществе стало необходимым условием даже минимальной социализации. Знания стали упорядочивать, систематизировать, они приобрели осязаемую материальную и не только материальную ценность. Благодаря структуре головного мозга человек оказался способен учиться – то есть знания воспринимать и анализировать (Савельев, 2014). В экономических отношениях, социальном доминировании, военном деле, научных прорывах особенно заметным преимуществом очевидно стали обладать индивиды, способные быстрее и полнее других переосмыслить массив чужого опыта, а после переосмысления выработать стратегию поведения с высокой степенью новизны, что в нашей модели DIKW уже можно отнести к мудрости. В любом случае, выделение закономерностей шло от некоторой гипотезы, выработанной системно или интуитивно. Принято считать, что на высокую новизну гипотез способны люди с развитым шизоидным радикалом (Пономаренко, 2006), мозг которых строит нетиповые, нестандартные образы. Мы вернемся к модели и DIKW и пониманию знаний чуть позже. Пока же обратимся к тому, последовательно учились выделению системных знаний при помощи информационных технологий.

Обратим внимание на следующее историческое наблюдение. Хозяйственная деятельность человека со времени формирования капитализма как общественно-экономической формации (вторая половина XIX века) характеризуется скачкообразным ростом объема обрабатываемых данных, их сложности, числа взаимосвязей. Особенностью этой области человеческой деятельности явилось то, что данные, рождаемые в ней, подвергались и подвергаются высокой степени формализации, что и привело к появлению основ финансового учета [7]. Кстати, в размытое многозначное понятие «управленческий учет» как раз входят методы фиксации и анализа хозяйственных операций. Оборотно-сальдовые ведомости, карточки счета и субсчета в бухгалтерском учете, отчеты по товарам, покупателям и т.д. - все это стало попыткой представить информацию в виде, «удобном» для анализа и принятия решений человеком. Дополнительно человечество придумывало методы и инструменты, помогающие ускорить устный и письменный счет и обработку – арифмометры, запись в столбик и т.д. Так было до начала эры информационных технологий.

## **§1.2 Интеллект и почему он искусственный**

Собственно, все развитие средств вычислительной техники поначалу воспринималось как быстрый и безошибочный арифмометр, способный помочь разуму

человека решать задачи, но не помогать добывать знания и, уж тем более, мудрость. Так было вплоть до 1950 года, до статьи Алана Тьюринга в журнале *Mind* [8], имевшем статус философского издания. В статье Тьюринг рассуждает о том, может ли мыслить машина, и как в слепом эксперименте машину от человека отличить. С этого момента начинается эпоха искусственного интеллекта.

Здесь мы будем придерживаться определения интеллекта как именно биологической способности человека вырабатывать решения на основе всей совокупности опыта, ощущений, инстинктов, с привнесением в решение черт характера. Таким образом можно утверждать, что интеллект – это продукт человеческого мышления, созданный на основе удивительного биологического механизма (Савельев, 2014), и так до конца не скопированный в техническом воплощении.

Так на стыке достижений сначала ламповой, а потом уже полупроводниковой электроники и сформулированного Тьюрингом научного вызова стали появляться машины, способные решать некоторые задачи не хуже человека. Задачи эти в большинстве своем относили к возможным только для человеческого интеллекта: игра в шахматы, распознавание речи, зрительных образов и т.д. Однако на деле выходило так, что конкретная машина могла решать только узкий класс задач, под которую она и создавалась. При этом интеллектуальная мощность машины была фиксирована и закладывалась на этапе ее создания. Полноценный интеллект, способный себя обучать и развиваться, так и не создали, в связи с чем к термину к 60-м годам заметно охладели.

Далее, к 70-м годам появляется термин «экспертная система», немного продливший волну интереса к реализации неутомимого машинного интеллекта. Ответить ожиданиям так и не получилось, все разработанные с таким названием информационные системы оказались классическими алгоритмическими машинами, реализующими пусть сложные и разветвленные, но все же конечные наборы инструкций. Вычислительная мощь аппаратных средств, а также методология автоматизации тем временем росла и крепла, предоставив бизнесу возможность накапливать и обрабатывать, пусть и инструкциями, невиданные доселе объемы информации.

В 2000-ые годы мечта вернулась. В статье *The New York Times* 20.07.2006 «Brainy Robots Start Stepping Into Daily Life» эксперт IBM предрек достижения новых горизонтов, даже представить которые мы не в силах [9]. Информационные системы становятся действительно самообучаемыми, приходит термин *Artificial Intelligence*. Умение обучаться в ходе выполнения своих задач и анализируя полученный результат – вот чему мы научили машину. Однако это еще не было полноценным переосмыслением полученного опыта.



### §1.3 Знания как актив

Рассмотрев модель DIKW из параграфа 1.1, можно заметить, что ценность знаний заметно выше, чем у информации и данных. При этом взаимосвязь информация-знания можно формализовать. Определяется она как качественный переход, уменьшающий неопределенность модели. Здесь оговорим, что неопределённость есть неотъемлемое качество любой модели и убрать ее окончательно не представляется возможным [10, с. 139]. Также обратим внимание, что современные компании признают знания своим основным активами [11]. Хорошо бы свои активы сберечь, а еще лучше приумножать, если они полезны.

Сдерживающим фактором развития технологии в настоящий момент является не отсутствие технических и организационных возможностей, а скорее то, что многим организациям на молодых рынках сложно представить себя компанией знаний – гораздо понятнее сырьевые или индустриальные активы, недвижимость и т.д. К тому же решение обратиться к своему опыту для поиска принципиально новых знаний психологически непросто дается рядовому человеку, что уж говорить об организациях как сложных системах. Но ситуация неослабевающего роста конкуренции во всех областях подвигает компании на поиск новых возможностей, в том числе и в переосмыслении своего опыта.

Для выделения знаний используются методы анализа, классифицируемые следующим образом [13, с. 18]:

- Классическая математика (мир объектов и их формальных моделей);
- Моделирование (теория нечетких множеств);
- Обучение (нейронные сети, генетические алгоритмы);
- Умный перебор (углубленный анализ данных – Data Mining).

Осветим каждый из методов чуть подробнее.

Строго говоря, все вышеуказанные методы извлечения знаний вышли из математики и описываются лучше всего в математических моделях. Под классической математикой для точности понимается аналитическая модель описания процесса. Здесь уместнее всего вспомнить физику школьного курса. Формула механического движения как раз и есть такая модель. Для процесса выбирается способ описания зависимости выходных параметров от входных (формула, таблица, график) и строится зависимость в интересующей нас области. Тогда, для нашего примера, мы можем обладать знанием о том, где будет наше материальное тело в разные моменты времени. Метод имеет отличное применение в сферах высокой детерминированности данных и процессов при наличии изученного соответствующего математического аппарата. Недостатком метода можно назвать необходимость наличия детерминированной гипотезы. То есть аналитике надо

подвергать какое-то предположение, чтобы выстроить решение. Дополнительно выяснилось, что строгими аналитическими моделями очень громоздко описывать логику решений человека, оперирующего нечеткими понятиями. Да и при высокой детерминированности математический анализ требует решения сложных систем дифференциальных уравнений, что неудобно, а подчас невозможно для процессов, быстро протекающих в реальном времени, когда необходимо выработать оперативно корректирующее воздействие.

### §1.4 Нечеткие множества

Нечеткие множества (Fuzzy Logic) предложил миру профессор Лотфи Заде в 1965 году [14 с. 338-353.]. В наше время их применение стало стандартным в целом ряде областей, включая бизнес. Проблему, которую решила теория, можно озвучить как «неформализуемые очевидные понятия». Скажем, для принятия решения о том, что, выходя на улицу, стоит одеться, нам достаточно ощущения «холодно». Не точное количество градусов Цельсия, не посчитанная более сложно формула восприятия погоды человеком. Строго говоря, даже не «холодно», а «холоднее, чем для майки». Аналитическая математика не в силах такое зафиксировать. Равно как и не способна зафиксировать качественные отличия типа «теплее», «холоднее» без меры и количества. Если не может формальная математика, то не может и машина, которая может оперировать в бинарном коде 0\1 (выполнено строгое условие или нет). Гениальность Заде состояла в том, чтобы вместо бинарной логики формализовать непрерывное пороговое состояние внутри дискретной. Так, отображение из множества  $X$ , элементы которого принимают значение 0 или 1 в такое же множество  $Y$  дает нам классическую двоичную логику, отлично моделирующую процессы в состояниях «да \ нет», «истина \ ложь» и т.д (см 15):.

$$f(X) \in \{0; 1\} \rightarrow Y \in \{0; 1\} \quad (1)$$

Здесь  $f(X)$  — двоичная (логическая) функция булевой переменной. На булевой логике стоит теория двоичных вычислений, отлично моделирующая работу вычислительных устройств вплоть до нашего времени, и пока необходимости в смене парадигмы построения вычислительных узлов не наблюдается (Набебин. 1996).

Заде предложил следующее:

Под нечётким множеством  $A$  понимается совокупность упорядоченных пар, составленных из элементов  $x$  универсального множества  $X$  и соответствующих степеней (функций) принадлежности  $\mu_A(X)$ . Прелесть здесь в том, что множество  $X$  — множество

значений от 0 до 1:  $X \in [0; 1]$  - не имеет четких значений, при этом  $\mu_A(X) \in [0; 1]$ . А то, 0 или 1 принимает значение в зависимости от некоторого значения (нормы)  $a$ :

$$\mu_A(X) = \begin{cases} 0, & \text{для } X < a; \\ 1, & \text{для } X > a. \end{cases} \quad (2)$$

Выражение (2) позволяет изящно разработать строгую модель для выражений типа «высокая цена». Можно понимать, что выше какого-то значения (нормы) цена за некоторый товар высока, хотя «справедливую» цену покупатель не в силах сформулировать. Так у улыбающейся продавщицы и купил бы выше, а у этой и даром не надо.

Продолжением теории является теорема аппроксимации нечетких множеств (Кофман, 1982), интерпретация которой позволяет утверждать, *что любая задача управления сводится к конечному числу «Если..., то», оперирующих с нечеткими условиями при заданных нормах*. Это качество нечетких множеств позволило свести задачи выработки решений в автоматизированных системах управления к гораздо более оптимальным алгоритмам, убрав необходимость в интегрировании сложных систем сложных дифференциальных уравнений с параметрами. Например, анализ конечного числа датчиков параметров плавления металла и обработка снятых значений средствами нечеткой логики заметно проще для автоматизированных вычислений, нежели решение сложной системы термодинамических уравнений при высокой скорости смены входных параметров.

В наше время нечеткие множества стали фактически стандартом моделирования человеческих решений в неформализуемых никак, кроме как некоторой интуитивной нормой ситуациях.

Систему, построенную на нечетких множествах можно улучшать, корректирую показатель нормы как управляемый параметр. Таким образом, можно утверждать, что система, построенная на приложении модели нечетких множеств, способная обучаться, пусть пока и не сама. Это оказалось очень удобно в управлении производственными процессами, при распознавании и принятия решений в системах, получающих данные от видео или аудионаблюдения, и даже при разработке систем, работающих с выделением строгих критериев в текстовых поисковых запросах («недорогая машина», «теплое пальто»).

## §1.5 Нейронные сети

Судьба теории нейронных сетей (или нейросетей) похожа на лихо закрученный сюжет средневекового романа. Началось все с создания модели работы головного мозга. Мак-Коллок (McCulloch) и Питс (Pitts) [16] в 1943 разрабатывают формальную модель на основе строгой логики (см. формула 1), которую дальше сильно развил Розенблат [17]. Понимание нейрофизиологии вышеуказанными учеными привело их к математическому моделированию работы нейрона – элемента центральной нервной системы, который на основании множества входных воздействий от иных нервных узлов (органы чувств или другие нейроны) через элементы связи (синапсы) переходит или не переходит в возбужденное состояние, в соответствии с которым передает или не передает сигнал на выход – аксон (см. Рисунок 2).

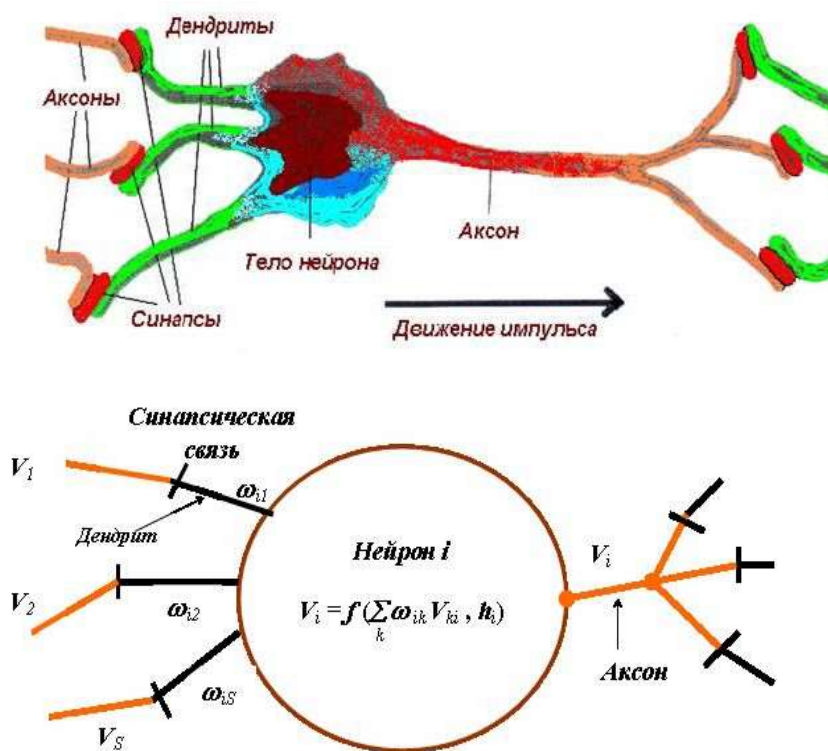


Рисунок 2. Изображение нейрона и его модель.

На рисунке 2 приведена еще одна сущность, описанная нейрофизиологами – дендрит. Дендрит представляет собой входной отросток нейрона и отделен от синапса, который представляет собой канал проводимости. Из рисунка видно, что математически возбуждение нейрона  $V_i$  на аксоне смоделирована функцией алгебраической суммы входных значений (синапсов  $V_{ik}$ ), нормированных дендритами  $\omega_{ik}$  порога срабатывания нейрона  $h_i$ . Для простоты полагают следующее:

$$V_i = \begin{cases} 1, & \text{если } \sum_{k=1}^S \omega_{ik} V_{ki} \geq h_i \\ 0, & \text{в ином случае} \end{cases} \quad (3)$$

Имея в виду при этом, что значения  $\omega_{ik}$  и  $h_i$  можно регулировать, изменяя уровень срабатывания нейрона. Если принять, что синапсы - это некоторые факты, а аксон - реакция на них, то нейрон моделирует простейшую последовательность «факты» - «следствие» или «реакция». Далее из нейронов можно выстроить некоторую систему, замыкая комбинации аксонов с ветвлением на другие дендриты, причем некоторые выходы возвращая обратно в виде связи - сеть или нейронную сеть. Для удобства работу с сетью представляют тактируемой (стробируемой), когда сигналы в одном такте работают с группой нейронов – слоем. Первый слой – входной, крайний – выходной. Схематичное изображение нейронной сети с двумя слоями приведено на Рисунке 3. Современные промышленные нейросети – многомерные, многослойные структуры с жесткой обратной связью, способные сразу обрабатывать несколько порций входных сигналов, разнесенных по слоям и тактам.

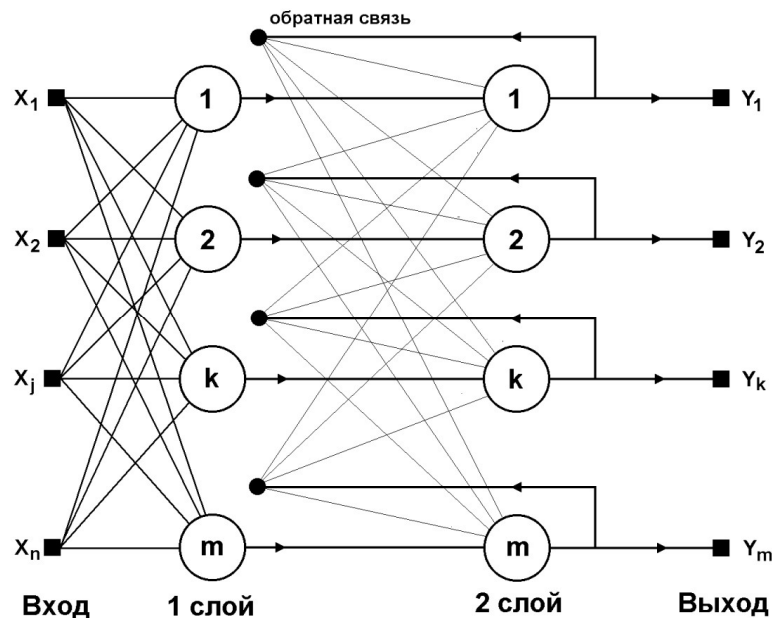


Рисунок 3. Двуслойная нейронная сеть.

Обратим внимание, что надежды на теорию возлагались серьезные, однако в 1969 году Минский и Пепперт показывают бесперспективность теории нейронных сетей [18], одновременно с этим трагически погибает Розенблатт. Научные и бизнес сообщества к теории нейронных сетей охладевают, несмотря продолжение исследований с большими усмехами в научных лабораториях. В 90 годы происходит возвращение внимания к нейросетям на волне возврата интереса военных (а после успеха Интернет к ним прислушиваются), технологическому скачку персональных компьютеров и теплomu отношению СМИ в связи с нейрофизиологическими истоками. Здесь «вспоминаются» успехи научных лабораторий в разработке нейросетей по распознаванию зрительных образов, рукописных текстов, а также главное качество нейросетей – самообучение. Если через нейросеть «прогонять» некоторую задачу, при которой возможно сравнение выхода нейросети с эталоном, то можно выработать некоторое корректирующее воздействие на вход, для следующего прогона. Если эту обратную связь сразу реализовать в модели, то получает возможность обучаться, раз за разом все лучше приближаясь к эталону. Опять же, нет необходимости прорешивать сложные системы уравнений, описывающих процессы. Дополнительными преимуществами нейросетей является устойчивость к частичному повреждению (функции деградируют в скорости и точности по сравнению с эталоном, но выполняются), параллелизм при решении многих задач и самоорганизация. Имея возможность выделять некоторые тренды и паттерны поведения систем, нейронные сети широко используются в бизнесе для:

- прогнозирования объемов продаж;
- управления производством;
- исследования рынка и поведения клиентов;
- риск-менеджменте и др.

Математика дает нам формальное доказательство справедливости использования нейросетей в различных процессах благодаря Теореме Колмогорова [19], из которой следует, что любую функцию (модель) от  $d$  переменных можно представить функциями одной переменной количеством  $d$  и операциями взвешенного суммирования (нейросеть – см. Рисунок 2 и формулу 3) количеством  $2d+1$ , причем количество функций  $d$  и операций суммирования  $2d+1$  необходимо и достаточно. Это означает, что нельзя обойтись меньшим числом функций и операций суммирования, но нет необходимости и в большем их количестве.

### §1.6 Генетические алгоритмы

Механизм поиска решений некоторых задач человеком можно определить, как предпочтительное стратегия «лучше что-то попробовать, чем ничего не делать». Такой подход удобен в условиях неопределенности вкупе с неотвратимой необходимостью решение принимать. Упрощенно можно утверждать, что однозначно определен только один факт, что некоторое текущее состояние системы точно губительно, но вот что точно делать – неясно. В природе такая задача, связанная с выживаемостью видов, реализуется алгоритмом наследования или генетическим алгоритмом. Подробная модель приведена ниже, пока же сразу оговорим недостатки генетических алгоритмов. Она состоит в том, что в живой природе в условиях борьбы за выживание удачным считается решение, которое достигает поставленной задачи, речи об оптимизации не идет. Достаточно посмотреть на формирование биологических видов. Скажем, такой вид, как современный человек – венец природы, – представляет собой удивительное сочетание сложности формирования в ходе эволюции с некоторой бессильной ограниченностью решений. Нейрофизиолог Савельев (Савельев, 2014) сетует, что при неповторимости головного мозга до степени невозможности построения его полной модели принцип его развития и формирования в эволюцию убог и непоследователен. Если бы при «выборе» биологических механизмов для реализации каких-то решений эволюции присутствовал современный образованный биолог, да еще мог бы на это выбор влиять, то мозг бы строился на гораздо более оптимальной биологической архитектуре. А из-за того, как сложилось, мы имеем набор присущих человеку слабостей, ограниченности энергетического ресурса головного мозга, старческий маразм и прочее. Однако человеческий вид выжил, да еще и совершил невероятный эволюционный скачок, что высшую биологическую задачу вида решает полностью. Природная реализация генетического алгоритма – биологическая эволюция (20), – зиждется на передаче устойчивых генетических признаков следующим поколениям. Выживают в условиях агрессивной и изменчивой окружающей среды те особи, чей набор признаков оказался лучше приспособлен к условиям, то есть реализовались некоторые «удачные» решения. При этом используется всего два принципа формирования генетических признаков у особи:

1. Скрещивание – из двух признаков родителей выбирается один. Обычно один из признаков может выбираться с более высокой вероятностью – доминантный, но может проявиться и второй – рецессивный согласно закону Менделя (Гайсинович, 1988). Так происходит закрепление «хороших» признаков, позволивших самим родителям дожить до появления потомства.

2. Мутация – скачкообразное «случайное» изменения генетически определяемого признака, несвязанное с признаками у родителей (там же). Некоторый природный эксперимент. Если мутация оказывается «удачной», то носители признака через механизм скрещивания закрепляют его у своего потомства. Заметим, что человек благодаря медицине и иным достижениям цивилизации, резко поменявшим требования окружающей среды к выживаемости и возможности воспроизведения потомства, изменил условия естественного отбора, но эволюцию человека как вида, не прекратил.

Генетические признаки биологических видов на нашей планете, определяющие от рождения внешний вид, специфику иммунитета, кодируются белковыми наборами в составе дезоксирибонуклеиновой кислоты (далее ДНК). Полная копия ДНК хранится в каждой клетке организма. Она представляет собой закрученную спираль из двух белковых наборов – см. рисунок 4. Белковые наборы соединяются друг с другом комплементарными парами белков. В наборах пар закодированы признаки, на основании которых строятся биологические ткани особи. При скрещивании на месте комплементарной пары берется пара одного из родителей, при мутации возникает другая пара, никак не связанная с такой же в ДНК родителей. Количество белковых пар, формирующих даже один простой признак, столь велико, что в живой природе набор признаков любой особи формируется на основании сочетания генов, полученных в ходе скрещивания и мутации. Если набор полученных признаков заведомо нежизнеспособен, а в случае человека – не позволяет вести полноценную жизни (синдром Дауна, гидроцефалия и т.д.), генетики говорят о врожденных уродствах, а в случае работы алгоритма – заведомо «плохие» решения.

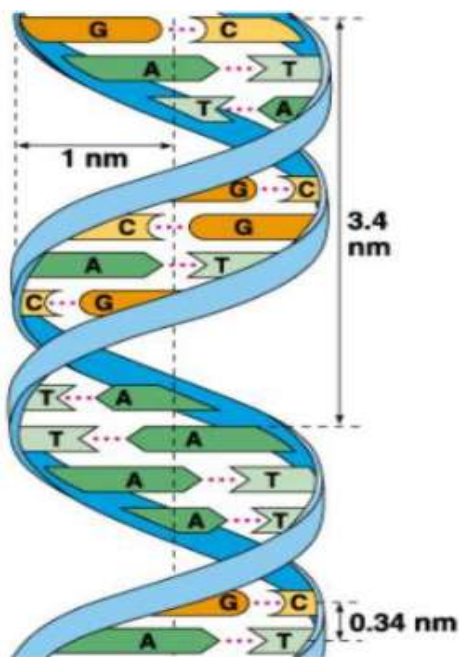


Рисунок 4. Изображение ДНК.



Математическую модель работы генетических алгоритмов можно описать на алгебре (множестве элементов и конечных операций над ними):

$\langle P_0, \lambda, l, s, \rho, f, t \rangle$ , Где:

$P_0 = (a_1^0, a_{12}^0, \dots, a_\lambda^0)$  - исходная популяция (родители),

$a_i^0$  для  $(1 \leq i \leq \lambda)$  - хромосома, ответственная за некоторый признак;

$\lambda$  - целое число (размер популяции);

$l$  - целое число (длина каждой хромосомы или число генов);

$s$  - оператор отбора;

$\rho$  - отображение, определяющее рекомбинацию (скрещивание и мутацию);

$f$  - функция оптимальности;

$t$  - счетчик поколений (критерий останова алгоритма).

Переложение разработанного природой механизма на поиск решений математической функции состоит в следующем. Выбирается некоторое решение (возможно – случайным образом), представляющее собой двумерную двоичную матрицу. Для простоты положим, что столбец матрицы – это некоторый параметр анализируемой функции (признак), а строка – это значение параметра в двоичном коде. Тогда элемент матрицы моделирует ген – двоичное число. Для начала работы выбирается (или случайным образом генерируется) матрица  $P_0$  - родитель. Далее хромосомный набор проходит стадию изменения через мутацию (случайное изменение элементов матрицы), а также через скрещивание, когда новое значение гена формируется на основании признака родителей. Каждое изменение определяется рекомбинацией  $\rho$ . Полученное решение (потомок) проверяется на «жизнеспособность» - насколько улучшилась функция оптимальности  $f$  по сравнению с имеющимся «лучшим» решением. Процесс итеративно повторяется фиксированное число раз либо по-другому, заведомо достижимому критерию останова работы алгоритма.

Недостатком эволюционного алгоритма является «слепой» поиск решений. Мы можем понимать на каком-то этапе, что какое-то поколение лучше остальных, но нет ни эталона для сравнения, ни аналитической возможности выделить.

Обоснованием применимости генетических алгоритмов является следствие из Теоремы Холланда (она же Теорема схем или Теорема шаблонов) [22], которое можно сформулировать следующим образом: *при реализации генетического алгоритма по множеству признаков среднее здоровье популяции (количество «хороших» решений) не уменьшается.*

Генетические алгоритмы широко используются в поисковой оптимизации, когда решение ищется в условиях высокой неопределенности. Опыт их применения показывает,

что при всей «случайности» метода существует ненулевая вероятность нахождения очень «хорошего» решения. Вероятность стараются повысить путем увеличения числа итераций рекомбинации (т.е. числа поколений), а также управляя такими параметрами, как вероятность мутации (сколько двоичных значений меняется за итерацию) и вероятность скрещивания (какие двоичные значения выбираются из родительских при выборе и на каком этапе). Генетический алгоритм не является самообучаемым, но показывает хорошие результаты при комбинировании с нейросетями, при помощи которых возможно корректировать параметры.

### §1.7 Data Mining

В заголовок параграфа вынесен англоязычный термин в связи с отсутствием удачного однозначного перевода на русский язык. Больше того, как и с термином Big Data, на волне спекуляций и поднятия его на знамя большим количеством организаций с сомнительной репутацией, термин был размыт и скомпрометирован в английском языке. Мы будем придерживаться практического определения, которое поддерживает, например, российская компания Base Group Labs [23], которая заслуживает отдельной теплой благодарности, о чем ниже. Итак, *Data Mining - это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из шагов Knowledge Discovery in Databases (KDD) – обнаружение знаний в накопленных данных.*

Основным отличием метода от описанных в параграфах 1.4-1.6 является то, что он использует целый набор алгоритмов обработки данных, причем число таких алгоритмов постоянно растет, да и сами алгоритмы развиваются и дополняются практиками. Конкретный пример применения этих методов может включать сложные последовательные и параллельные схемы обработки данных, вложенные внутрь одного алгоритма дополнительными обработками разными алгоритмами на разных участках и прочее. Как следует из вышеприведенного определения, основная ценность Data Mining проявляется в поиске нетривиальных знаний, то есть знаний, по существованию которых невозможно или сложно построить гипотезу. Широкий интерес к технологии, помимо причин, описанных в параграфе 1.3, обусловлен и ростом быстрого действия доступных широкому кругу компаний вычислительных средств. В главе 3 при выборе инструментария проведения эксперимента мы рассмотрим наиболее популярные решения.

В перечень математически описанных проблем, решаемых средствами Data Mining, входят:

1. *Классификация* – это отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

2. *Регрессия*, в том числе задачи прогнозирования. Установление зависимости непрерывных выходных от входных переменных.

3. *Кластеризация* – это группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность этих объектов. Объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем больше похожи объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация.

4. *Ассоциация* – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события X следует событие Y. Такие правила называются ассоциативными. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis).

5. *Последовательные шаблоны* – установление закономерностей между связанными во времени событиями, т.е. обнаружение зависимости, что если произойдет событие X, то спустя заданное время произойдет событие Y.

6. *Анализ отклонений* – выявление наиболее нехарактерных шаблонов.

Искусство Data Mining сводится к тому, чтобы бизнес-задачу или даже более абстрактную бизнес-цель свести к конечному последовательному набору решений вышеуказанных задач. Причем искусство анализа заключается именно в сведении входных данных к удобному для работы алгоритмов виде, а после – к подготовке данных для обработки. При этом совершенно «нелогичное», «неочевидное» применение алгоритмов по отношению к входным данным может дать очень полезные результаты. Рассмотрим используемый в эксперименте метод - ассоциативные правила. Строгая математическая модель метода состоит в следующем.

Пусть  $I = \{i_1, i_2, \dots, i_n\}$  – множество элементов, входящих в транзакцию, где транзакция суть набор событий, объединенных каким-то признаком, называемым идентификатором транзакции.  $D$  – множество транзакций.

Ассоциативным правилом называется импликация  $X \gg Y$  (читается "X дает Y" или "из X следует Y"), где  $X \in I$ ,  $Y \in I$  и  $X \cap Y = \emptyset$ .

Правило  $X \gg Y$  имеет *поддержку*  $s$  (support), если  $s\%$  транзакций из  $D$  содержат  $X$  и  $Y$ ,  $\text{supp}(X \gg Y) = \text{supp}(X \& Y)$ .

*Достоверность* правила показывает, какова вероятность того, что из  $X$  следует  $Y$ . Правило  $X \gg Y$  справедливо с достоверностью (confidence)  $c$ , если  $c\%$  транзакций из  $D$ , содержащих  $X$ , также содержат  $Y$ ,  $\text{conf}(X \gg Y) = \text{supp}(X \& Y) / \text{supp}(X)$ .

*Лифт* – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом:  $\text{lift}(X \& Y) = \text{conf}(X \& Y) / \text{supp}(Y)$ . Значения лифта, большие единицы, показывают, что условие появляется более часто в транзакциях, содержащих и следствие, чем в остальных.

Методика появилась и была развита при решении задачи анализа базы чеков в крупных розничных магазинах, где идентификатором транзакции являлся уникальный номер чека, а событиями в транзакции – наименования товара в чеке. Выделенные правила позволили выявить закономерности, которые не могли быть предсказаны маркетологами в силу отсутствия гипотез. При этом стимулирование покупателей к выдерживанию правил (размещение товаров, составляющих ассоциативное правило, рядом, совместные комплекты и т.д.) показывало заметный рост продаж. Тогда уже их «заметили» и даже как-то объяснили маркетологи. Изящность метода состоит в том, что подготовить в виде транзакций с идентификаторами можно практически любые данные. Например, у нас есть ряд слабосвязанных признаков клиентской базы: пол, возраст, регион проживания. Каждый из признаков можно представить конечным дискретным набором: пол мужской или женский, возраст разбить на несколько интервалов, регионы проживания тоже могут быть бесконечным рядом. Идентификатором транзакции у нас будет уникальный код пациента, а транзакциями – факт вхождения в какой-то из дискретных наборов. После мы ищем ассоциативные правила в наборе транзакций. Верный алгоритм не сможет выделить противоречивые правила в силу того, что транзакция из набора взаимоисключающих признаков всегда включает только один и делать дополнительную проверку нет необходимости, при этом правила выстроятся именно по другим сочетаниям. Выяснится, например, что у нас в клиентской базе, женщины тяготеют (но не входят очевидно) к одной возрастной группе.

Простым, хорошо изученным, и при этом быстро сходящимся алгоритмом выделения ассоциативных правил является алгоритм Apriori [24]. Лучше всего, как и другие алгоритмы дискретной математики, он иллюстрируется на примере.

Пусть мы имеем набор транзакций, представленный в Таблице 1. Идентификатор транзакции (номер) в данном случае не уникален, мы пишем в таблицу каждую транзакцию в новую строку. Во второй столбец впишем элемент, встречаемый в транзакции, в третий – количество (например – единиц товара).

Таблица 1. Обычный вид базы данных транзакций.

Номер транзакции	Наименование элемента	Количество
1001	A	2
1001	D	3
1001	E	1
1002	A	2
1002	F	1
1003	B	2
1003	A	2
1003	C	2
...	...	...

Для того, чтобы было возможно применить алгоритм, необходимо провести предобработку данных: во-первых, привести все данные к бинарному виду; во-вторых, изменить структуру данных. Тогда база транзакций примет вид согласно Таблице 2.

Таблица 2. Нормализованный вид базы транзакций.

TID	A	B	C	D	E	F	G	H	I	K	...
1001	1	0	0	1	1	0	0	0	0	0	...
1002	1	0	0	0	0	1	0	0	0	0	...
1003	1	1	1	0	0	0	0	0	1	0	...

Количество столбцов в таблице равно количеству элементов, присутствующих в множестве транзакций  $\{D\}$ . Каждая запись соответствует транзакции, где в соответствующем столбце стоит 1, если элемент присутствует в транзакции, и 0 в противном случае. Заметим, что исходный вид таблицы может быть отличным от приведенного в таблице 1. Главное, чтобы данные были преобразованы к нормализованному виду, иначе алгоритм не применим.

Итак, данные преобразованы, теперь можно приступить к описанию самого алгоритма. Алгоритм работает в два этапа. На первом шаге необходимо найти часто встречающиеся наборы элементов, а затем, на втором, извлечь из них правила. Количество элементов в наборе будем называть размером набора, а набор, состоящий из  $k$  элементов, –  $k$ -элементным набором.

Выявление часто встречающихся наборов элементов – операция, требующая много вычислительных ресурсов и, соответственно, времени. Примитивный подход к решению данной задачи – простой перебор всех возможных наборов элементов. Это потребует  $O(2^{|I|})$  операций, где  $|I|$  – количество элементов. Аргіогі использует одно из свойств поддержки, гласящее: поддержка любого набора элементов не может превышать минимальной поддержки любого из его подмножеств. Например, поддержка 3-элементного набора {Хлеб, Масло, Молоко} будет всегда меньше или равна поддержке 2-элементных наборов {Хлеб, Масло}, {Хлеб, Молоко}, {Масло, Молоко}. Дело в том, что любая транзакция, содержащая {Хлеб, Масло, Молоко}, также должна содержать {Хлеб, Масло}, {Хлеб, Молоко}, {Масло, Молоко}, причем обратное не верно.

Это свойство носит название анти-монотонности и служит для снижения размерности пространства поиска. Не имея мы в наличии такого свойства, нахождение многоэлементных наборов было бы практически невыполнимой задачей в связи с экспоненциальным ростом вычислений.

Свойству анти-монотонности можно дать и другую формулировку: с ростом размера набора элементов поддержка уменьшается, либо остается такой же. Из всего вышесказанного следует, что любой  $k$ -элементный набор будет часто встречающимся тогда и только тогда, когда все его  $(k-1)$ -элементные подмножества будут часто встречающимися.

Все возможные наборы элементов из  $I$  можно представить в виде решетки, начинающейся с пустого множества, затем на 1 уровне 1-элементные наборы, на 2-м – 2-элементные и т.д. На  $k$  уровне представлены  $k$ -элементные наборы, связанные со всеми своими  $(k-1)$ -элементными подмножествами.

Рассмотрим рисунок 5, иллюстрирующий набор элементов  $I = \{A, B, C, D\}$ . Предположим, что набор из элементов  $\{A, B\}$  имеет поддержку ниже заданного порога и, соответственно, не является часто встречающимся. Тогда, согласно свойству анти-монотонности, все его супермножества также не являются часто встречающимися и отбрасываются. Вся эта ветвь, начиная с  $\{A, B\}$ , выделена фоном. Алгоритм в этой ветке работать не будет. Использование этой эвристики позволяет существенно сократить пространство поиска, что сказывается на быстродействии алгоритма.

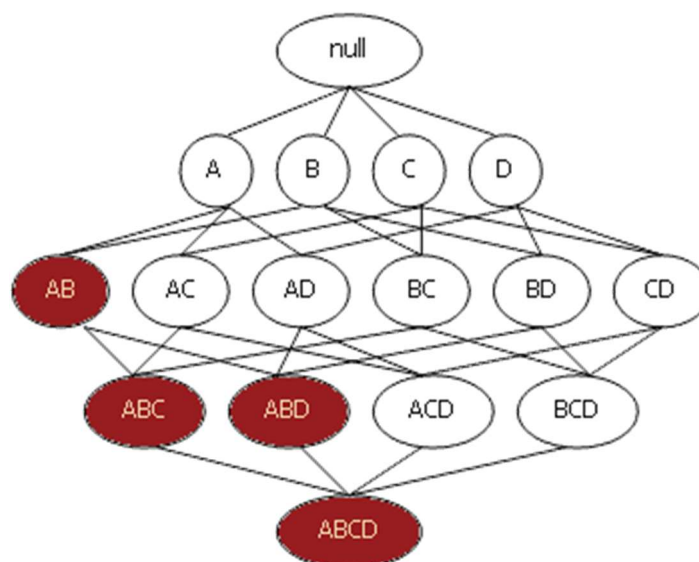


Рисунок 5. Дерево алгоритма Apriori.

Вполне очевидно, как ассоциативные правила использовать для выделения закономерностей в покупательских предпочтениях или иных транзакционных данных. Однако привести к транзакционному набору можно практически и набор параметров, признаков и даже некоторые численные значения, в которых нужно найти зависимости. Перед переходом к ассоциативным правилам важно выделить некоторый идентифицирующий общий признак, который будет являться идентификатором для транзакции. А остальные данные преобразуются в однотипный набор транзакций. Рассмотрим пример данных, на первый взгляд неподходящие под набор транзакций. Пусть мы ведем некоторую базу данных по разноцветным шарикам, учитывая их цвет, вес, цену материал и поставщика. Пример нескольких строк в базе приведен в Таблице 3. Можно видеть, что разные данные представлены различными типами, некоторые вообще являются непрерывными значениями. Также обратим внимание, что данные нормализованы, то есть поле Артикул является уникальным ключом в таблице.

Таблица 3. Пример разнородных признаков.

Артикул	Название	Цена, руб	Вес, г	Поставщик	Материал
001	Желтый шар	100	148	Завод «Звезда»	Пластик
002	Зеленый шарик	120	250	Фабрика Игрушек	Метал
003	Черный шарок	250	50	Комбинат шаров	Дерево
...	...	...	...	...	...

Для сведения данных к таблице транзакций с данными сведения в столбцах Цена и Вес проведем через операцию квантования, когда конечный набор непрерывного свойства заменяется на вхождение в некоторый интервал. После факты вхождения, а также признаки полагаем равноправными транзакциями. Тогда получаем таблицу следующего вида:

Таблица 4. Признаки как набор транзакций

Артикул (ID)	Транзакция
001	Желтый шар
001	Цена от 100 до 200 руб.
001	Вес от 101 до 150 г
001	Пластик
001	Завод «Звезда»
002	Зеленый шарик
002	Цена от 100 до 200 руб.
002	Вес более 150 г
002	Метал
002	Фабрика Игрушек
003	Черный шарок
003	Цена более 201 руб.
003	Вес менее 99 г.
003	Дерево
003	Комбинат шаров

Таким образом, на обработанных данных доступны методы нормализации и выделения ассоциативных правил. Заметим, что операция квантования непрерывных задач весьма критична с точки зрения выбора границ интервалов, и делать ее нужно очень зряче, исходя из бизнес-требований.

Как уже указывалось выше, и генетические алгоритмы и нейросети (особенно их модификацию – сети Кохонена) также могут входить в цепочку обработки при конкретном анализе данных и включаются в перечень Data Mining. Вообще список алгоритмов незакрытый, и в различных программных решениях в модуле Data Mining присутствуют разные наборы методов, что также размывает семантическую точность термина.



Дополнительно обратим внимание на следующий факт. При применении методов Data Mining на выбранном массиве данных технических проблем гораздо больше в самих данных: их упорядоченность, полнота, непротиворечивость. Достаточно вспомнить, что при внедрении информационных систем в государственных ведомствах проблемой стали данные [25]. Вообще степень зрелости организации благотворно сказывается на адекватности и полноте данных в ее информационных системах, что в, в свою очередь, дает на вход анализа Data Mining модель с меньшими несоответствиями. Поэтому в Data Mining как процесс входит стадия подготовки и очистки данных перед самым анализом.

### **§1.8 Будущее: децентрализованные базы данных**

В 2008 году в сети Интернет расходитя статья за авторством Сатоси Накамото, описывающая принципы работы протокола денежной единицы Bitcoin [26]. Идея состояла в том, что участники сделки в специальной базе данных фиксировали факт передачи денег от одного участника другому. Это похоже на классический перевод денег между счетами в банке. Новизной же было то, что у этой базы возникали два новых принципиальных свойства: децентрализованность и полная история все событий. Реализация этих свойств достигалась с помощью того, что копия все базы данных была у каждого из участников, и содержала бы историю всех операций всех участников. Для этого все изменения в базе данных отражались в виде криптографических операций, с расчетом контрольной суммы каждой операции. Таким образом все транзакции в такой базе данных выстраиваются в цепочку блоков, каждый из которых рассчитан специальной математической функцией. Эта функция относится к классу криптостойких хеш-функций [27].

Любой желающий может проверить, что транзакции не были изменены никем, рассчитав хеш-функцию самостоятельно. При наличии нескольких участников (децентрализация) подделка блоков вместе с хеш-функцией возможно только при сговоре более половины участников. В ином случае участники рассчитают разные результаты хеш-функций и блок не будет подтвержден. Такой способ организации базы данных получил название *блокчейн (blockchain – цепочка блоков)*. Способ синхронизации состояния базы данных называется *алгоритмом консенсуса*.

Долгое время идея Накамото воспринималась лишь как принципиально новый вид валюты – криптовалюта, и первой реализацией стала криптовалюта Биткойн [26]. Мы оставляем за пределами настоящей работы рассуждения о правомерности признания его денежной единицей. Нам здесь важно то, что блокчейн получил заметное распространение. Реализация децентрализованный базы данных для биткойна имела ряд ограничений, в частности язык описания событий поддерживал очень ограниченный набор инструкций. С

учетом этих ограничений Виталик Бутерин в 2013 году анонсирует новую платформу Ethereum [28], которая была предложена как универсальный блокчейн для разработок децентрализованных приложений. Платформа имеет продвинутый внутренний язык программирования – Solidity, на котором пишутся специальные инструкции – смартконтракты. Данные и инструкции в Ethereum обладают теми же принципиальными свойствами: децентрализация и полная история событий. При чем тут анализ данных? А вот при чем.

Свойства децентрализованности и полной истории имеют несколько важнейших следствий, которые представляют поистине невероятные возможности:

1. Неподделываемость и прозрачность любых действий участников. Все действия, совершенные в блокчейне, фиксируются в цепочке блоков, защищаются контрольной хешсуммой и не могут быть удалены или изменены кем-либо.
2. Распространение всех событий в базе данных одновременно с подтверждением. Любой участник, синхронизируя базу данных, имеет полную и непротиворечивую картину всего случившегося в базе, подтвержденную каждым участником сети.
3. Участники, использующую базу данных, для идентификации используют средства криптографической защиты, с помощью которых, а также системы электронных ключей возможно получить криптостойкую защиту данных одновременно с доступностью зашифрованной информации всем участникам.

Для примера разберем только один проект с использованием вышеуказанных свойств – Patientory ([www.patientory.com](http://www.patientory.com)). Представим себе человека, который решает разные проблемы со здоровьем в разных клиниках. В силу требований законодательства (что в РФ, что в ЕС или США) пациент предоставляет сведения о своем здоровье только лечащему его доктору. Правом на всю историю своего лечения обладает только он сам. В связи с этим в каждой клинике по его первому обращению сызнова заполняется анамнез заболеваний. Даже такие значимые вещи, как группа крови и аллергия на препараты может быть недоступна в экстренном случае. В сложных с медицинской точки зрения при случаях при переходе пациента клиники оформляют специальный документ – выписку из истории болезни. Однако вся медицинская информация очень важна для врача. А используя методы анализа данных, можно получить информацию, которая поможет врачу принять правильное решение, сберечь жизнь и здоровье пациента.

Проект Patientory разработал архитектуру приложения, работающего с использованием блокчейна Ethereum. Вся медицинская информация о пациенте хранится в блокчейне: диагнозы, протоколы докторов, сведения о процедурах и манипуляциях.

Сопоставление набора данных конкретному пациенту шифруется закрытым ключом и тоже помещается в блокчейн. Более того, распоряжение этим закрытым ключом делегируется пациенту. Каждая клиника, подключенная к приложению, хранит у себя копию всей базы данных как участник децентрализованной системы. В случае необходимости пациент, обращаясь в новую клинику, предоставляет врачу закрытый ключ, при помощи которого врач оперативно получает информацию из единой карты пациента, дополняя ее при необходимости сведениям о проводимых процедурах. Пациент имеет возможность поменять закрытый ключ, покинув клинику. В этой схеме обратим внимание на следующие факты, имеющие отношение к анализу данных:

1. При использовании алгоритмов Data Mining, помогающих принять решение, связанное со здоровьем пациента, анализ гарантированно имеет на входе полную и непротиворечивую информацию.
2. Используя прозрачные методы обезличивания информации в таких системах, можно иметь постоянный доступ до обновляемой, достоверной базы данных, что позволяет использовать методы Data Mining на широкой, «питательной» фактами выборке.
3. Обмен полученными знаниями в таких системах будет мгновенным и всеобщим, при этом блокчейн позволяет выработать прозрачные механизмы поощрения участников на выработку и внедрение знаний.

На момент написания работы автор участвовал в проектировании сообщества клинических центров, которые намеревались организовать работу с пациентами и друг другом при помощи формальных правил, записанных в смартконтракты блокчейн. Децентрализованность и полнота всей истории, а также дальнейшее развитие технологий блокчейн дает основания полагать, что в ближайшее время будут разработаны принципиально иные подходы к организации информации, когда лучшей защитой будет полная прозрачность и открытость. Тогда знания будут доступнее и полнее, чем сегодня. Это вдохнет новую жизнь в развитие математических методов анализа данных, алгоритмы самообучения информационных систем принятия решений.

## **2 ДАННЫЕ В КОММЕРЧЕСКОЙ МЕДИЦИНЕ**

### **§2.1 Здравоохранение как коммерческая деятельность**

Как было указано выше, автор руководил разработкой информационной системы, автоматизирующей деятельность многопрофильной коммерческой клиники. Это позволило изучить работу медицинского центра с точки зрения организации процессов в медицинском центре и некоторых особенностях управления.

Коммерческая медицина в РФ имеет интересную историю. Советский Союз предоставлял всю медицинскую помощь населению на безвозмездной основе. В перечень медицинских услуг для населения в связи с этим, например, не входили услуги косметологии. Такой вид, как пластическая хирургия, был доступен только по строгим медицинским показателям. Ряд медицинских организаций финансировались курирующими предприятиями и ведомствами в виде социальной нагрузки. Высококвалифицированная узкоспециализированная помощь была сосредоточена в крупных медицинских НИИ. Право на бесплатное образование и медицинскую помощь защищалось советским государством как фундаментальные достижения социалистического строя. Тем не менее в 1986 году офтальмолог с мировым именем Святослав Федоров организует Межотраслевой научно-технический комплекс «Микрохирургия глаза». Это учреждение получило право предоставлять ряд услуг населению на возмездной основе, вкпе с этим определять размеры заработной платы персонала. Для того времени такие полномочия были неслыханными в системе медицины, да и вообще в какой-либо деятельности. По факту это был первый официальной коммерческий медицинский центр.

С 1990 года вместе с другими отраслями образуется коммерческий сектор медицинских услуг. Долгое время он формировался за счет введения услуг на коммерческой основа при организациях, входящих в структуру Министерства здравоохранения. Вместе с общим ростом технологической составляющей в медицинской деятельности неукоснительной росли требования к техническому обеспечению медицинского учреждения любого профиля, что требует поддержания дорогостоящих и в капитальных и в операционных затратах основных фондов. Такое было под силу уже сложившимся центрам при отраслях и государственных ведомствах.

Тем не менее формирование небольших коммерческих центров также шло, и быстрее всего формировался сегмент дорогостоящих услуг околomedicalной направленности: пластическая хирургия, косметология, косметическая стоматология – именно в этих отраслях коммерческая медицина становилась заметно современнее государственной по оснащению и методологии. Центрами создания успешных коммерческих клиник чаще всего оказывались личности докторов, чьи заслуги или

методики обретали признание, и при этом не находились в сегменте обязательной или связанной с вопросами жизни и смерти отрасли: тот же Федоров, Дикуль [29], и т.д.

В формировании коммерческой медицины РФ также начали участвовать международные медицинские компании, в основном в лабораторной деятельности. Примерно до 2000 года развитие коммерческой медицины сдерживали низкая покупательная способность населения, несформированный рынок медицинского страхования, как государственного, так и коммерческого, общее недоверие населения к оказанию врачебных услуг за деньги, как в силу ментальности, так и в силу часто распространённой мотивации коммерческой клиники к навязыванию услуг [30]. Дополнительно ощущалась нехватка квалифицированного управленческого персонала. В силу традиций клинические учреждения возглавлял медицинский работник, в связи с чем успех клинического центра часто зависел от административных талантов центральной медицинской фигуры.

С 2000 года в РФ формируется устойчивый зрелый рынок добровольного медицинского страхования. Медицинская страховка становится важным элементом мотивации кадров в организациях всех типов. Дополнительно в РФ появляется школа менеджмента в медицине. Городское население, особенно молодежь, осознает здоровье как ценность, имеющая вполне понятное материальное воплощение. Профилактика заметно дешевле и более предсказуема, чем лечение хронических заболеваний. Стоит заметить, что в РФ, все равно, нет традиций регулярного посещения врача, особенно мужчинами, однако исследование этого факта находится за пределами работы. Важно то, что на 2017 год коммерческая медицина является вполне заметной отраслью экономики. К сожалению, дополнительным фактором, способствующим этому, является очевидное ослабление роли государственной медицины, доступной на бесплатной основе. Сокращается число стационаров, медицина тяготеет к концентрации в крупных высокотехнологичных центрах федерального масштаба, с падением количества врачей и кабинетов в поликлинической деятельности.

## **§2.2 Процессный подход в медицине**

Как и любая коммерческая организация, типовой медицинский центр ставит своей задачей извлечение прибыли. С общей зрелостью рынка в нем ужесточается конкуренция за пациентов. Здесь возникает заметное неэтическое противоречие. Оно состоит в том, что, получая деньги за услуги, любая клиника объективно заинтересована в увеличении среднего чека, что достигается расширением числа услуг и манипуляций для пациента. Противоречие это актуально не только для РФ и разрешается комплексом мер.

За сдерживание накручивания услуг выступает страховая компания. В медицине растет доля услуг, оказываемых по системе добровольного медицинского страхования, когда пациент или его работодатель покупает фиксированный по цене полис на фиксированный период времени, в течение которого лечебное учреждение, как подрядчик по отношению к страховой, обязуется предпринять все меры для излечения пациента. Тут страховая компания стоит на страже своих интересов и отслеживает обоснование назначений. Однако возникает опасность перекоса в другую сторону, и врач может находиться под давлением и не решиться на неочевидные исследования.

Другим фактором является растущие на рынке коммерческих медицинских услуг конкуренция и консолидация [30], которая все равно очень низка. При этом потенциалом для роста продаж обладает только база пациентов, которые оплачивают услуги коммерческой клиники сами – так называемая кассовая медицина. Ниже мы рассмотрим еще одну ценность именно этой группы пациентов. Здесь же обратим внимание на то, что обращающийся в медицинское заведение на регулярно основе пациент становится ценнее одного даже высокого чека. В связи с этим клиника уже заботится о формировании у потребителя ощущения ценности своих услуг, и стимуляция к назначению дополнительных процедур и манипуляций эту ценность заметно снижает.

Конкуренция, а также растущие требования регуляторов (Минздрав, Роспотребнадзор) формируют на рынке высокий порог вхождения, связанный с высокой стоимостью основных фондов, как по инвестициям, так и по содержанию. От медицинского центра для присутствия на рынке требуется наличие дорогостоящего диагностического оборудования, докторов с высокой квалификацией и регалиями, налаженная логистика лабораторных анализов. В связи с этим растут необходимые вложения в фонды и персонал, и тогда становится очень острой проблема их утилизации. В борьбе за утилизацию, например, компьютерных и магниторезонансных томографов сложилась вполне наглядная картина: автор предлагает читателю в любой поисковой системе в Интернете поискать услуги компьютерной или магниторезонансной терапии крупного города РФ и убедиться, сколь большое число центров работает в удлинённом графике вплоть до круглосуточного.

Вышеуказанные факторы дали толчок развитию аутсорсингов сервисов по наполнению клиентской базы коммерческих клиник: агрегаторы Яндекс.Здоровье, Zabota.ru, DocDoc и другие. Клиники, подобно ритейлу и другим B2C направлениям, проводят рекламные и мотивационные кампании, формируют программы различные программы лояльности – в общем, используют весь арсенал средств борьбы за пациентов.

Конкуренция формирует высокие требования к сервисной составляющей оказания медицинской услуги в коммерческой клинике. Тут нужно отметить следующее.

Особенностью медицинских услуг, связанных со здоровьем, сопряженных с болью и необходимостью пациенту прилагать свои усилия в лечении, является то, что субъективное ощущение пациента может быть очень далеко от объективных показателей качества проведенных манипуляций. Человек, склонный к минимизации своих усилий, охотнее верит «чуду» или доктору, который его обещает. В связи с этим в медицине, и прежде всего – в коммерческой, обитает большая группа околonaучных и далеких от медицины методов, пользующихся большой популярностью. Ярким примером является гомеопатия – очень широко распространенный способ лечения, совершенно далекий от так называемой доказательной медицины. К сожалению, в погоне за прибылью медицинские центры, получившие лицензии и укомплектованные действующими докторами, вполне часто прибегают к немедицинским методам. Помимо очень неэтичного обмана такое увлечение приводит к очень печальным последствиям для жизни и здоровья.

Вышеизложенное отвлечение было приведено для подчеркивания того факта, что сервис, становясь важной конкурентной составляющей медицинской клиники, по сути своей, является некоторым антагонистом качества медицинских услуг, которые должны оцениваться по показателям здоровья конечного больного. В связи с этим, нацеливать деятельность медицинской организации строго на удовлетворенность пациента не представляется возможным. Здесь сильна позиция регулятора и всего медицинского сообщества, и снижать такое регулирование очень опасно. В РФ до сих пор развиты традиции самолечения, и даже регулирование не имеет должного влияния. Скажем, препараты-антибиотики, включая детские, в большой своей части продаются без требования врачебного рецепта, что является нонсенсом для европейской или североамериканской медицины.

Для преодоления вышеуказанных разногласий, а также некоторой формализации деятельности докторов всемирной организацией здравоохранения был предложен единый классификатор заболеваний – МКБ [31]. Суть этого классификатора в том, что доктор, в случае постановки предварительного или окончательного диагноза назначает некоторый фиксированный набор манипуляций для уточнения диагноза или проведения лечения. В настоящий момент времени действует классификатор 10-ого пересмотра. Принятие следующего классификатора МКБ-11 ожидается в середине 2018 года. Министерство здравоохранения признает МКБ в качестве эталона для оценки действий врача. В случае конфликтной ситуации между пациентом или его представителями и доктором уполномоченные Министерством здравоохранения эксперты и правоохранительные органы оценивают решения врача по близости к стандартам МКБ. Дополнительно, такая

формализация упростила взаимоотношения между страховыми компаниями и клиниками, а также управление деятельностью самой клиники.

В поликлинической деятельности (оказываемой лабораторно и не связанной с тяжелыми хроническим и трудно излечимыми заболеваниями) доктор может в очень большой степени опираться на требования стандарта МКБ, формируя диагностические и лечебные манипуляции исходя из его требований. Это позволяет в большой степени алгоритмизировать врачебную деятельность, а также оценивать работу доктора более объективно. Таким образом, сервисная составляющая сосредотачивается в сопутствующей деятельности медицинского центра: логистике, информировании, удобстве посещения врача, общей вежливости персонала и так далее. Стало быть, клинике в борьбе за пациента необходимо формировать удовлетворенность именно этими аспектами оказания услуг. При этом, при должном информировании пациента, доктор вполне обосновывает свои назначения требованием открытого стандарта, предупреждая у пациента мнения о чрезмерном навязывании и необоснованности манипуляций.

Дополнительным источником увеличения среднего чека с пациента является сервис возврата. Есть даже аутсорсинговые компании, специализирующиеся на такой услуге. Имеется ввиду кампания по работе с пациентами, побывавшими в клинике, в целях привлечения их (возврата) на иные поликлинические услуги. К сожалению, у среднего человека здоровье с возрастом не улучшается. И здесь профилактика и плановые обследования поддерживаются врачебным сообществом. Коммерческие клиники за счет сервисной составляющей вполне конкурируют с клиниками системы государственного обязательного медицинского страхования: время работы, отсутствие очередей, необходимости хождения по многим кабинетам. Здесь вполне проявляется процессная ориентированность коммерческих клиник, не входя с противоречием с медицинскими требованиями и требованиями регулятора. Значит имеется сугубо рыночная, конкурентная составляющая деятельности медицинского центра, которая может и должна быть объектом управления, и значит и приложения методов извлечения знаний. При этом, как мы покажем ниже, и медицинской составляющей анализ данных может быть очень даже полезен.

### **§2.3 Анализ данных и медицинская тайна**

Взаимоотношения врача и пациента, помимо стандарта МКБ, регулируются законодательством в части защиты данных пациента. В общемировой практике присутствует понятия медицинской или врачебной тайны. Юридическое понятие врачебной тайны, в целом, близко в разных странах. Главным постулатом является то, что единственным владельцем и распорядителем медицинских сведений о себе является сам



пациент, равно как и правом принимать решение в части медицинских манипуляций. Мы не будем углубляться в практику применения, понятие дееспособности и законного представителя. Достаточно того, что данные о диагнозах, результатах диагностических процедур и проведенным манипуляциям пациента защищаются законом и не могут быть без ведома пациента куда-либо переданы. Это накладывает на анализ медицинских данных определенное ограничение. Коммерческий медицинский центр, действуя в рамках договора оказания медицинских услуг, обрабатывает данные пациента и отвечает за их сохранность. Доктор, проводящий конкретные медицинские процедуры, исходит из положений такого договора как представитель медицинского центра. Аналитик данных, в свою очередь, должен быть также допущен до обработки медицинских данных пациентов. Закон (323-ФЗ) исходит из положения «однозначной определимости пациента». То есть простое обезличивание данных не является защитой медицинских данных от компрометации. Например, очень редкое заболевание, даже без указания метрик пациента, может явиться возможностью однозначно его идентифицировать, что является нарушением врачебной тайны. В связи с этим, помимо обезличивания метрик, при анализе данных пациентов медицинских центров необходимо защищать входные данные аналитических алгоритмов, а также работать с большими выборками по количеству пациентов, рассматриваемому периоду, без необходимости не использовать в персонифицированном анализе отличные от метрик идентифицирующие признаки – пол, возраст и т.д.

При этом в медицинской деятельности как ни в какой другой методы анализа данных могут принести совершенно ощутимую, измеримую пользу, поскольку непосредственно влияют на здоровье человека. Об этом в следующей главе.

## **§2.4 Потенциал новых методов анализа и обработки информации**

Для дальнейшей определенности рассмотрения и принимая во внимание вышеизложенное разделим данные, накапливаемые коммерческой клиникой, на две принципиально разные составляющие. Первую формируют данные о диагнозах, манипуляциях и результатах диагностических процедур. Будем называть их медицинскими. Они защищаются требованиями к врачебной тайне. Вторые образуют сведения о пациентах как о клиентах – физических лицах, обратившимся в клинику за услугами. Мы понимаем, что интересен анализ, заходящий в оба подмножества, но сначала рассмотрим их отдельно в свете возможностей анализа данных.

Рассмотрим медицинские данные. Сначала мы зафиксируем, что медицинская деятельность опирается на ряд фундаментальных наук: биологию, химию, физику и их краевые приложения в изучении жизнедеятельности человеческого организма и влиянии на

него. Как и любая наука, здесь главенствует научный доказательный подход. Новые виды заболеваний, влияющих на них микроорганизмов, лекарственных средств проходят процедуры сначала фундаментальных исследований, а после - полноценных научных испытаний на живых организмах, а после и на людях с требованием соблюдения стандарта GCP [32]. С другой стороны, даже после таких испытаний очень важным источником обновления знаний в медицине является практика. Не зря крупные медицинские клиники федерального масштаба фактически являются научными центрами, где одновременно с лечением заболеваний по глубокой специализации постоянное изучение накопленного опыта. В этой связи методы анализа данных могут помочь научным работникам в формировании выжимки из такого опыта. Сейчас такие центры для научных исследований используют массивы медицинских данных и опираются на вычислительные средства как на облегчающие работу с большим количеством сведений. Очень полезным инструментом Data Mining оказывается для статистических выборок, способов очистки данных, формированию устойчивых закономерностей, что помогает формализовать диагностические признаки, учесть побочные эффекты, выявить иные слабые отклонения, имеющие принципиальные значения в медицинской сфере. Организации, специализирующиеся на клинических исследованиях в международном масштабе, а также в странах золотого миллиарда, проявили интерес к методам Data Mining с начала 2000-х годов [33].

В сложных клинических случаях врач имеет право обратиться к группе своих коллег, собирая консилиум – совет компетентных специалистов, которые уполномочены вынести некоторые рекомендации лечащему доктору. Решение все равно принимает лечащий доктор, неся полную ответственность. Идея такого совета очевидна: расширить поле вариантов для лечащего доктора за счет практики и компетенций других докторов. По сути своей, если вести учет практических случаев в информационной системе специальным образом, а потом, по требованию, предоставлять лечащему доктору уже выжимку выделенных знаний средствами Data Mining, то хоть ценность такой информации и воспринимается ниже (хотя здесь вопрос скорее психологический), но получить доступ до таких знаний можно заметно быстрее, при этом зафиксировать на основе, например, выделения ассоциативных правил, совершенно новые закономерности. Дополнив базу данных как сырой практики, так и подтверждённых выделенных правил результатом своих изысканий и фактическим исходом лечения, врач передает уже следующему коллеге формализованную выжимку из знаний. Такая итеративная информационная система может формулировать рекомендуемые необходимые шаги в диагностике и лечении с такими вероятностными весами, что в критических случаях систему может использовать и

непрофильный специалист, волею случая вынужденный принимать необходимое решение на месте. Имею мнение, что, когда такая модель позволит спасти чью-то жизнь, можно смело утверждать, что Data Mining однозначно полезное изобретение человечества.

Еще одним приложением алгоритмов анализа данных является построение самообучаемых систем, подсказывающих врачам действия по массиву относительно стандартных случаев. Здесь используются алгоритмы нечетких множеств и нейросети. На текущем уровне развития технических средств в мире нет известной системы, которая может заменить врача в диагностике [34]. Хотя ошибка подобных систем после первичного самообучения уже не превышает ошибки среднестатистического врача, этические и юридические вопросы не позволяют делать какую-либо информационную систему ответственной за врачебное решение. Наиболее вероятным представляется такое применение самообучаемых систем, когда они становятся механизмом проверки и поддержки принятия решений, как было описано выше. В этом случае некоторая диагностическая система строит вероятностное поле исходов диагностики, которое является дополнительной информацией для лечащего врача. Особенно это важно в том случае, когда на базе выделенных ассоциативных правил будет выделена слабая, но устойчивая зависимость, определяющая, что некоторый набор признаков может свидетельствовать о ранней стадии тяжелого заболевания. В этом случае система может порекомендовать врачу сделать дополнительные исследования дабы исключить заболевания либо диагностировать на ранней стадии.

Что касается второго вида выделяемых данных. Все вышеуказанные применения методов анализа данных приложены к медицине как к специфической, имеющей огромное гуманитарное и экономическое значение отрасли человеческой деятельности. Но не будем забывать, что нам интересна как объект приложения исследований коммерческая клиника, решающая наравне с вышеперечисленными вопросами и задачу получения, прибыли от ведения деятельности. Как мы уже говорили, конкурентная борьба клинических центров ведется как в качестве оказания медицинских услуг, так, в большей степени, и в сервисной составляющей сопровождения пациента. Поддержание активной базы пациентов, у которых клиника ассоциируется с созданной ценностью, становится одной из важнейших задач. Таким образом можно держать высокий уровень утилизации специалистов, дорогостоящего оборудования, койко-мест стационара. Задачи эти традиционно решаются маркетологами, и поддержать их в этих решениях можно методами Data Mining. Используя алгоритмы классификации и выделения знаний, можно выделить в поведении пациентов, зафиксированных в некоторых фактах и признаках, устойчивые правила, с которыми уже удобно работать маркетологу в своих мотивационных задачах. Ведь в коммерческой

медицине особенно маркетологу чуть сложнее, чем в работе с классическими розничными покупателями. Поведение последних по отношению к продукции и услугам массового сегмента гораздо лучше поддается гипотезам, больше открытых источников данных, гораздо устойчивее и более открытая обратная связь. В вопросах, касающихся здоровья людей, присутствует врачебная тайна, общее нежелание обсуждать что-либо с кем-то, кроме своего доктора. А как было сказано выше, далеко не всегда качественная медицинская услуга соответствует удовлетворенности пациента, хотя лечебная задача была решена. Именно с такой ситуацией столкнулся автор в исследовании, описанном в главе 4.

Есть еще одна ценность, которую может сформировать анализ данных в коммерческой медицине. Речь идет о таких знаниях, которые интересны фармакологическим компаниям. Прогнозирование спроса на препараты или даже фармакологические группы составляет в фармакологии большую проблему. В частности, из исследования DSM за 2015 год [35] видно, что у лидеров фармакологического рынка более половины выручки формируют розничные аптечные сети. Говоря иначе, компании зарабатывают на продаже лекарств, которые покупают граждане в аптеках. В масштабах производственной компании, потери на возвратах, невыкупленных лекарствах с истекшим сроком, а также неудовлетворенный повышенный спрос является причиной заметных издержек. Другой вопрос, что передача информации о диагнозах и врачебных назначениях не может передаваться фармакологическим компаниям в прямом виде, поскольку нарушает врачебную тайну. Однако, проведенные внутри клиники исследования, результатом которых уже являются совершенно обезличенные выделенные ассоциативные правила или прогнозы, построенные на иных алгоритмах, охотно покупаются у клиник, пациентопоток которых составляет от сотен уникальных посещений в месяц.

Традиционно выжимку по заболеваниям, а также по фактологии назначений в обезличенной форме приобретают компании, продающие услуги добровольного медицинского страхования. Скорость корректировки своих ценовых предложений и постоянный анализ своих рисков и вероятности является для них необходимыми условиями выживания на рынке, как и в любой сфере страхования.

### **3 ВЫДЕЛЕНИЕ ЗНАНИЙ В БАЗЕ ДАННЫХ КОММЕРЧЕСКОЙ КЛИНИКИ**

#### **§3.1 Объект исследования**

В главе 2 приведен анализ возможностей методов анализа данных в медицине. Автору, предлагающему анализ данных в качестве меры повышения качества работы клинического центра, показалось, что в исследовании выделенных нами подмножества медицинских данных больше как юридических сложностей, так и скептицизма врачей. При этом на момент проведения работ в маркетинге Data Mining был уже более привычен. В связи с этим было решено продемонстрировать результативность методов анализа данных на выделении неочевидных правил в признаках, присущим пациентам. В этом случае ожидался достаточно явно интерпретируемый результат, который продемонстрирует административному руководству клинического центра возможности и работоспособность методов анализа данных.

Как уже говорилось выше, на момент написания работы автор состоял в должности технического директора (руководителя разработки) компании, которая разрабатывала программное обеспечение, автоматизирующее деятельность коммерческой медицинской организации – медицинскую информационную систему (далее МИС). В разрабатываемую МИС входил постоянно растущий в функциональности блок экономической аналитики деятельности клиники. До момента обращения внимания на Data Mining МИС была способна статически анализировать процент невозвратных пациентов, высчитывать средний чек и другие очевидно выделяемые показатели для руководителей и акционеров. Опытные менеджеры вполне могли работать с такими показателями для выработки управленческих решений, но этим, собственно, дело и было ограничено. Дальнейшие действия целиком определялись компетенциями менеджера как в интерпретации этих показателей, так и в определении взаимосвязей. Среди организаций, использующих МИС, нашлась клиника, обозначившая интерес к новым методам анализа данных, а также к использованию искусственного интеллекта в своей деятельности. Вышеуказанная организация представляет собой сеть, состоящую из 6 клинических центров, объединенных под одним брендом и администрируемая единой управляющей компанией (далее УК). УК формировала стратегию создания у потребителя ценности за счет ориентирования всей деятельности клиники на пациента. На все центры МИС разделяла единую базу пациентов (медицинских карт), стандарты лечения, единый коллцентр. Активно использовались методы возврата пациента и удержания имеющейся клиентской базы. Автором работы было предложено провести эксперимент по профилированию базы пациентов для выделения признаков и закономерностей, присущих наиболее экономически активным

пациентам. УК огласилась на проведение эксперимента в надежде выделить неочевидные знания из признаков и поведения пациентов, с тем, чтобы также убедить врачебную часть руководства в применимости методов анализа данных и в медицинской деятельности.

Ожидаемый результат был также интересен руководству компании-разработчику МИС для получения компетенций, необходимых для разработки в программном обеспечении блока анализа и выделения знаний как части подсистемы поддержки решений врача. Дополнительный интерес был вызван внимание фармакологических компаний к выжимке данных по диагнозам и лекарственным назначениям. К моменту инициации проекта УК вела переговоры о передаче информационной базы фармакологическим компаниям, а разработчик МИС изучал варианты обезличивания данных пациентов в передаваемой выборке. Поскольку алгоритмы Data Mining позволяют добиться на основании данных обезличенные знания высокого передела, которые ценны фармакологическим компаниям, то можно продавать именно их без нарушения законодательства в части защиты персональных данных.

В ходе подготовки и проведения эксперимента были выявлены следующие трудности и ограничения:

1. Отсутствие выделенного ресурса в части специализированного аналитика под задачу. Автору, в условиях высокой текущей нагрузки, необходимо было самостоятельно или с минимальным отвлечением команды разработки провести полноценный эксперимент с интерпретацией данных.

2. Отсутствием в распоряжении как УК, так и разработчика МИС программного решения с лицензией, позволяющей проводить анализ данных с очевидной коммерческой выгодой.

3. Необходимостью провести эксперимент в таких условиях и таким инструментарием, чтобы он был воспроизводим в дальнейшем без крупных финансовых затрат.

4. Дефицит теоретической литературы по тематике. На сайтах поставщиков аналитического ПО присутствует документация по применению соответствующего инструмента, с подробным описанием шагов. Полноценного же учебника с фундаментальными основами по ассоциативным правилам и вообще по Data Mining обнаружить не удалось. Можно утверждать, что знания про то, как добывать знания, даются с трудом.

С учетом требований и ограничений переходим к этапу выбора методов решения задачи. Сначала выберем программные инструментарий для решения поставленных задач

из вариантов, доступных для разработки либо в готовом виде на рынке специализированного ПО.

### **§3.2 Обзор решений и выбор инструментария для эксперимента**

Для проведения эксперимента необходимо начать с выбора программного решения, причем даже до нисходящего разбиения задачи на этапы для решений. Связано это со следующим. Как мы уже говорили, обычно в конкретном экземпляре эксперимента по выделению знаний больше проблем не с работой алгоритма, а с подготовкой выборки из исходной базы данных, очистки, приведения к формату, необходимому для работы алгоритма. Эти интерфейсные детали целиком определяются инструментарием в выбранном решении, в связи с чем, исходя из требований ПО, и надо определять состав работ и спецификацию выборки данных.

Сначала рассмотрим вариант вообще провести эксперимент без стороннего ПО. Вариант вполне работоспособный, поскольку для эксперимента был выбран алгоритм выделения ассоциативных правил Apriori, который открыт, описан и даже имеет ряд доступных бесплатных библиотек. Для его реализации необходимо выделить ресурс человеко-часов разработки, при этом даже в случае создания некоторого «разового» программного решения под задачу подразделение разработки получает необходимую ценную практику. Дополнительные удобства такой реализации в том, что разработка глубоко погружена в контекст модели данных и программных интерфейсов МИС, откуда берутся исходные данные для анализа, что упрощает этап подготовки, загрузки и очистки данных для работы алгоритма. К сожалению, вариант был отклонен в свете приведенных выше ограничений. Выделить человекочасы разработки не представлялось возможным. Как раз результат выделения знаний и предполагалось использовать для ресурсного маневра в подразделении разработки.

Среди коммерческих продуктов, предоставляющих алгоритмы Data Mining были выделены специализированные решения. Дело в том, что Data Mining включается в серьезные программные решения ERP и CRM направленности в качестве отдельного модуля. Преимущества вполне понятны: работа алгоритмов ведется в одном контуре данных, нет необходимости в интерфейсной обработке данных, их очистке и подготовке для работы алгоритма. В нашем случае данные находились в работающей в клинике МИС, и переходить из-за эксперимента к новой информационной системе, конечно же, не имело смысла. Но для себя отметим, что что серьёзные корпоративные системы автоматизации бизнеса определяют Data Mining как commodity для ведения бизнеса.

Обратимся к имеющимся на рынке программным решениям. Заметим, что в работе приведены изученные автором в ходе подготовки эксперимента программные средства, в связи с чем обзор не претендует на полноту.

Первым был рассмотрен продукт RapidMiner [36]. В настоящее время это полноценная платформа, написанная на языке Java, представляющая целый набор программных решений. Для работы с данными необходим RapidMiner Studio – собственно среда для анализа. Остальные продукты формируют экосистему, предлагая сервер для больших объемов, корпоративное хранилище данных для анализа, и, конечно, средства подготовки данных за отдельные деньги. Дополнительно предлагают пользоваться их облачным сервисом хранения вместо локального сервера. В бесплатной версии доступны почти все алгоритмы Data Mining, но ограничена исходная выборка для анализа. Максимальна бесплатная версия обрабатывает 10 000 строк. Это не так уже много, поскольку практически все алгоритмы на вход требуют данные в строгом формате, в виде плоской ненормализованной таблицы. Программа, для упрощения работы, визуализирует последовательность обработки в виде конструктора блоков с входами и выходами. Представляется вполне подробная документация по решению задач, но, как и говорилось, нет теоретических основ. Внешний вид рабочего окна представлен на рисунке 6.

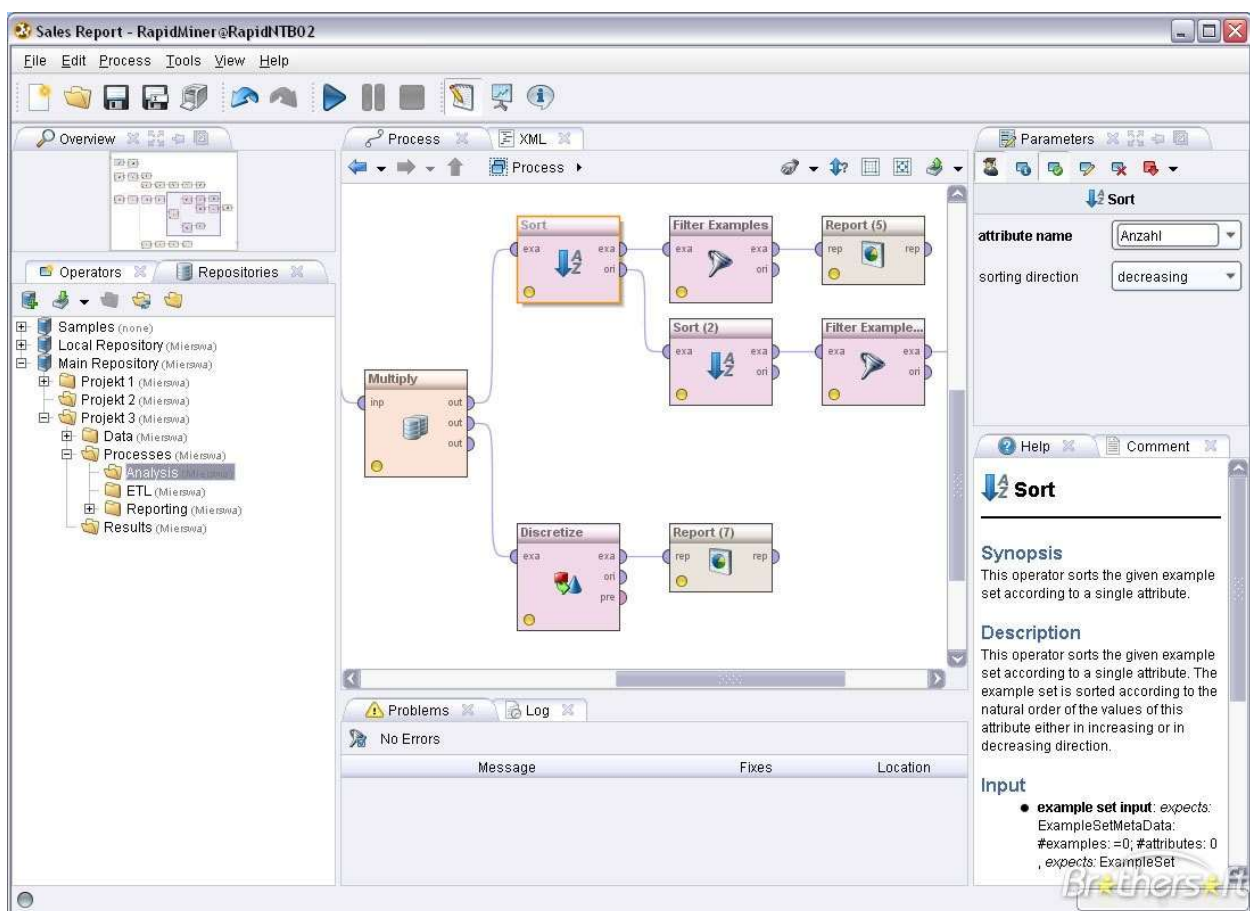


Рисунок 6. Пример анализа в RapidMiner.



Прохождение алгоритмов программируется путем соединения блоков в визуальном конструкторе. Каждый блок означает этап работы с данными, при этом этапом может являться и алгоритм выделения знаний, и очистка данных, их фильтрация или иная обработка. Решение, по личному наблюдению автора, является флагманским благодаря сочетанию высокой емкости по алгоритмам и нацеленности на простоту работы. Было рекомендовано научным руководителем работы, а также широко представлено на специализированных ресурсах. С этим решением даже проходила работа. Решение не было выбрано потому, что ограничение в 10 000 строк мешало эксперименту с ассоциативными правилами, дополнительно возникли трудности с работой с русскоязычной информацией в выборке. Но для дальнейшей возможной работы решение было отмечено за удобство, хорошую поддержку практических кейсов, большое сообщество поддержки.

Мощным решением для профессионалов является проект R [37]. По сути R представляет собой специализированный язык программирования алгоритмов Data Mining и некоторой графической оболочки к нему. Позиционирует себя как фактический стандарт реализации алгоритмов Data Mining. Здесь можно согласиться, поскольку специализированный язык описания, по определению, способен точнее и полнее описать решение задачи. Разрабатывается сообществом, но во главе с Lucent Tech. Является открытым по исходным кодам (open source) решением. Очень востребован в случае реализации внешних библиотек для имеющихся программных решений, а также для научных разработок. В силу невозможности погружаться в новый, глубокий технический контекст, решение было отклонено. Все интерфейсные задачи, подготовка данных и работа с результатами за пределами решения технически сложны и лежат на разработчике. Типовое окно работы программы представлено на рисунке 7. Как мы видим, для «зрячего» проведения эксперимента, получения результата, а также для решения возникающих технических проблем необходимо работать с кодом задачи на языке. Формирование графиков или выборок сродни ручному подключению дополнительных библиотек. Как мы уже сказали, такое решение дает гораздо больше гибкости, но и требует большего погружения. За таким решением необходимо закреплять высокоуровневого аналитика с навыками объектно-ориентированного программирования и глубокими знаниями теории Data Mining. Продукт не сопровождается теоретической литературой, но имеет хорошую поддержку сообщества, равно как и богатую открытую библиотеку практических кейсов. Даже в случае проведения эксперимента на решении вопрос о его внедрении в компанию разработчик МИС или в УК осложнился бы поисками или подготовки соответствующего специалиста, что сильно подняло бы совокупную стоимость владения решением (ТСО).

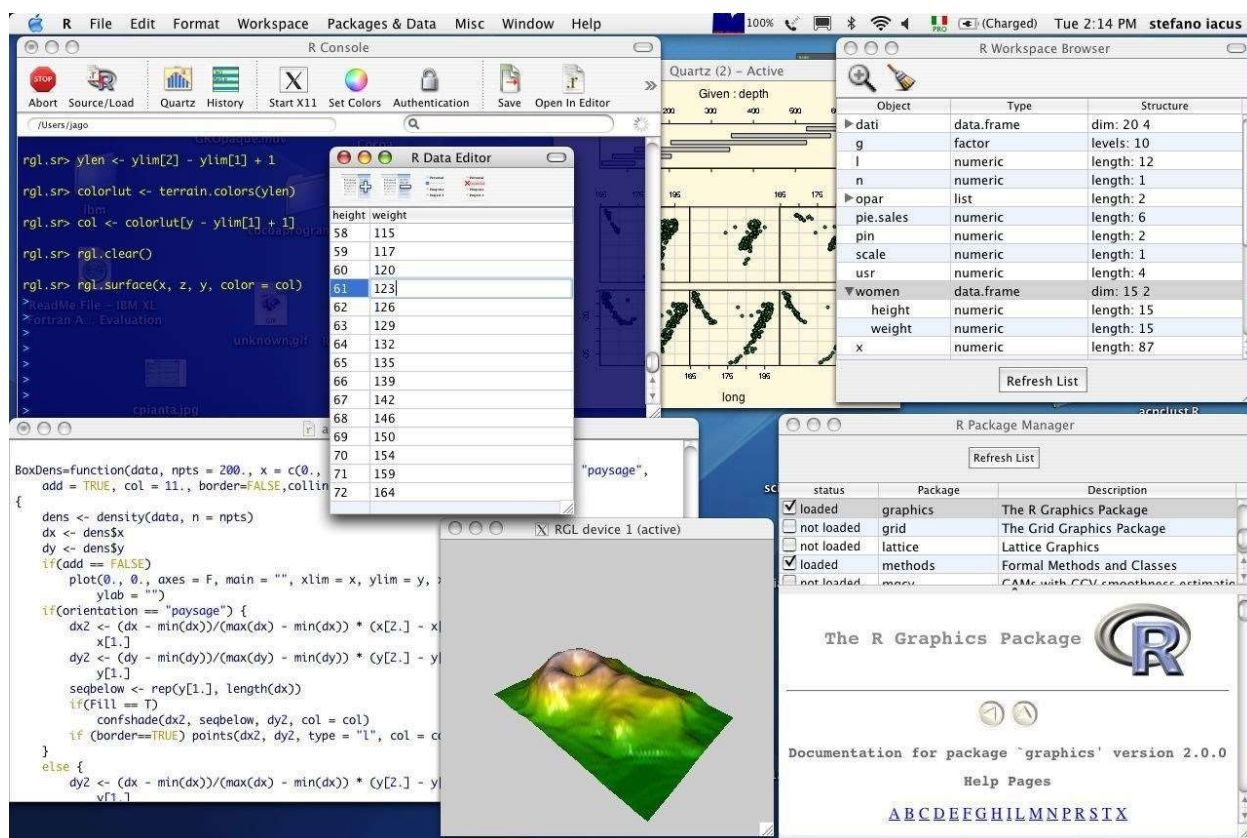


Рисунок 7. Рабочее окно R в MacOs.

Следующее рассмотренное решение – открытая (open-source) платформа Orange [38]. Платформа написана на языке Python. Представляет собой удачный гибрид визуального конструктор с элементами, из которых моделируется эксперимент, и тонкой настройки экземпляров анализа вплоть до алгоритмов на языке Python. Имеет хорошо документируемые требования к формату входных данных, есть дополнительные компоненты формирования запроса к SQL базе. На рисунках 8 и 9 приведены примеры рабочего окна приложения, демонстрирующие возможность работы с исходными данными и возможности обработки слабоструктурированных данных для анализа. Решение было рассмотрено и отклонено из-за нехватки времени на освоение. Выбранное ниже решение оказалось намного проще. Тем не менее, Orange очень интересен именно для освоения методов Data Mining студентами или начинающими специалистами, имеет хорошую поддержку методической документацией, дружелюбным сообществом, готовыми примерами. Возможно, поэтому Orange имеет широкое распространение в академической среде. Дополнительно выделим чуть более широкую, чем в других решениях, инструментальную поддержку генетических и других алгоритмов, востребованных в генетике. Среди благодарностей преобладают обращения генетических лабораторий.

Рисунок 8. Пример работы с входными данными в Orange.

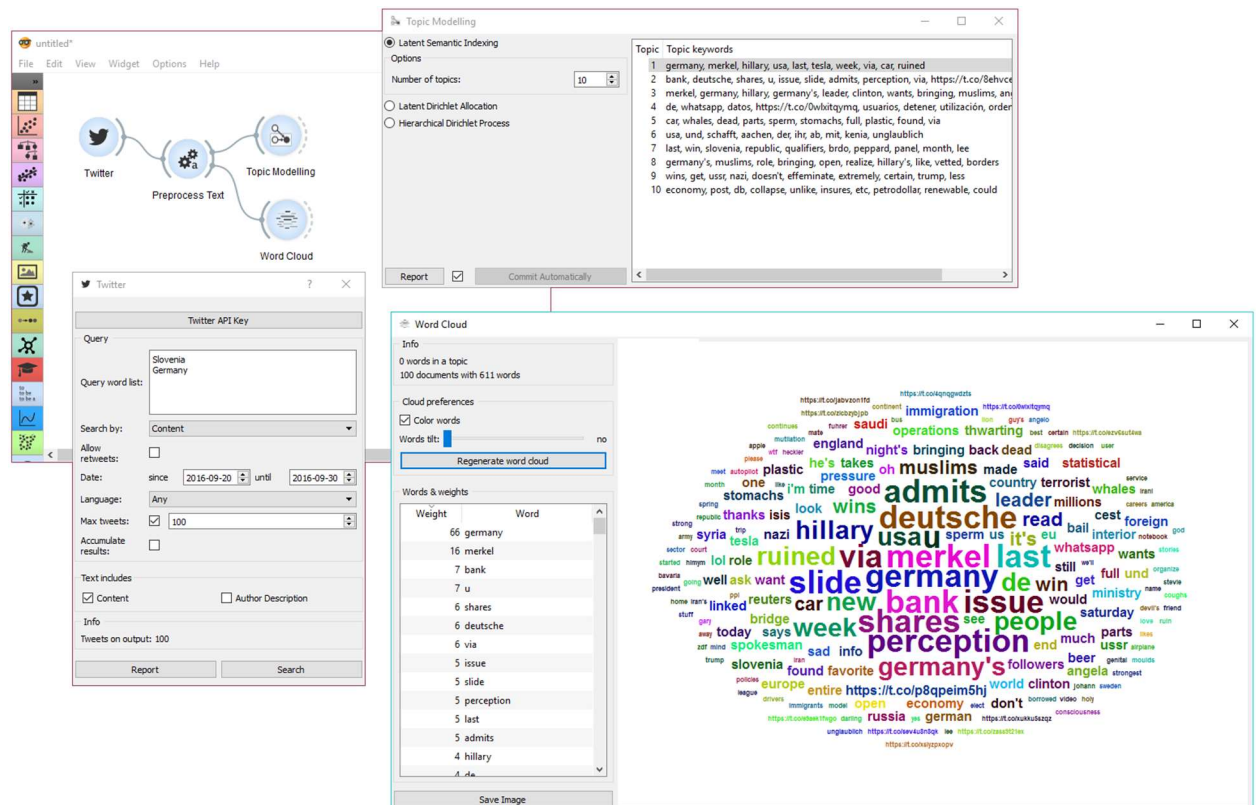


Рисунок 9. Пример обработки слабоструктурированных данных в Orange.

Из полностью коммерческих продуктов рассмотрим продукт компании StatSoft: Statistica DataMiner [39]. Продукт представляет собой модульный «конструктор», из которого опционально набирается требуемый функционал. Продукт коммерческий, при этом имеет удобно проработанный интерфейс загрузки входных данных с различных источников. Как мы видим из названия и компании-разработчика и самого продукта, StatSoft ранее разрабатывал пакеты статической обработки информации, а теперь дополнил их методом Data Mining. Компания предоставляет широкую методическую поддержку по своим продуктам, включая документацию. Вообще, судя по направлениям деятельности, консалтинг по работе с продуктом и данными составляет заметную часть бизнеса. Это в целом заметная для поставщиков аналитического ПО тенденции: продавать не только продукт, но решение задачи, разовое или регулярное. Примеры рабочих окон ПО Statistica представлены на рисунке 10. Обратим внимание, что методы продукта в большей части написаны на языке R, рассмотренном выше. Также отметим высокое внимание разработчика к медицине, к сопровождению доказательных исследований и выборок в клиническом анализе. К сожалению, функционала продукта явно не хватало для выбранного эксперимента, демонстрационную лицензию поставщик не предоставил, но очень настойчиво предлагал консалтинг. Не бесплатно.

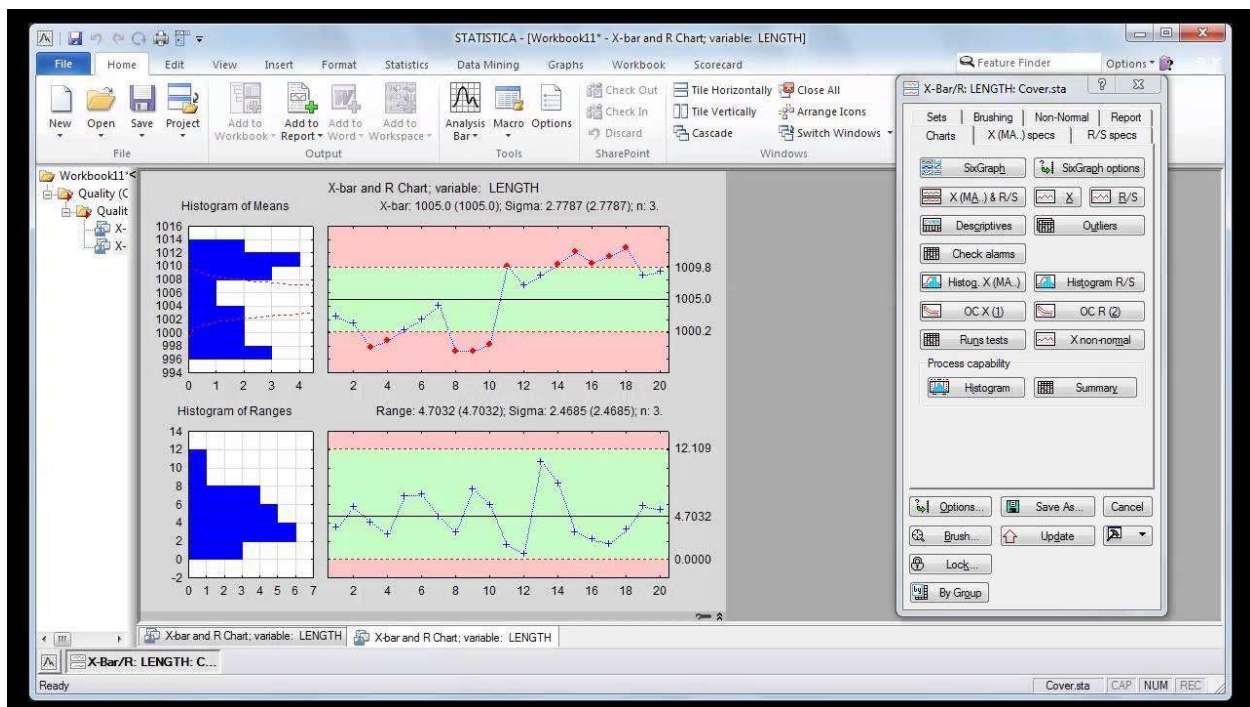


Рисунок 10. Рабочее окно программы «Statistica».

Переходим к решениям российских разработчиков. Тут можно сделать некоторое широкое отступление. Если окинуть взором новые технологии, меняющие мир, в том числе Data Mining, то они все больше лежат в плоскости сугубо интеллектуальных усилий. А еще эти технологии сами по себе являются локомотивом всей науки, особенно прикладной. Как мы уже говорили выше, научный прорыв фактически являет собой новые, неизведанные знания, полученные в ходе научных изысканий. Даже если изыскания эти приложены к сугубо экспериментальным данным с материальными телами, то Data Mining способен ускорить обработку и выделение результата, а если мы говорим о выделении закономерности, на которую следует обратить внимание, и в которой потенциально возможен научный прорыв, так он для того и предназначен. Тем отраднее было автору обратить внимание, что в России есть умы, активно и продуктивно интересующиеся темой. Как непосредственно знакомый с разработкой программных решений, автор прекрасно отдает отчет, что объем инвестиций имеет значение. Он превращается в человекочасы квалифицированных аналитиков, программистов, тестировщиков, и позволяет быстрее и эффективнее добиться поставленной цели. Здесь мы, конечно же, уступаем западным компаниям и инвестфондам. Бюджеты затрат на алгоритмы Data Mining исчисляются в миллиардах долларов США [40]. Но также автор убежден, что в эпоху совершенно новых осмыслений работы с информацией российская академическая школа и, если угодно, генетическое богатство наших математиков и программистов может явить нам тот самый шанс перехода экономики на «новую нефть», под которой понимают знания.

Традиционным флагманом в РФ считается корпорация Яндекс и ее направление Yandex Data Factory [41]. Коробочного ПО компании не предлагает, а продает свои услуги, или, вернее, «решения проблем» средствами Data Mining. Изначально направление росло, решая задачи самой компании Яндекс, но универсальность накопленных компетенций сподвигли подразделение выделиться и выходить на рынок. Маркетинговая стратегия нацелена на крупные сырьевые и промышленные компании. Некоторым отличием Yandex Data Factory от интеграторов, предлагающих разработку решения на заказ на том же языке R или поставку зарубежного коробочного ПО, является заявленная открытость решения и участие в нем большого сообщества. На деле продукт оказался гораздо более закрытым, чем мощные готовые решения с облачными модулями у тех RapidMiner или Orange. В частности, примеров работы найти не удалось, все результаты работы представляют собой закрытые отчеты для заказчиков. Воспользоваться в эксперименте решением не удалось. В качестве документации по решению самым информативным открытым источником оказался презентационный Whitepaper (краткий обзор продукта) на английском языке [42]. Автор надеется, что компания Яндекс доформирует свое решение до коробочного и будет

активно развивать открытое сообщество участников. Такой формат работы кажется более перспективным и полезным.

Решение российской компании Base Group Labs – ПО Deductor Studio [43]. ПО позиционируется как платформа, направленная на максимальное упрощение работы аналитика, не требуя навыков программирования и глубокого погружения в теорию. Вместо конструктора блоков Deductor Studio предлагает работать с деревьями, в которых каждый этап работы моделируется элементом дерева, что позволяет вводить вариации в цепочки алгоритмов, а также визуализировать любой этап обработки при необходимости. Компания оказывает очень широкую методическую поддержку как по самому продукту, так и по методам Data Mining, открыто публикуя теоретические основы анализа данных. Заявленная нацеленность на простоту работы обязала разработать удобные средства ввода данных, включая подключения к SQL базам источника. Продукт составляет часть экосистемы, к которой предлагается мощный сервер корпоративного хранилища данных, серверная платформа для больших объемов, а также продукт интеграции с корпоративной шиной данных. В этой экосистеме Deductor Studio позиционируется как ПО для АРМ аналитика, которое вполне способно решать задачи и локально. Для нашего эксперимента в свете ограничений продукт оказался очень подходящим. К тому же компания Base Group Labs пошла навстречу автору работы и предоставила полноценную лицензию на время проведения эксперимента. Таким образом, именно это решение и было выбрано для проведения всех работ по выделению знаний.

### **§3.3 Описание эксперимента**

Определив инструмент, перейдем к решению задачи. Как мы уже говорили, определив ПО и интерфейсные требования, можно переходить к планированию состав работ. Deductor Studio имеет удобные средства импорта данных из формата MS Excel, что и определило, как мы будем формировать данные. Итак, был составлен следующий план эксперимента - см. Таблицу 5. Напомним, что автор работы руководил подразделением разработки в компании поставщике МИС, а заказчиком эксперимента и потребителем результата является управляющая компания сети клиник, которой интересен результат. Мы привели это для понимания ответственных сторон в 4 колонке таблицы. План работ даже в таком простом виде позволил разграничить ответственность и лучше управлять ходом эксперимента.



Таблица 5. План работ по проведению эксперимента.

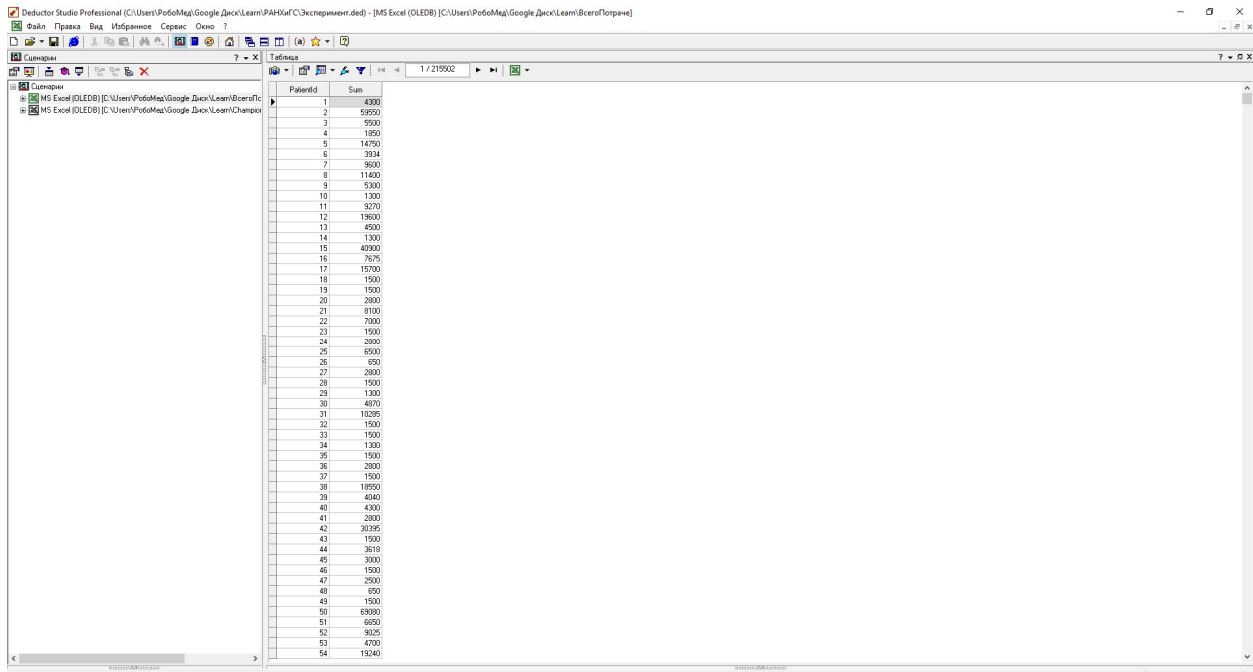
№ этапа	Перечень работ	Ожидаемый результат	Ответственный
1	Определение бизнес-вопроса или зоны интереса для приложения Data Mining.	Определена область приложения эксперимента	Отдел маркетинга УК
2	Формирование выгрузки данных из МИС в формате XLS	Выгружена таблица с данными в формате XLS	Разработчик МИС
3	Формирование словесного плана работы алгоритмов	Словесный план работ сформирован и передан в УК	Разработчик МИС
4	Согласование словесного плана работ из п.3	Словесный план работ согласован УК	УК
5	Проведение эксперимента	Выделен результат (правила)	Разработчик МИС
6	Прием и интерпретация результата	Результат принят и признан \ не признан полезным	УК

Далее будем двигаться по плану, комментируя действия и результаты на каждом шагу, а также решаемые технические трудности.

На шаге 1 маркетинговое подразделение УК сформировало зону интересов в виде некоторого бизнес-вопроса. Он звучал следующим образом: «Каковы немедицинские признаки, свойственные группе пациентов, которая принесла наибольшую сумму денег клинике в течение 2017 года». Прежде всего, размер суммы отдел определить не смог, сформировав его достаточно нечетко. Работа с грациями была отложена на следующие шаги, на текущем этапе такая постановка задачи определила, что мы будем выгружать из базы данных МИС на шаге 2. Было сделано следующее. Сначала проанализировали объем выплаченных в пользу клиник сумм, и выгрузили таблицу согласно рисунку 11. В выборку попали:

- Patient ID – идентификатор пациента, тождественно равный номеру карты в сквозной базе карт центра;

• Sum – общая сумма выплаченных в кассу клиники денежных средств. Общая базовая стоимость оказанных услуг может отличаться от полученных денег в связи со скидками, акциями и т.д.



PatientId	Sum
1	4300
2	5950
3	590
4	1850
5	14750
6	2534
7	8600
8	11400
9	5300
10	1300
11	5270
12	18600
13	4500
14	1300
15	40800
16	7675
17	15700
18	1500
19	1500
20	2800
21	8100
22	7000
23	1500
24	20900
25	8500
26	850
27	2800
28	1500
29	1300
30	4870
31	10295
32	1500
33	1500
34	1300
35	1500
36	2800
37	1500
38	18950
39	4040
40	4300
41	2800
42	20395
43	1500
44	3618
45	3000
46	1500
47	2500
48	650
49	1500
50	63000
51	6050
52	9025
53	4700
54	15240

Рисунок 11. Таблица принесенных в кассу клиники денежных средств.

На данном этапе нам интересно выделить подмножество самых «ценных» по принесенным деньгам пациентов. Как мы уже сказали, выделить такое подмножество формально со стороны заказчика не представлялось возможным. Эту задачу мы тоже возложили на DataMining, проведя по множеству карт и сумм операцию квантования, и, для наглядности – операцию разложения по столбцам (создание кросс-таблицы). Исходя из закона Парето (20% всех пациентов максимально полезны по принесенной выручке), квантование проходило на 5 интервалов. Ветка сценариев Deductor Studio приведена на рисунке 12, фрагмент получившейся таблицы – на рисунке 13.



Рисунок 12. Ветка выделения подмножества полезных пациентов.



PatientId	<...>	0	3200	4500	7500	15200
1			1			
2						1
3				1		
4		1				
5					1	

Рисунок 13. Фрагмент таблицы после квантования и разложения

Далее было выделено подмножество карт, входящих в максимальную по выручке квантовую группу, и направлено в разработку. По этим картам были выгружены признаки, формирующие таблицу, представленную на рисунке 14. В перечень полей попали:

- Номер карты - идентификатор пациента в клинике. У одного физического всегда одна медицинская карта;
- Пол – вполне очевидное поле и достаточно важный признак пациента;
- Дата рождения – тоже очевидное поле, нужно для соотнесения в возрастную группу;
- Источник рекламы – здесь из конечного списка выбирается некоторый информационный источник, сообщивший о клинике пациенту;
- Постоянная скидка – устанавливаемый жестко на пациента параметр. Исходно может быть связан с рядом факторов, формального признака формирования не имеет;
- Количество посещений – общее количество приходов в клинику на пациенте. Параметр, который, пусть и опосредовано, но показывает некоторую возвратность пациента в медицинский центр.

Номер карты	Пол	F3	Источник рекламы	Постоянная скидка	Количество посещений
15	М	27.07.1980	Другое	0	2
42	Ж	05.10.1950	Другое	15	3
71	Ж	01.03.1964	Другое	0	2
160	М	14.07.1972	Другое	10	62
307	Ж	02.02.1974	Другое	5	2
348	М	23.09.1968	Другое	0	7
429	Ж	29.08.1950	Другое	10	10
489	Ж	13.08.1967	Другое	0	1
605	Ж	20.12.1961	Другое	10	34
639	Ж	11.10.1960	Другое	0	5
663	М	09.04.1950	Другое	0	2
696	Ж	27.06.1970	Другое	0	24
1123	Ж	05.01.1949	Другое	10	6
1239	Ж	02.10.1961	Другое	15	40
1382	М	02.02.1964	Другое	0	2
1382	М	02.02.1964	Другое	0	146
2003	Ж	04.05.1969	Другое	15	7
2017	Ж	28.01.1961	Другое	0	1
2034	М	11.07.1976	Другое	0	11
2606	М	24.06.1966	Другое	0	3
2629	М	03.07.1979	Другое	0	1
2787	М	16.06.1962	Другое	0	1
3067	Ж	17.07.1973	Другое	15	68
3130	М	13.06.1967	Другое	0	7
3553	Ж	18.09.1964	Другое	15	26
3683	Ж	17.07.1961	Другое	10	13
3697	Ж	12.06.1957	Другое	0	2
3699	М	01.11.1971	Другое	0	36
3625	Ж	30.11.1941	Другое	0	1
3701	Ж	28.06.1965	Другое	10	1
3766	Ж	04.06.1971	Другое	0	1
3823	Ж	11.02.1970	Другое	0	2
3841	Ж	23.10.1945	Другое	0	1
3918	М	11.05.1968	Другое	0	3
4023	М	29.10.1964	Другое	0	2
4044	Ж	01.03.1962	Другое	15	1
4357	Ж	26.11.1972	Другое	0	36
5039	Ж	04.02.1967	Другое	0	5
5133	Ж	08.11.1944	Другое	0	2
5480	М	18.06.1968	Другое	0	3
5667	Ж	23.02.1964	Другое	0	3
5675	Ж	15.08.1976	Другое	0	3
5715	Ж	01.03.1965	Другое	10	12
5823	М	18.10.1957	Другое	0	1
6169	М	09.01.1962	Другое	0	2
6348	Ж	07.07.1956	Другое	15	3
6463	Ж	22.08.1968	Другое	0	75
6562	М	25.05.1962	Другое	10	11
6769	М	10.02.1966	Другое	0	1
6950	Ж	19.03.1962	Другое	0	3
7116	Ж	31.05.1960	Другое	0	1
7188	Ж	24.11.1962	Другое	10	13
7190	Ж	28.01.1961	Другое	10	10
7225	М	20.04.1960	Другое	10	11

Рисунок 14. Исходная выборка данных эксперимента.

Для выделения неочевидных зависимостей (связей) признаков между собой необходимо свести все интересующие нас признаки к набору транзакций. Deductor Studio работает с ненормализованной таблицей транзакций на входе. Для формирования такой таблицы непрерывные численные значения (скидка и количество приходов) и дата рождения было подвергнуто операции квантования и преобразования с тем, чтобы при сведении в единую базу транзакций не перепутать очевидность признаков вида «скидка 15%» и «количество посещений 15», а также для удобства интерпретации результата в отчете. Скриншот ветви обработки и итоговая таблица перед выделением ассоциативных правил представлены на рисунках 15 и 16. Обратим внимание, что вопрос о том, как проводить границы интервалов квантования для возраста и количества посещений, в целом, остается открытым. Так, пациентов по возрастным группам, возможно, с определенной целью разобьет медицинский менеджер, а количество посещений может быть привязано к какой-то программе лояльности и шкалу можно взять оттуда. В нашем случае, по умолчанию, квантование происходило методом, когда интервал нарезается на равные доли.

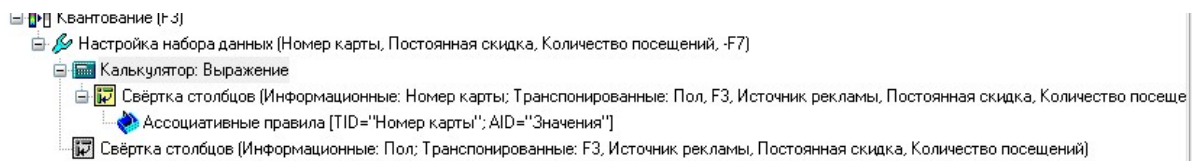


Рисунок 15. Ветка обработки набора данных и выделения ассоциативных правил.

	Номер карты	Заголовки	Значения
	266629	Пол	Ж
	266629	Ф3	от 10.08.1989 до 26.04.1991
	266629	Источник рекламы	Другое
	266629	Количество посещений	4
	266629	Выражение	Скидка 10
	154670	Пол	Ж
	154670	Ф3	от 02.12.1973 до 08.08.1975
	154670	Источник рекламы	Другое
	154670	Количество посещений	8
	154670	Выражение	Скидка 10
	58125	Пол	М

Рисунок 16. Фрагмент таблицы транзакций перед выделением ассоциативных правил.

Эта таблица и была подана на вход алгоритма выделения ассоциативных правил. Идентификатором транзакции выступила карта пациента (идентификатор пациента), а транзакциями – значения после операции свертки таблицы. Ассоциативные правила, для демонстрационной задачи, формировались следующими параметрами (Рисунок 17):

- Поддержка - от 0,5 до 25 %. выделить и совсем слабые и близкие к очевидным зависимостям. Завышать поддержку нет необходимости, дабы не работать с очевидными и без того правилами.
- Достоверность для фильтрации правил в «рабочем» интервале от 20 до 90%.

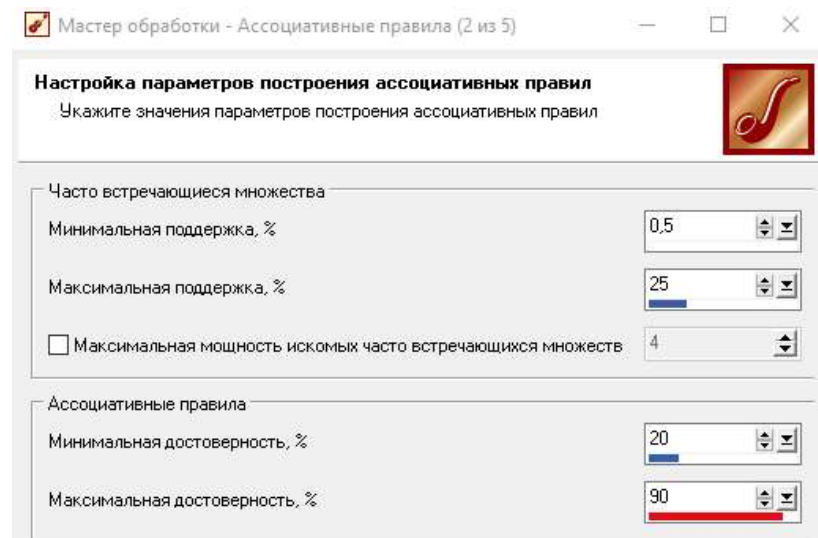


Рисунок 17. Настройки алгоритма выделения ассоциативных правил.

Регулируя вышеуказанными параметрами настройки работы алгоритма, можно выбирать некоторый показатель «неочевидности» закономерностей. Правила, находящиеся выше показателя границы поддержки в 25% уже близки к очевидным, и методы Data Mining для их выделения не требуются.

Алгоритм выделил ассоциативные правила, приведенные в таблице 5. Перед нами выделенные зависимости между признаками, присущими пациентам в группе самых «полезных» по выручке. Заметим, что следствие «рекомендации знакомых» превалирует в виде неявного правила для разных возрастных групп. Также есть зависимости в количестве посещений по разным возрастам, но тут все проще: 15 посещений требуется для некоторых типовых планов лечения, присущим заболеваниям, к которым склонны люди пожилые. Кстати, вот это вот «слабое», но объясненное правило развеяло некоторый скептицизм у заказчика. От него «поверили» и в другие правила, выделенные алгоритмом. Рассмотрим, как бизнес новые знания интерпретировал и применил.

Таблица 5. Выделенные ассоциативные правила

Фильтр: Без фильтрации							
Правил: 31 из 31							
№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	11	Рекомендации знакомых	94	0,75	26,78	1,265
2	2	12	Рекомендации знакомых	85	0,68	24,36	1,151
3	3	13	Рекомендации знакомых	67	0,53	24,72	1,168
4	4	14	Рекомендации знакомых	69	0,55	27,49	1,299
5	5	15	Рекомендации знакомых	334	2,66	20,08	0,949
6	6	до 15.06.1942	15	201	1,60	44,87	3,384
7	7	от 15.06.1942 до 22.09.1948	15	102	0,81	22,82	1,721
8	8	от 22.09.1948 до 20.06.1952	15	91	0,73	20,27	1,529
9	9	18	Рекомендации знакомых	64	0,51	31,68	1,497
10	10	20	Рекомендации знакомых	97	0,77	26,65	1,259
11	11	5	Рекомендации знакомых	232	1,85	21,56	1,019
12	12	от 08.12.1993 до 30.01.1998	5	128	1,02	28,57	3,330
13	13	от 30.01.1998	5	133	1,06	29,69	3,460
14	14	7	Рекомендации знакомых	106	0,85	22,27	1,052
15	15	8	Рекомендации знакомых	94	0,75	20,80	0,982
16	16	9	Рекомендации знакомых	88	0,70	22,11	1,044
17	17	до 15.06.1942	Рекомендации знакомых	136	1,08	30,36	1,434
18	18	от 08.12.1993 до 30.01.1998	Рекомендации знакомых	120	0,96	26,79	1,265
19	19	от 10.03.1972 до 02.12.1973	Рекомендации знакомых	93	0,74	20,71	0,978
20	20	от 10.08.1989 до 26.04.1991	Рекомендации знакомых	98	0,78	21,92	1,036
21	21	от 15.06.1942 до 22.09.1948	Рекомендации знакомых	112	0,89	25,06	1,184
22	22	от 17.02.1987 до 26.04.1988	Рекомендации знакомых	90	0,72	20,04	0,947
23	23	от 19.09.1984 до 03.12.1985	Рекомендации знакомых	94	0,75	21,08	0,996
24	24	от 22.09.1948 до 20.06.1952	Рекомендации знакомых	104	0,83	23,16	1,094
25	25	от 25.04.1968 до 23.05.1970	Рекомендации знакомых	95	0,76	21,16	0,999
26	26	от 26.04.1988 до 10.08.1989	Рекомендации знакомых	107	0,85	23,94	1,131
27	27	от 26.04.1991 до 08.12.1993	Рекомендации знакомых	104	0,83	23,16	1,094
28	28	от 27.03.1955 до 30.08.1957	Рекомендации знакомых	90	0,72	20,04	0,947
29	29	от 30.01.1998	Рекомендации знакомых	163	1,30	36,38	1,719
30	30	от 30.08.1957 до 16.12.1959	Рекомендации знакомых	92	0,73	20,58	0,972
31	31	от 31.12.1963 до 21.02.1966	Рекомендации знакомых	91	0,73	20,36	0,962

### §3.4 Ценность полученного результата

Большой набор правил, которые, при слабой поддержке явили следствием источник рекламы как «рекомендации знакомых» был очень полезен маркетинговому отделу заказчика. Дело в том, что прямыми запросами и статистикой формировалась картина, показывающая эффективность наружной рекламы. Наш же эксперимент показал, что ряд возрастных групп среди самых полезных пациентов пришли в сеть клиник по рекомендации. Это правило в буквальном смысле перевернуло мировоззрение УК. Еще до завершения данной работы в УК произошли изменения в маркетинговой стратегии. Ее направили на мотивацию лояльности имеющейся клиентской базы. Дополнительно, отсутствие в выделенных правилах зависимостей, связанных с ценовой мотивацией, групповыми предложениями и т.д. позволило снять ресурсы с этих направлений. «Сарафан» в качестве привлечения пациентов в коммерческой медицине работает гораздо лучше.

Понятное практическое применение результата и подтверждение слабых зависимостей другими методами продемонстрировали бизнесу всю ценность выделения знаний из собственного опыта. Если даже фактически учебный прогон данных через один алгоритм без сценарной работы, когда можно было работать одновременно в ряде веток с разными параметрами, дал ощутимый результат, то ценность методичной работы по выделению знаний трудно переоценить. И УК, и компания-разработчик приняли решение ставить Data Mining в качестве регулярной активности. В компании-разработчике МИС в стратегической карте продукта появился блок поддержки решений врача, а также блок Data Mining аналитики CRM.

На основании вышеизложенного можно утверждать, что выделенные правила представляют для бизнеса целый набор ценностей: от более зрячей маркетинговой стратегии до полноценного стратегического рыночного преимущества за счет знаний.

## 4 ЗАКЛЮЧЕНИЕ

В ходе выполнения настоящей работы автором были решены следующие принципиальные задачи:

1. Доказана ценность процесса выделения знаний во всей информации, которой располагает предприятие, в нашем случае – медицинская клиника.

2. Продемонстрирована работа алгоритмов Data Mining как эффективного инструмента выделения знаний на полноценном живом примере.

Как мы указывали в параграфе 4.4, даже учебный прогон методов анализа данных предоставил ощутимую финансовую ценность бизнесу, позволив скорректировать маркетинговую стратегию под реальные закономерности, выделенные в базе пациентов. С этой точки зрения цели, поставленные нами во введении, можно считать достигнутыми. Дополнительно, в ходе работы, были получены следующие сопутствующие полезные результаты:

1. Изучены доступные программные решения, на примере Deductor Studio продемонстрирована принципиальная возможность использовать недорогой дружелюбный программный инструментариий специалисту маркетингологии или аналитику информационных баз данных с минимальной подготовкой. Была даже сброшена некоторая магия с терминов Big Data, Data Mining и т.д.

2. Получены рабочие практики по выделению знаний в медицинских информационных системах на объектах коммерческой медицины.

3. Автором получены практические навыки, способные помочь в организации и управлении подразделениями, ведущими разработку ПО в области Data Mining, оказывающими услуги анализа данных, предоставляющими консультирование в вопросах использования методов Data Mining.

Полученные в ходе работы данные, и их применение показали, что неочевидные знания на текущем этапе, в коммерческой клинике среднего масштаба являются осязаемой ценностью, а их поиск и выделение, в этом случае, является значимым конкурентным преимуществом на конкурентном рынке. Организация процесса регулярного выделения знаний с помощью построения самообучаемых информационных систем, как мы показали, является технически реализуемой задачей. Эти системы и знания, который они находят, а после дополняют и улучшают их качество на итеративной основе, способный помочь принятию более качественных, оперативных решений организационного уровня, в маркетинге, вспомогательных процессах, в самой клинической деятельности. Анализ данных способен постоянно, проактивно снабжать специалистов и управленцев полем вариантов решений, сценарных прогнозов высокого качества, диагностировать отклонения

на ранних стадиях, сразу определяя природу таких отклонений. При этом и методы анализа данных и системы на их основе по своей природе будут и себя, и выдаваемые данные, и получаемые знания постоянно улучшать.

В ходе жесткой эволюции в бизнесе к новой среде приспособятся те и только те, кто осознает, что знания становятся дороже многих других ресурсов. Поначалу работа со знаниями будет принципиальным конкурентным преимуществом, а позже станет необходимым условием существования на рынке во всех сферах. Средства Data Mining в скором времени сравняются по применимости в ПО автоматизации учета небольших компаний и будут покупаться в составе программных наборов вида «Все ПО для небольшой фирмы».

Автор оптимистично смотрит в будущее, в котором знания обретут такую ценность и значимость, а исходных данных будет накоплено такое количество и в таком качестве, что фундаментальные проблемы человечества, вроде препаратов против вируса СПИД или генетических методов борьбы с раком будут решаться не столько в гениальных гипотезах или смелых экспериментах, а в поле правильного применения методов анализа данных.

Вместе со стремительной наступающими децентрализованными приложениями и базами данных анализ данных будет идти глобально, постоянно и открыто, и также глобальны и открыты будут его результаты. Обучаемые информационные системы, с открытыми принципами функционирования, заменят ряд государственных и межгосударственных институтов.

Когда-то в СССР существовала вера, что все производство и потребление в масштабах огромной страны возможно спланировать. С исчезновением СССР было принято верить в абсолютное заблуждение такого подхода. Теперь же мы видим, как средства анализа данных, помноженные на невероятные средства сбора всей информации, вновь приближают времена, когда на рынок будут адаптивно выводиться продукты, меняющие свои свойства каждую партию, каждый экземпляр процесса. Скорости обратной связи систем работы с покупателями будут нарастать, продолжат прибавлять объемы и глубина захвата сохраняемой и непрерывно анализируемой информации о деятельности компаний. В таком мире качество планирования выпуска продукции, даже в условиях нарастающей новизны и сложности, и с учетом рыночной конкуренции, будет похоже на социалистическое планирование.

Не за горами то время, когда условный кладовщик на складе не будет смотреть в отчетах складской информационной системы цифры товарных остатков и их распределение по площадям. Информационная система будущего будет способна сразу выдавать прогнозы поступления и отгрузки товарных позиций, рекомендовать направить еще не

поступивший товара на оптимальное складское место. Кажется, еще живые люди будут грузить эти коробки, а вот куда и как их грузить, будет решать информационная система, опираясь на свои знания и знания тех самых грузчиков. Квалифицированные доктора нового времени будут наблюдать не пациентов, а управлять отклонениями в обучаемых информационных системах поддержки принятия решений врачом.

Знания в мире будущего станут новой нефтью, новым золотом и принципиально новым оружием обеспечения стратегических интересов государств.



## 5 СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

- 1) Некрасов С.И., Некрасова Н.А. Философия науки и техники: тематический словарь справочник. Учебное пособие. – Орёл: ОГУ, 2010. – 289 с.
- 2) Zipf G.K. Human behavior and the principle of least effort. Cambridge: Univer. Press, 1949.
- 3) Doran G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. Management Review, Volume 70, Issue 11(AMA FORUM).
- 4) Савельев С. В. Нищета мозга. — Москва: Веди, 2014.. — 192 с.
- 5) David Weinberger The Problem with the Data-Information-Knowledge-Wisdom Hierarchy - Harward Business Review. — 2010.
- 6) Практическая характерология с элементами прогнозирования и управления поведением. Методика "семь радикалов». - Москва: Феникс, 2006. – 49 с.
- 7) Пачоли Л. Трактат о счетах и записях / Под ред. Я. В. Соколова. — Москва: Финансы и статистика, 1994. — 320 с
- 8) Turing, Alan (October 1950), «Computing Machinery and Intelligence», Mind LIX (236): 433—460, doi: 10.1093/mind/LIX.236.433NYT
- 9) John Markof, «Brainy Robots Start Stepping Into Daily Life», «The New York Times» Thursday, July 18, 2006
- 10) Ленин В. И. Полное собрание сочинений в 55 томах. Том 18. Москва: Издательство политической литературы, 1958. – 412 с.
- 11) Информационный менеджмент. Фантом, обретающий плоть. Executive.ru. Электронная статья [Электронный ресурс]. – 2011. URL: <http://www.executive.ru/community/articles/1490449/>
- 12) Рыжов А.П. Advanced information technologies RU 2016 1 Презентация курса «Аналитические системы». - Москва. 2016
- 13) Zadeh L. A. Fuzzy sets // Information and Control. — 1965. — Т. 8, № 3. — P. 338-353.
- 14) Кофман, А. Введение в теорию нечетких множеств / А. Кофман. – Москва: Радио и связь, 1983. – 432с.
- 15) Набебин А.А. Логика и пролог в дискретной математике». – Москва: Издательство МЭИ. 1996. - 432 с.
- 16) Мак-Коллок У.С., Питтс В. Логическое исчисление идей, относящихся к нервной активности. Пер. с англ - . Москва: Издательство иностранной литературы. - 1956.
- 17) Розенблат Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга. Москва:Мир. – 1965.

- 18) М. Минский, С. Пейперт. Персептроны. — Москва: Мир. - 1971.
- 19) Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного и сложения // ДАН СССР. 1957. Т. 114, № 5. С. 953–956.
- 20) C. Darwin n the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life
- 21) Гайсинович А.Е. Зарождение и развитие генетики. — М.: Наука, 1988. — 424 с.
- 22) J. H. Holland. Adaptation in Natural and Artificial Systems. The MIT Press, reprint edition, 1992.
- 23) Data Mining — добыча данных. Basegroup.ru. Электронная статья [Электронный ресурс]. – URL: <https://basegroup.ru/community/articles/data-mining>
- 24) Apriori — масштабируемый алгоритм поиска ассоциативных правил. Basegroup.ru Электронная статья [Электронный ресурс]. – URL: <https://basegroup.ru/community/articles/apriori>
- 25) Государство переходит на электронный документооборот. Cnews.ru. Электронная статья [Электронный ресурс]. – 2012 URL: <http://www.cnews.ru/reviews/free/gov2012/articles/articles2.shtml>
- 26) Questions about Bitcoin. Bitcontalk. Электронная статья [Электронный ресурс]. – 2009. URL: <https://bitcointalk.org/index.php?topic=13.msg46#msg46>
- 27) Herbert Hellerman. Digital Computer System Principles. - N.Y.: McGraw-Hill, 1967, 424 с
- 28) Официальная страница проект Ethereum. [Электронный ресурс]. URL: <https://blog.ethereum.org/>
- 29) Биография В.И. Дикуля. Электронная статья [Электронный ресурс]. URL: <https://www.dikul.net/dikul/>
- 30) Взгляд на перспективы развития рынка частных медицинских услуг в РФ в 2017–2019 гг. Результаты исследования КПМГ. Kpmg.ru Электронная статья [Электронный ресурс]. – 2016. URL <https://assets.kpmg.com/content/dam/kpmg/ru/pdf/2017/03/ru-ru-research-on-development-of-the-private-medical-services-market-v1.pdf>
- 31) Международная классификация болезней 10-го пересмотра (МКБ-10). [Электронный ресурс]. URL: <http://mkb-10.com>
- 32) Стандарт GCP — Надлежащая клиническая практика. Электронная статья [Электронный ресурс]. URL: <https://gmpnews.ru/terminologiya/standart-gcp/>

33) Demand for Data Analytics in Healthcare. Электронная статья [Электронный ресурс]. URL: <https://www.usfhealthonline.com/resources/healthcare/demand-for-data-analytics-in-healthcare/>

34) Чугрюмова И.Г. Медицинская система принятия решений с использованием нейронной сети. – Харьков: Медицинская литература. – 2004.

35) Фармацевтический рынок России. Итоги 2015 года. DSM Group. Электронная статья [Электронный ресурс]. – 2016. URL: [http://www.dsm.ru/docs/analitics/Annual\\_report\\_2015\\_DSM\\_web.pdf](http://www.dsm.ru/docs/analitics/Annual_report_2015_DSM_web.pdf)

36) <https://rapidminer.com/>

37) <https://www.r-project.org>

38) <https://orange.biolab.si/>

39) [http://statsoft.ru/products/STATISTICA\\_Data\\_Miner/data-mining-more.php](http://statsoft.ru/products/STATISTICA_Data_Miner/data-mining-more.php)

40) Double-Digit Growth Forecast for the Worldwide Big Data and Business Analytics Market Through 2020 Led by Banking and Manufacturing Investments, According to IDC. Электронная статья [Электронный ресурс]. – 2016. URL: <https://www.idc.com/getdoc.jsp?containerId=prUS41826116>

41) <https://yandexdatafactory.com/ru/>

42) Manufacturer's Guide for Applying Machine Learning. YandexFactory. Электронная статья [Электронный ресурс]. – URL: <https://yandexdatafactory.com/white-papers/machine-learning-manufacturing-practical-guide-applying-machine-learning-business-operations/>