МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
«Высшая школа экономики и управления»
Кафедра «Информационные технологии в экономике»

ПРОЕКТ ПРОВЕРЕН
Рецензент, генеральный директор ООО «Софт-Фьюче»,
_____ (А.В. Берестов)
«____» _____2019 г.

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий кафедрой, д.т.н., с.н.с.
_____ (Б.М. Суховилов)
«____» _____ 2019 г.

Сейсмологическая статистическая аналитическая система на основе алгоритмов машинного обучения

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ
ЮУрГУ–38.04.05.2019.893.ПЗ.ВКР

Руководитель проекта, доцент, к.т.н.
_____ (О.С. Буслаева)
«____» _____ 2019 г.

Автор проекта,
студент группы ЭУ-244
_____(А.Е. Станогин)
«____» _____ 2019 г.

Нормоконтролер, доцент, к.т.н.
_____ (Е.В. Бунова)
«____» _____ 2019 г.

Челябинск 2019

АННОТАЦИЯ

Станогин А.Е. Сейсмологическая статистическая аналитическая система на основе алгоритмов машинного обучения. Челябинск ЮУрГУ, ЭУ – 244; 2019. – 61 с., 4 ил., 17 табл., библиогр. список – 33 наим.

Аналитические системы на основе Data Mining и машинного обучения применяются в сейсмологии более 10 лет [2]. Они позволяют извлекать из сейсмологических показателей скрытые закономерности на основе которых можно прогнозировать землетрясения. Выполнение такой работы человеком трудоемко, требует привлечения специалистов высокой квалификации и не всегда возможно[4]. Данное утверждение объясняет актуальность дипломной работы.

Целью диссертационной работы является разработка информационной системы комплексной оценки параметров геодинамических событий горнопромышленных регионов.

В диссертации поставлены и решены следующие задачи:

1. Сформулированы функциональные требования для информационной системы обработки геодинамических событий горнопромышленного региона.

2. Разработана информационная модель регламентированного сбора из распределенных источников геодинамических данных различных форматов, включающая их последующий анализ на основе оригинальных методов обработки пространственной информации.

3. Разработана комплексная модель обработки данных, обеспечивающая проведение оценки геодинамической обстановки в различных регионах Челябинской области.

Результатами работы являются: проект по разработке математического и программного обеспечений сейсмологической аналитической системы.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Сейсмологические события на протяжении всего времени существования человечества приносили людям ущерб. В последние десятилетия ситуация дополнительно обостряется промышленным освоением сейсмоактивных районов и разработкой месторождений глубоко залегающих полезных ископаемых, добыча которых часто сопровождается мощными проявлениями горного давления, необходимостью оценки текущего состояния массива горных пород, вмещающего инженерные сооружения. Таким образом, в настоящее время для принятия решений, направленных на обеспечение безопасности населения, снижение ущерба и проведение превентивных мероприятий, требуется осуществлять прогноз как естественной природной, так и техногенной сейсмичности. Исходной информацией для прогноза, как правило, являются сейсмические каталоги, содержащие сведения о слабых событиях (их месте, времени и силе), предваряющих сильные.

Основные подходы к решению проблемы прогноза сейсмических событий появились только в XX веке. При этом в задачи прогноза входят определение силы ожидаемого сейсмического события, места и времени его возникновения.

Наибольшие успехи в настоящее время достигнуты в области долгосрочного прогнозирования (оценке долгосрочного сейсмического режима), когда оценивается средняя опасность возникновения сильных событий за длительные промежуток времени в протяженных пространственных областях. Они связаны с именами В.И. Бунэ, М.В. Гзовского, А.А. Гусева, Б. Гутенберга, В.И. Кейлис-Борока, СВ. Медведева, И.Л. Нерсесова, Ю.В. Ризниченко, Ч.Ф. Рихтера, М.А. Садовского, В.И. Уломова, С.А. Федотова и многих других исследователей. В настоящее время имеются достижения в развитии подходов к среднесрочному прогнозу, когда достаточно определенно говорится о наиболее опасных пространственных областях и указывается относительно продолжительный период, когда следует ожидать возникновение сильных события. Особый вклад в это направление внесли А.Д. Завья-

лов, В.С. Куксенко, В.А. Мансуров, Г.А. Соболев, Т.Л. Челидзе и другие. Для землетрясений проблема краткосрочного прогноза до сих пор не может считаться решенной, однако в последние десятилетия активно развивается направление краткосрочного прогнозирования горных ударов.

Все это свидетельствует об актуальности проведения исследований по дальнейшему развитию подходов, методов и алгоритмов прогноза сильных сейсмических событий.

Объектом исследования является сейсмологическая аналитическая система. Предметом исследования – методы интеллектуального анализа данных для решения задач в области сейсмологии.

Целью работы является разработка проекта реализации программного и математического обеспечения сейсмологической аналитической системы.

Задачи магистерской работы:

1) анализ информационных сейсмологических систем;

2) классификация задач и обзор научных работ, посвящённых анализу данных в сфере сейсмологии;

3) постановка задачи для проведения исследования и разработки математического и программного обеспечений;

4) обзор существующих методов интеллектуального анализа данных;

5) анализ научных работ по использованию механизмов машинного обучения в сейсмологии и описание примеров использования механизмов;

6) разработка математического и программного обеспечений сейсмологической аналитической системы;

7) разработка плана коммерциализации проекта.

ГЛОССАРИЙ

В таблице ниже (таблица 1) представлен перечень терминов и сокращений, используемых в работе.

Таблица 1 – Глоссарий

Термин, сокращение	Определение
Б	Бинарный
Д	Дата
К	Количественный
Математическое обеспечение	Совокупность математических методов, моделей, алгоритмов обработки информации, используемых при решении задач в информационной системе (функциональных и автоматизации проектирования информационных систем)
МД	Мало данных. Меньше порога в 1500 записей. Порог выбран эмпирически на основании просмотра исходных данных.
НПИ	Нет полезной информации для эксперимента.
ПД	Персональные данные. Нет доступа для проведения анализа. Кроме того, не несут полезной аналитической информации.
С	Строка
ХД	Хранилище данных
ЦПТ	Чисто с плавающей точкой
ЦЧ	Целое число

ГЛАВА 1 СЕЙСМОЛОГИЧЕСКИЕ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ СИСТЕМЫ

1.1 Анализ информационных сейсмологических аналитических систем

Рассмотрим состояние и направление развития некоторых современных сейсмических сетей, осуществляющих мониторинг землетрясений в глобальном, национальном и региональном масштабах. Это полезно с точки зрения выявления перспективных направлений развития систем мониторинга сейсмичности. Глобальная сейсмическая сеть (GSN) представлена цифровыми сейсмическими станциями, имеющими широкую полосу и большой динамический диапазон регистрации сигналов. Руководство GSN [3] осуществляет Корпорация научно-исследовательских институтов по сейсмологии – IRIS [4]. Глобальная сейсмической сеть разработана с целью обеспечения мирового сейсмологического сообщества данными наблюдений. В сети GSN под управлением Альбукерской сейсмологической лаборатории (USGS/ASL) функционирует более 80 станций. Вторым оператором сети – Калифорнийским университетом в Сан-Диего (IRIS/IDA) установлено около 80 станций. Кроме того, в состав GSN входит ряд станций, принадлежащих университетским сетям и другим членам корпорации IRIS. Общее количество таких станций 15. Таким образом, 13 в сети GSN работает около 150 станций, располагающихся более чем в 80 странах на всех континентах. Благодаря этому резко улучшилось равномерность покрытия станциями земного шара и, как следствие, значительно возросло количество данных [5].

Концепция открытого доступа к станционным данным является основным принципом формирования GSN. С большинства станций данные наблюдений поступают через Интернет, орбитальные спутниковые каналы или по выделенным телефонным линиям. GSN работает в тесном взаимодействии с Национальной сейсмической сетью США (USNSN), которая эксплуатируется Геологической служ-

бой США. Станции GSN и USNSN подобны, а сбор данных, их анализ и распространение координируются USGS и IRIS. Европейско-Средиземноморский сейсмологический центр (EMSC) создан с целью быстрого (близкого к реальному времени) определения эпицентров потенциально разрушительных землетрясений Европейско-Средиземноморского региона. В настоящий момент членами EMSC [6] являются 6 ключевых членов, представляющих 5 стран, 34 активных члена, представляющих 25 стран, и 3 члена по праву – международные организации ESC, ORFEUS и ISC. Для срочного оповещения о землетрясениях в Европейско-Средиземноморском регионе установлен уровень магнитуд 5.0 и выше. Центр ведет работу непрерывно и круглосуточно. Если в обычном режиме Центр получает данные от 41-ой сети (более пятисот станций), то в службе срочного оповещения принимают участие сети, которые дают более качественные данные в оперативном режиме. В настоящее время приблизительно 30 сейсмологических сетей передают данные в EMSC в режиме близком к реальному времени. Девятнадцать сетей расположены в европейских странах. В работе EMSC принимает участие Геофизическая служба РАН, которая передает станционные данные и результаты сводной обработки [5]. Данные сетей передаются в виде сообщений по электронной почте и автоматически анализируются в EMSC. Для тех данных, которые вызваны землетрясением, превышающим уровень пороговой магнитуды, центр производит локализацию и определяет магнитуду.

Для землетрясений, которые происходят в пограничных областях стран Европейско-Средиземноморского региона, точность и надежность данных центра выше, чем любой сети из стран этого региона. Срочное 14 донесение EMSC немедленно рассылается в правительственные органы Европейских стран, международные организации. База данных EMSC в настоящее время еще находится в стадии разработки. Цель состоит в том, чтобы данные, собранные и хранимые в EMSC, стали легко доступными научным учреждениям и специалистам. Данные будут включать следующую информацию: все полученные EMSC сообщения с данными и все срочные донесения, посланные EMSC; все бюллетени, полученные EMSC; и подготовленный в EMSC Европейский бюллетень. В заключение следует отметить,

что EMSC своей собственной сейсмической сети не имеет. Работа центра базируется на данных наблюдательных сетей и станций, принадлежащих различным странам. Данные предоставляются на основе членства стран в EMSC или на основе взаимных интересов организаций по обмену данными. Национальный центр информации о землетрясениях (NEIC) является частью Геологической службы США [7].

Главной задачей центра является максимально быстрое и, насколько возможно, более точное определение основных параметров сильных и разрушительных землетрясений, происходящих на территории США и во всем мире. NEIC немедленно распространяет эту информацию заинтересованным национальным и международным организациям, ученым, специалистам и широкой публике. Следующей задачей центра является создание базы сейсмических данных и обеспечение доступа к ней широкого круга исследователей. Источниками для создания базы являются данные современных цифровых национальных и глобальных сейсмических сетей, отдельных обсерваторий, станций и институтов более чем из 80-ти стран мира, получаемые в соответствии с международными соглашениями. NEIC является национальным центром не только текущих данных, но и постоянно пополняемого архива информации о землетрясениях. Третьей задачей NEIC является развитие исследований, направленных на улучшение определения местоположения землетрясений и понимание механизма землетрясений. Такие работы Центра направлены, прежде всего, на уменьшение сейсмической опасности. Они возможны благодаря тесному и длительному международному сотрудничеству NEIC с научно-исследовательскими институтами, национальными и региональными сетями. 15 Национальный центр непрерывно и круглосуточно ведет работу по точному и быстрому определению местоположения и энергии значительных землетрясений в США и во всем мире. Эта информация сообщается правительственным федеральным агентствам и агентствам штатов, ответственным за реакцию на чрезвычайные ситуации, национальным и международным средствам массовой информации, научным группам и частным гражданам, запрашивающим такую информацию. Когда разрушительное землетрясение происходит за рубежом, информация о землетрясении передается персоналу американских посольств и консульств в странах,

где произошло землетрясение, а также гуманитарному департаменту Организации Объединенных Наций. NEIC дает срочные донесения о землетрясениях с магнитудой 4.5 и более на территории США и с магнитудой 6.0 – 6.5 и более (а также о тех, которые вызвали разрушения) по всему миру. В настоящее время персонал NEIC ежегодно определяет параметры и публикует информацию более чем о 20 тыс. землетрясений [5]. Геофизическая служба РАН.

Сейсмическая сеть, которая ответственна за сбор сейсмических данных на территории России, имеет иерархическую трехуровневую структуру. В нее входят телесейсмическая сеть (OBN) и 10 региональных сейсмических сетей. В состав некоторых из них в свою очередь входят локальные сети. В общей сложности в единой сети сейчас работают более 250 сейсмических станций и 10 центров сбора и обработки данных [8]. Организационно объединяет и координирует работу всех сетей Геофизическая служба Российской академии наук, которая обеспечивает производство наблюдений, текущую обработку данных, издание сейсмологических каталогов и бюллетеней, предоставление данных для исследований в области наук о Земле. Филиалы Геофизической службы обеспечивают сейсмический мониторинг территорий отдельных регионов. Геофизическая служба взаимодействует с международными и национальными сейсмологическими центрами с целью обмена данными и интеграции в мировую систему сейсмических наблюдений. Геофизическая служба наряду с научными исследованиями в области сейсмического мониторинга и развитием новых средств и методов производства наблюдений обеспечивает оперативное оповещение центральных и местных органов власти, а также других ведомств и организаций о землетрясениях и их возможных последствиях. 16 В состав телесейсмической сети (OBN), центр которой находится в г. Обнинске, входит около 40 станций. Все станции имеют широкополосные каналы регистрации. Подавляющая часть из них оснащена цифровым оборудованием, но есть еще несколько станций, которые используют короткопериодные и длиннопериодные каналы с записью на фотобумагу. Цифровая регистрация на 12 станциях производится оборудованием, предоставленным корпорацией сейсмологических институ-

тов США (IRIS). Его характеристики аналогичны характеристикам станций Глобальной сейсмической сети GSN. Реализована передача данных по каналам связи, в том числе в режиме близком к реальному времени. Центр телесейсмической сети регулярно получает в таком режиме данные с более чем 40 отечественных и зарубежных станций, располагающихся на разных континентах. Кроме того, центр имеет доступ к ряду зарубежных баз данных и с задержкой до нескольких месяцев получает данные всех станций Глобальной сейсмической сети [5]. Программное обеспечение центра создано как результат многолетних усилий с использованием собственных разработок и достижений сейсмологических центров США, Австралии и других стран. Оно позволяет реализовать практически все современные методы обработки данных, включая производство сбора данных в различных режимах, автоматическое детектирование и ассоциацию фаз, определение параметров событий в интерактивном режиме, формирование бюллетеня сейсмических событий [9].

Обнаружение сейсмических событий и оповещение о тревоге сейсмической опасности является главной целью служб срочных донесений (ССД), которые непрерывно и круглосуточно функционируют в составе Геофизической службы РАН и её филиалов, расположенных в сейсмоактивных регионах. Службы постоянно ведут анализ сейсмической обстановки. Хотя большая часть территории России находится в стабильном континентальном регионе Земли, около 20% ее площади расположено в сейсмоопасных зонах с интенсивностью землетрясений 7 баллов и выше. Наибольшую опасность несут землетрясения в зонах активных тектонических процессов на территориях Камчатки, Сахалина, Прибайкалья, Северного Кавказа, Алтая. Донесения о произошедших на территории РФ и мира или в зонах ответственности регионов о сильных разрушительных землетрясениях и их последствиях оперативно передаются в центральные и местные органы исполнительной власти, заинтересованным ведомствами 17 и организациям, в центральные и территориальные органы МЧС. Донесения содержат информацию о землетрясениях и прогнозируемых возможных их последствиях. В Геофизической службе РАН и не-

которых ее филиалах широко используется автоматизированное рабочее место сейсмолога «WSG». «WSG» представляет собой программный комплекс, включающий основной программный модуль «WSG» («Windows Seismic Grafer») и набор вспомогательных сервисных утилит [10, 11]. Основные вычислительные процедуры, предназначенные для обработки сейсмических сигналов и получения оценок параметров гипоцентров сейсмических событий как по записям одной станции, так и по группе станций, сосредоточены в программном модуле «WSG». Практика работы с этим программным модулем показала, что он может быть использован в качестве рабочего места сейсмолога как на отдельных сейсмических станциях, так и в обрабатывающих центрах. Региональные сейсмические сети ГС РАН имеют разный уровень оснащенности сейсмическим оборудованием, средствами связи. Используемая сейсмическая аппаратура и программные средства разрознены. Отсутствует единый протокол сбора, хранения и обработки сейсмических данных.

В литературе не встречается однозначно определённое понятие сейсмологической информационно-аналитической системы.

Прежде всего определим понятие и классификацию сейсмологической информационной системы (СИС). Различные определения МИС и классификации МИС приведены в работах [3] и [4]. Например, в [4] даётся следующее определение: «Совокупность информационных, организационных, программных и технических средств, предназначенных для автоматизации сейсмологических процессов и(или) организаций».

С.А. Гаспарян [3] определяет СИС, как одну из форм организации сейсмологической деятельности, позволяющая персоналу при соответствующей технологической поддержке использовать комплекс математических и технических средств, обеспечивающих сбор, хранение, обработку, анализ и выдачу сейсмологической информации».

Отсюда можно сделать вывод, что СИС является автоматизированной системой, обеспечивающей сбор, хранение, обработку, анализ и выдачу сейсмологической информации [4].

С.А. Гаспарян [3] определяет следующую классификацию СИС:

1. Технологические информационные сейсмологические системы (ТИСС);

2. Банки информации сейсмологических служб (БИСС);

3. Статистические информационные сейсмологические системы;

4. Научно-исследовательские информационные сейсмологические системы;

5. Обучающие (образовательные) информационные сейсмологические системы.

Для технологических информационных систем объектом описания является сейсмологическая активность (землетрясение), пользователем - сейсмолог, информация интегрируется на уровне одного землетрясения.

Банки информации сейсмологических служб обеспечивают информационную поддержку отношений совокупность землетрясений – сейсмологи. Основанием для деления банков информации на виды является широта охвата обслуживаемого населения.

Статистические информационные сейсмологические системы обеспечивают информационную поддержку отношений популяция (в смысле населения обслуживаемого региона) – органы, управляющие системой сейсмологического обслуживания. Деление статистических информационных систем на виды основано на различии объектов описания, представленных в статистических отчетах ЛПУ и территориальных органов управления здравоохранением.

Научно-исследовательские информационные сейсмологические системы позволяют рассматривать объекты и документы науки. Разделение на виды основано на различиях объектов описания.

Обучающие информационные сейсмологические системы обеспечивают информационную поддержку отношений обучаемые – преподаватели.

Образовательные информационные системы разделяются на виды в соответствии с педагогическими принципами оценки уровня освоения знаний учащимся.

Таким образом, на основании общего понятия информационно-аналитической системы [5] и понятия сейсмологической информационной системы, приведённого выше, можно определить, что сейсмологические информационно-анали-

тические системы (СИАС) – это комплекс аппаратных, программных средств, информационных ресурсов, методик, которые используются для обеспечения автоматизации аналитических работ для решения задачи сферы сейсмологии. Также к классу сейсмологических информационно-аналитических систем можно определить статистические информационные системы и научно-исследовательские информационные сейсмологические системы.

Фундаментом любой, в том числе и сейсмологической, информационно аналитической системы является – аналитическое программное обеспечение. Для определения понятия «аналитическое программное обеспечение» в качестве исходной информации можно использовать доклады известных информационных агентств (IDC, Gartner), а также некоторые материалы российских авторов. В мировой практике принято использовать термин Business Intelligence (BI), что на русский язык может быть переведено как деловой интеллект. Это понятие объединяет различные средства и технологии анализа и обработки данных масштаба предприятия. Наиболее подробное описание систем, относящихся к категории BI, содержится в аналитическом докладе Gartner «Infrastructure and Applications Worldwide Software Market Definitions. 2002». В этом документе содержится традиционная классификация систем класса BI, построенная, главным образом, с точки зрения программной архитектуры. Далее рассмотрены основные элементы классификации Gartner, даны определения, отражающие не только техническую, но и экономическую сущность каждого сегмента классификации.

Итак, Gartner выделяет следующие сегменты рынка BI:

– средства построения хранилищ и витрин данных (data warehouse);

– инструменты оперативной аналитической обработки (On-Line Analytical Processing, OLAP) и прочие средства многомерного анализа;

– информационно-аналитические системы (Enterprise Information Systems, EIS) и системы поддержки и принятия решений (Decision Support Systems, DSS);

– средства интеллектуальной добычи данных (data mining);

– инструменты конечного пользователя для выполнения запросов и построения отчетов (query and reporting tools).

1.2.1 Хранилища данных

Один из авторитетных специалистов в этой области – Б.Инмон (Bill Inmon) определяет хранилища данных (ХД) как «предметно-ориентированные, интегрированные, стабильные, поддерживающие хронологию наборы данных, организованные для целей поддержки управления, призванные выступать в роли «единого и единственного источника истины», обеспечивающего менеджеров и аналитиков достоверной информацией, необходимой для оперативного анализа и принятия решений» [6]. Ценность ХД для экономистов заключается в следующем: ХД – это некая база данных масштаба предприятия, которая содержит определенную аналитическую информацию, обеспечивает ее оперативное представление в удобном для пользователя виде и обладает структурой, учитывающей отраслевую специфику деятельности организации. Типичные представители программных продуктов этой категории: SAP Business Warehouse (SAP), Informatica.

1.2.2 OLAP-средства

Под термином OLAP, как правило, понимают системы аналитической обработки данных в режиме реального времени. OLAP-системы обеспечивают решение многих аналитических задач: анализ ключевых показателей деятельности, маркетинговый и финансово-экономический анализ, анализ сценариев, моделирование, прогнозирование и т.д. Такие системы могут работать со всеми необходимыми данными, независимо от особенностей информационной инфраструктуры компании. С точки зрения пользователя, отличие OLAP-системы от хранилища данных заключается в предметной (а не технической) структурированности информации, при этом пользователю предоставляется возможность оперировать привычными экономическими категориями и понятиями. К типичным представителям программных продуктов этого класса относятся: Hyperion Essbase (Hyperion Solutions

Corporation), Oracle OLAP (Oracle), MS Analysis Services (Microsoft), Business Objects (Business Objects), Cognos PowerPlay (Cognos), MicroStrategy.

1.2.3 Информационно-аналитические системы

Этот класс аналитических систем включает множество разнообразных продуктов, основная задача которых – предоставить конечные решения для менеджеров-аналитиков. Например, для банковской сферы реализованы методики дистанционного анализа, внутреннего и внешнего анализа, анализа прибыльности, рейтинговой оценки надежности банка (CAMEL), расчет рейтинга надежности банка (на основе методики В.С.Кромонова), расчет лимита межбанковского кредитования (на основе методики КБ «Европейский Трастовый Банк»), GAP-анализ.

1.2.4 Средства интеллектуальной добычи данных

Средства интеллектуальной добычи данных (data mining). Программные продукты, относящиеся к этой категории, обеспечивают поиск полезных данных в огромных массивах информации. Иными словами, такие программные продукты позволяют аналитику получить качественно новую информацию, не содержащуюся в источнике данных явным образом. Здесь используются популярные методы математического анализа данных: фильтрация, дерево решений, ассоциативные правила, генетические алгоритмы, нейронные сети, статистический анализ.

В качестве примера вывода, полученного с помощью средств data mining, приведем результат анализа базы данных биллинговой системы оператора сотовой связи: «в предыдущем месяце наибольшее число продаж самого популярного тарифного плана приходится на клиентов в возрасте от 18 до 27 лет во временном интервале с 10 до 14 часов». Эта информация не хранится в базе данных явно, однако такие результаты могут быть получены после проведения процедуры анализа,

при помощи одного из вышеперечисленных методов или их комбинации.

Таким образом, системы data mining помогают аналитику сформировать качественные выводы, которые обычный человек не в состоянии получить стандартными методами исследования данных (во всяком случае, не так быстро, как программа). Как правило, функции интеллектуального извлечения данных встраиваются в OLAP-системы. Типичные представители фирм-разработчиков: Hyperion Essbase (Hyperion Solutions Corporation), Oracle Data Mining (Oracle), SAS (SAS Institute).

1.2.5 Инструменты конечного пользователя

Инструменты конечного пользователя для выполнения запросов и построения отчетов (query and reporting tools). Такие системы обеспечивают функции построения запросов к информационно-аналитическим системам (в пользовательских терминах), интеграцию данных из нескольких источников, просмотр данных с возможностью детализации и обобщения, построение полноценных отчетов и их печать. Они предназначены для пользователей, обладающих «продвинутыми» техническими навыками. При этом профессиональных знаний в области информационных технологий не требуется, тем не менее, для экономистов такие средства не всегда бывают удобны. Как правило, модули, содержащие функции Query & Reporting, входят в состав многих OLAP-систем, но есть и отдельные программные продукты этого класса. Таким образом, четко провести грань между OLAP и Query & Reporting невозможно. Характерный пример – приложение Hyperion Essbase, которое аналитики относят к обоим классам.

В заключение подведем некоторые итоги классификации.

Во-первых, очевидно, что отнести тот или иной программный продукт к какому-то одному классу не всегда возможно, поскольку многие системы позволяют решать аналитические задачи нескольких категорий. К числу «многофункциональных» можно отнести системы таких мировых производителей, как Hyperion

Solutions Corp., Cognos, Business Objects, Microsoft. Эти компании являются лидерами мирового рынка систем делового интеллекта, их продукты также активно продаются в России. Типичным примером универсальной системы может служить Hyperion Essbase – аналитическая платформа класса OLAP, предназначенная для решения довольно широкого круга задач. Будучи OLAP-системой, Hyperion Essbase также решает часть задач, относящихся к информационно-аналитическим системам, средствам интеллектуального извлечения данных, а также обеспечивает функции программных средств построения запросов и отчетов. Кроме того, в некоторых случаях Hyperion Essbase может использоваться в качестве хранилища данных, а также в качестве аналитической «прослойки» в крупных компаниях, где данные распределены по многим информационным источникам.

Во-вторых, в настоящее время наибольшим спросом на рынке пользуются хранилища данных, OLAP-средства и системы data mining. Они обладают богатыми аналитическими возможностями, в том числе в части финансовых и статистических функций, которые постоянно развиваются и улучшаются. При этом они позволяют хранить и обрабатывать большие объемы информации.

В-третьих, при выборе аналитической системы необходимо учитывать степень простоты освоения и эксплуатации программы пользователями-экономистами, не владеющими техническими знаниями в профессиональном объеме. Иначе говоря, программный продукт должен быть настраиваемым под конечных пользователей и требовать при этом минимальной поддержки со стороны технических специалистов. Например, упомянутый выше Hyperion Essbase позволяет обеспечить всю рутинную работу, оставив аналитику только ту часть, которая касается собственно анализа и представления данных.

В-четвертых, при выборе аналитической системы также следует учитывать ее приспособленность к решению конкретных, интересующих конечного пользователя задач. В лучшем случае это реализуется в виде готовых отраслевых решений в конкретной предметной области.

1.2 Классификация задач и обзор научных работ, посвящённых анализу данных в сфере сейсмологии

Современную сейсмологию невозможно представить без использования точных и надёжных методов анализа и прогнозирования [7]. На основании научных работ в данной сфере можно определить следующие основные классы задач анализа данных в сфере сейсмологии:

1. Задачи сейсмологической диагностики;
2. Задачи анализа изображений;
3. Задачи классификации и кластеризации;
4. Задачи предсказания (например, предсказание катаклизмов).

Далее приведены примеры научных работ, посвящённых анализу данных в сфере сейсмологии по выделенным задачам, а также дано краткое описание каждой из задач.

1. Задачи сейсмологической диагностики

В последние годы благодаря применению современных методов интеллектуального анализа данных Data Mining, действующих на основе правил, формализующих экспертные знания, стало возможным получение хороших результатов в сейсмологической диагностике [8].

Например, в работе [9] для решения задач предсказания ураганов с наибольшей эффективностью используются методы наивного байесовского классификатора (NB [10]), хотя его отличие от других методов несущественно. В данном исследовании для анализа данных использовалась система RapidMiner.

Другой пример, в работе [11] для решения задачи прогнозирования землетрясений с наибольшей эффективностью используются методы Data Mining – методы исчисления вероятностей, байесовы и нейронные сети.

2. Задачи анализа изображений (топография и т.п.)

В сейсмологических информационных системах наибольший объем занимают есть изображения, например, топографические снимки. Это огромная часть

сейсмологических данных, которыми надо эффективно управлять [1].

Использование метода главных компонент PCA [12] (principal component analysis) для локализации анатомических частей сетчатки продемонстрировали в работе [13].

3. Задачи классификации и кластеризации

Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные, в соответствующем понимании, группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству [22].

4. Задачи предсказания

На текущий день в связи с развитием электронных сейсмологических карт, созданием межрегиональных сейсмологических баз данных в сфере сейсмологии происходит накопление большего объёма сейсмологических данных [1]. Данная тенденция позволяет решать задачи прогнозирования на основе анализа данных о землетрясениях.

1.3 Описание задачи

Для исследования использована база данных с информацией о мониторинге землетрясений на территории РФ за последние 20 лет. Сбор информации был начат в конце 90-х годов, и позже оцифрован [28]. Данные используются сотрудниками для исследований, но поскольку анализ выполняется подручными средствами (MS Excel), то соотношение результативности и трудозатрат низкое. На текущий момент база данных содержит примерно две тысячи записей. Потребителю нужна система, позволяющая упростить анализ данных.

База данных содержит в себе сейсмологические статистические данные о мониторинге землетрясений в совокупности с сопутствующей демографической информацией (возраст, пол, место жительства и т.д.). На основе анализа базы данных возможно решить задачу прогнозирования.

Выводы по главе 1

В данной главе определено понятие сейсмологической информационно-аналитической системы – комплекс аппаратных, программных средств, информационных ресурсов, методик, которые используются для обеспечения автоматизации аналитических работ для решения задачи сферы сейсмологии. Проведён анализ аналитического программного обеспечения, определена его сегментация:

– средства построения хранилищ и витрин данных (data warehouse);

– инструменты оперативной аналитической обработки (On-Line Analytical Processing, OLAP) и прочие средства многомерного анализа;

– информационно-аналитические системы (Enterprise Information Systems, EIS) и системы поддержки и принятия решений (Decision Support Systems, DSS);

– средства интеллектуальной добычи данных (data mining);

– инструменты конечного пользователя для выполнения запросов и построения отчетов (query and reporting tools).

На основании характеристики каждого класса можно сделать вывод, что наиболее богатыми аналитическими возможностями обладают хранилища данных, OLAP-средства и системы data mining. Данное утверждение также подтверждается анализом задач и обзором научных работ, посвящённых анализу данных в сфере сейсмологии, в ходе которого выделено четыре основных класса задач:

1. Задачи сейсмологической диагностики.

2. Задачи анализа изображений (топография и т.д.).

3. Задачи классификации и кластеризации.

4. Задачи предсказания (например, предсказание землетрясения).

По каждому из классов приведены примеры конкретных аналитических задач в области сейсмологии, большая часть и которых была решена средствами интеллектуальной добычи данных, что позволяет говорить о широкой распространенности использования механизмов интеллектуального анализа данных в сейсмологии.

Кроме того, на основе проведённого анализа можно сделать вывод, что для различных задач сейсмологической аналитики эффективными оказываются различные методы, выбор которых связан со значительными затратами времени специалистов в области анализа данных и не может быть сделан специалистами. Это подтверждает необходимость автоматизации процесса анализа с помощью специализированной системы, которая будет требовать от сейсмологов минимальных экспертных знаний в области интеллектуального анализа данных.

ГЛАВА 2 МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДЛЯ РЕШЕНИЯ ЗАДАЧ В ОБЛАСТИ СЕЙСМОЛОГИИ

2.1 Обзор существующих методов интеллектуального анализа данных

Целью DataMining является нахождение таких моделей, которые не могу быть найдены обычными методами. И существует два вида моделей: предсказательные и описательные [30].

Предсказательные модели: позиционируются на наборе данных с известными результатами. И используются для предсказания результатов на основании других наборов данных. Это модели классификации (описывают правила, по которым можно отнести описание объекта к одному из классов) и модели последовательностей (они описывают функции, по которым можно прогнозировать изменение непрерывных числовых параметров).

Описательные модели: они уделяют особое внимание сути зависимостей в наборе данных, взаимному влиянию различных факторов, построению эмпирических моделей. Являются легкими для восприятия человеком.

Согласно классификации по стратегиям, задачи Data Mining подразделяются на следующие группы:

– обучение с учителем;

– обучение без учителя;

– другие.

Категория обучение с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование.

Категория обучение без учителя представлена задачей кластеризации.

В категорию другие входят задачи, не включенные в предыдущие две стратегии.

Data Mining – это не один метод, а совокупность большого числа различных

методов обнаружения знаний. Базовыми методами, которые может найти технология DataMining, согласно В. А. Дюку являются [31]:

1. Ассоциация – применяется, когда несколько событий связаны между собой. Например, исследования показали, что 59 % купивших чипсы берут также и газированную воду, а если есть скидка на такой комплект, то газированную воду приобретают в 79 % случаев. Если менеджеры располагают подобными данными, то им достаточно легко оценить действенность предполагаемой скидки. Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм Apriori.

2. Классификация – выявление черт, которые будут характеризовать группу, к которой принадлежит объект, на основе обучения на уже классифицированных объектах. Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks).

3. Кластеризация – отличается от классификации тем, что группы заранее не известны и средства DataMining самостоятельно выявляют различные однородные группы данных. Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей – самоорганизующихся карт Кохонена.

4. Последовательность – применяется при существовании цепочки событий, связанных во времени. Например, при приобретении квартиры в течение месяца приобретается кухонная плита в 49 % случаев, а в течение трех недель - холодильник в 73 %.

5. Прогнозирование – создание или нахождение шаблонов, которые будут истинно показывать тенденция поведения необходимых показателей по временным рядам. При помощи них можно предсказать поведение системы в будущем. Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

2.2 Примеры использования механизмов машинного обучения в сейсмологии

2.2.1 Предсказание землетрясений (пример 1)

Исследование описывалось в 2010 году [26].

В исследовании использовался инструмент для анализа данных Tanagra (Lumière University Lyon 2).

Использовалось 3 алгоритма. Набор данных состоял из 3000 записей с 14 признаками и был разделен на тренировочный и тестовый в соотношении 70 / 30. Результаты приведены в таблице ниже (см. таблица 2).

Таблица 2 – Результаты исследования

№ п/п	Инструмент	Точность (%)	Время (мс)
1	Naïve Bayes	52,33	609
2	Decision List	52	719
3	KNN	45, 67	1000

2.2.2 Предсказание землетрясений (пример 2)

Исследование описывалось в 2008 году [27].

Описанная система названа Intelligent Heart Disease Prediction System и реализована на .net фреймворке.

Использовалось 2 классических алгоритма и нейронная сеть. Набор данных состоял из 900 записей с 15 признаками и был разделен на тренировочный и тестовый в сочетании приблизительно 50 / 50. Результаты приведены в таблице ниже (таблица 3).

Таблица 3 – Результаты исследования

№ п/п	Инструмент	Точность (%)
1	Naïve Bayes	86,53
2	Decision Tree	89
3	Neural Network	85,53

2.2.3 Предсказание землетрясений (пример 3)

Исследование проводилось в 2010 году [25].

В исследовании использовался инструмент для анализа данных Weka (University of Waikato).

Набор данных состоял из 909 записей с 13 признаками. Результаты приведены в таблице ниже (таблица 4).

Таблица 4 – Результаты исследования

№ п/п	Инструмент	Точность (%)
1	Naïve Bayes	96,5
2	Decision Tree	99,2
3	Classification via clustering	88,3

2.2.4 Масштаб применения Data Mining в сейсмологии (пример 4)

В исследовании 2013 года, проанализированы данные об инструментах Data Mining применяемых в сейсмологии [26].

В том же источнике описано исследование для предсказания успешности искусственного оплодотворения (IVF – In vitro fertilization).

Для начала, используется теория приближенных множеств (таблица 6).

Таблица 5 – Теория приближённых множеств

Факт	Предсказание		
	Успех	Неудача	Точность (%)
Успех	17	4	80.952
Неудача	26	10	27.777
Точность (%)	39.5349	71.4286	47.368

Затем искусственная нейронная сеть с обратным распространением (таблица 7).

Таблица 6 – Искусственная нейронная сеть с обратным распространением

№ п/п	Показатели ошибки	Предсказание	
		Неудача	Успех
1	MSE	0.209522132	0.212860733
2	NMSE	1.164459543	1.18301446
3	MAE	0.23114814	0.25780224
4	Min Abs Error	9.90854E-07	6.66044E-06
5	Max Abs Error	1.015785003	0.998857054
6	R	0.498099362	0.498099362
	Точность (%)	73.07692308	75

Затем комбинация этих методов (таблица 8).

Таблица 7 – Комбинация теории приближённых множеств и искусственной нейронной сети с обратным распространением

№ п/п	Показатели ошибки	Предсказание	
		Неудача	Успех
1	MSE	0.092835478	0.110601021
2	NMSE	0.378803726	0.451293836
3	MAE	0.14313612	0.191653959
4	Min Abs Error	0.002563409	0.005851654
5	Max Abs Error	1.055555499	1.055555556
6	R	0.789058201	0.789058201
	Точность (%)	89.23076923	91.83673469

Сравнение точности методов представлено в таблице 9.

Таблица 8 – Сравнение точности методов

	Теория приближенных множеств	Нейронная сеть	Комбинация
Точность в предсказании успеха	47	73	90

Выводы по главе 2

Часто параметры по умолчанию оказываются для алгоритмов, реализованных в современных мощных библиотеках машинного обучения достаточными для достижения качественного результата, что подтверждает тезис о возможности их использования различными специалистами, не имеющими глубоких знаний в алгоритмах машинного обучения. Но таким специалистам требуется предоставить доступ к инструментам, на что и нацелена описываемая система.

Как указано выше, в настоящее время, различные инструменты анализа данных доступны бесплатно и представлены в различных программных библиотеках.

Создание простой аналитической системы, которая будет давать сейсмологическим аналитикам доступ к этим библиотекам определенно востребовано.

Система должна иметь достаточно простой и расширяемый интерфейс для доступа к базе данных, в которой хранится вся информация. Расширяемость интерфейса может достигаться не инструментами в пользовательском интерфейсе, а понятным и доступным кодом, чтобы после передачи системы заказчику он мог расширить интерфейс (вводимые поля) своими силами. Система управления базами данных (СУБД) также должна выбираться с расчетом расширения.

Система должна иметь в своем составе инструменты интеллектуального анализа данных. При этом следует разделить часть системы, отвечающую за ввод данных и анализ данных.

Для обучения и проверки выборку необходимо разделить в соотношении 70/30 на обучающую и тестовую соответственно и выполнить классификацию с использованием Naïve Bayes, Decision Tree, Random Forest, Neural Network, KNN.

ГЛАВА 3 РАЗРАБОТКА МАТЕМАТИЧЕСКОГО И ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ СИСТЕМЫ АНАЛИЗА ДАННЫХ

3.1 Разработка программного обеспечения системы анализа данных

Для простоты реализации и расширения система реализуется на языке программирования Python 3, который является одним из самых популярных [33] и имеет значительное количество доступных библиотек машинного обучения. Система разделена на две основные части:

1. Подсистема ввода, хранения и управления данными.
2. Подсистема анализа данных.

Общая схема системы представлена ниже (рисунок 1) в виде компонентной диаграммы нотации UML.

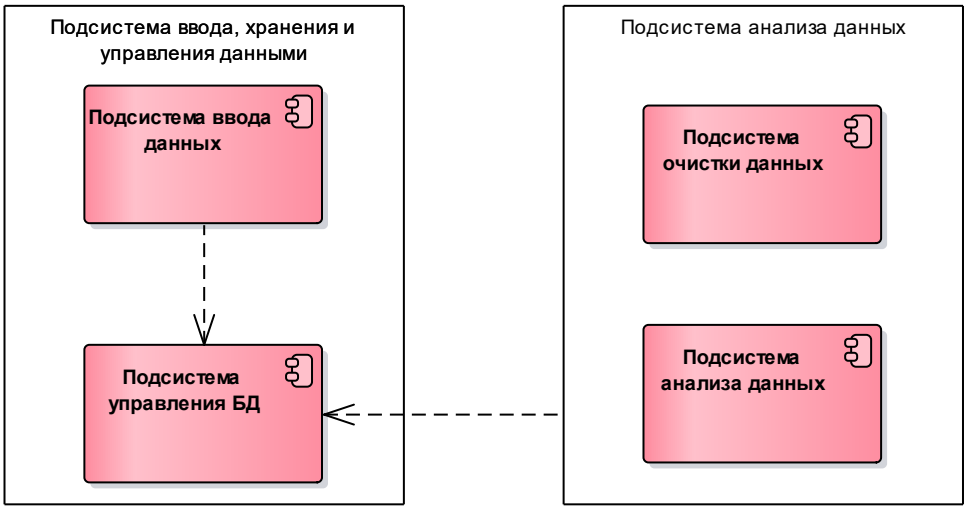

Рисунок 1 – Общая схема системы

Система анализа данных представляет собой web-сервис, который будет доступен по внутренней сети больницы и будет принимать и отдавать данные в заданном формате.

Подсистема ввода, хранения и управления данными реализована на основе

фреймворка Django работающего на основе Python 3. Подсистема анализа данных реализуется с нуля, с использованием различных библиотек для решения конкретных задач в контексте системы. Так, например, для машинного обучения используются библиотеки Scikit Learn, TensorFlow, Pandas, Numpy, для доступа к системе через сеть библиотека Flask, и т.д.

В качестве СУБД для хранения данных выбрана PostgreSQL, поскольку фреймворк Django, на основе которого реализована подсистема ввода, хранения и управления данными имеет встроенную поддержку PostgreSQL, она относится к категории свободного программного обеспечения и в сети интернет доступно много информации об этой СУБД.

Набор данных, хранимых в системе обеспечивает требования Постановления N 444 от 11 мая 1993 года Совета Министров Российской Федерации «О федеральной системе сейсмологических наблюдений и прогноза землетрясений».

3.2 Разработка математического обеспечения системы анализа данных

3.2.1 Описание исходных данных

Для решения задачи создания информационной системы сбора и обработки сейсмической геодинамической информации требуется провести сравнение форматов ее представления.

Так Центр данных ИГИ НЯЦ РК и ISC используют для обмена данными IASPEI Seismic Format (ISF). Веб-сервисы IRIS в качестве ответа на запрос отправляют пользователю данные в формате QuakeML. Геофизическая служба РАН предоставляет данные в собственном формате, при этом форма представления данных для региональных каталогов сильно различается по регионам. Рассмотрим каждый формат более подробно. IASPEI Seismic Format (ISF). ISF является утвержденным IASPEI форматом для обмена параметрическими сейсмологическими

данными (гипоцентры, магнитуды, фаза прибытия и т.д.). Он принят в качестве стандарта в августе 2001 года. Является расширением стандарта "Международная система мониторинга 1.0" (IMS1.0), который был разработан для обмена данными, использующимися для мониторинга Договора о всеобъемлющем запрещении ядерных испытаний [59]. Пример ISF представлен на рисунке 16. Вся представленная в бюллетене информация сгруппирована по нескольким секциям: название бюллетеня, название сейсмического события, характеристики сейсмического события, блок магнитуд (список значений с указанием типа магнитуды) и блок, содержащий сведения о фазах сейсмической волны. Может присутствовать дополнительная информация о сейсмических станциях и времени прихода волны на станцию. Описание блока характеристик сейсмического события представлено в таблице 4.

```
Event  9339007  CENTRAL KAZAKHSTAN

Date          Time        Err   RMS Latitude Longitude  Smaj  Smin  Az Depth   Err Ndef Nsta Gap  mdist  Mdist Qual   Author      OrigID
2009/12/05 04:37:02.47   3.86  2.22  51.2122   72.7761  60.9  12.3  30   0.0  -1.0    9    0 175   1.87   7.64 a i uk spep      9339007

Magnitudes
mb      2.52
mpv     2.14
class   6.13

Sta     Dist  EvAz Phase        Time      TRes  Azim AzRes   Slow   SRes Def   SNR       Amp   Per Qual   mb     mpva   class   ArrID
VOSZ    1.87 324.4 Pn       04:37:32.275  -3.6                          T__              0.5  0.60 mc_   2.15   1.79           93390155
BVAO    2.33 321.9 Pn       04:37:42.688   0.5 266.9 126.9  37.7   24.0 T__   2.5        0.5  0.45 mc_   2.46   2.02           93390156
VOSZ    1.87 324.4 Sn       04:38:01.488   0.3                          T__              0.9  0.30 mc_                  5.14   93390157
KURK    3.71  95.4 Pn       04:38:02.089   0.9                          T__              0.4  0.70 md_   2.51   2.15           93390158
KURBB   3.67  97.0 Pn       04:38:05.509   4.8 265.8 -15.6  12.9   -0.9 T__              0.9  0.55 md_   2.96   2.59           93390159
BVAO    2.33 321.9 Sn       04:38:13.411   0.9 140.0  -0.0  23.5   -1.2 TA_  10.0        2.9  0.65 md_                  5.99   93390160
KURK    3.71  95.4 Sn       04:38:45.307  -1.6                          T__              1.8  0.80 md_                  6.27   93390161
KURBB   3.67  97.0 Sn       04:38:45.521  -0.4 185.9 -95.5  26.6    1.9 T__              5.7  0.80 md_                  7.11   93390162
MK31    7.64 121.6 Lg       04:41:03.866  -2.0 316.1   7.4  21.2  -10.6 TA_   4.1        1.2  0.85 mc_                         93390163

Event  9339008  NORTHERN XINJIANG, CHINA

Date          Time        Err   RMS Latitude Longitude  Smaj  Smin  Az Depth   Err Ndef Nsta Gap  mdist  Mdist Qual   Author      OrigID
2009/12/05 04:59:08.36   3.59  0.00  44.7786   81.8402  29.0  19.5   8   0.0  -1.0    2    0 175   2.04   2.04 a i uk spep      9339008
```

Рисунок 2 – Бюллетень в формате ISF

Наибольший объем данных по сейсмическим событиям, зарегистрированных Геофизической службой РАН, собран в виде сейсмологического каталога, пример которого представлен на рисунке 17. Данное издание составляется на основе следующих потоков входящей информации:

ежедневных оперативных сводок с данными опорных сейсмических станций России и СНГ, поступающих по электронной почте и по телетайпным каналам связи;

оперативных каталогов и сводок с региональных сейсмических станций, по-

ступающих по электронной почте из филиалов ГС РАН (Дагестанского, Камчатского, Сахалинского, Магаданского и СевероОсетинского), из филиалов ГС СО РАН (Алтае-Саянского, Байкальского и Якутского), а также из лабораторий ГС РАН (Кавминводской и Воронежской);

ежедневных оперативных сводок из 10 стран мира: Финляндии, Англии, Дании, Польши, США, Германии, Румынии, Болгарии, Чехии, Словакии, Венгрии, поступающих по электронной почте;

станционных сводок, создаваемых в ГС РАН (г. Обнинск) при обработке волновых форм, поступающих с цифровых сейсмических станций России и зарубежных стран в режиме, близком к реальному времени, и по каналам Интернет;

станционных бюллетеней сейсмических станций России и СНГ, поступающих по e-mail и почте;

бюллетеней Национального центра данных о землетрясениях США (PDEEDR, NEIC, США) и Международного центра данных (REB, СТВТО, Австрия).

Таблица 9 – Описание признаков в исходном наборе данных

№	Кодовое имя	Описание	Тип данных	Кол-во не пустых	Примечание
1	Date	Дата события в формате yyyy/mm/dd	ЦЧ	-	Не участвует Одно значение во всех записях
2	Time	Время события в формате hh:mm:ss.ss	ЦЧ	-	Не участвует. ПД
3	Err	Ошибка в определении времени возникновения	С	-	Не участвует. ПД

№	Кодовое имя	Описание	Тип данных	Кол-во не пустых	Примечание
		события, в секундах (пустое, если внесено исправление)			
4	RMS	Среднеквадратичная невязка времени в очаге, в секундах	С	-	Не участвует. ПД
5	Latitude	Широта, градусы	С	-	Не участвует. ПД
6	Longitude	Долгота, градусы	С	-	Не участвует. ПД
7	Smaj	Большая полуось 95% доверительного эллипса ошибок определения эпицентра, км	С	-	Не участвует. ПД
8	Smin	Малая полуось 95% доверительного эллипса ошибок определения эпицентра, км	Д	1917	Будет преобразовано в признак «Возраст на момент постановки диагноза»
9	Az	Азимут большой полуоси, градусы	ЦЧ	1929	
10	Depth	Глубина очага, км	Д	1657	Не участвует. НПИ

№	Кодовое имя	Описание	Тип данных	Кол-во не пустых	Примечание
11	Err	Ошибка в определении глубины очага, в км	Д	1618	Не участвует. НПИ
12	Ndef	Число определений фаз	Д	1921	Будет преобразовано в признак «Возраст на момент постановки диагноза»
13	Nst	Число станцию участвующих в определении характеристик события	Д	219	Не участвует. НПИ, МД

Продолжение таблицы 10

№	Кодовое имя	Описание	Тип данных	Кол-во не пустых	Примечание
14	Gap	Разрыв в охвате азимута, градусы	ЦЧ	929	Не участвует. МД
15	mdist	Расстояние до ближайшей станции, градусы	Д	359	Будет преобразовано в целевой признак «Факт смерти»
16	Mdist	Расстояние до самой удаленной	С	12	Не участвует. МД

		станции, градусы			
17	Qual	Тип события	ЦЧ	786	Не участвует. НПИ, МД
18	Author	Автор	ЦЧ	380	Не участвует. МД
19	OrigID	Номер события, его идентификатор	ЦЧ	984	Не участвует. НПИ
20	MPSP	Средние значения магнитуд по продольным волнам зарегистрированным короткопериодной аппаратурой	ЦЧ	540	Не участвует. МД

Продолжение таблицы 10

№	Кодовое имя	Описание	Тип данных	Кол-во не пустых	Примечание
21	MPLP	Средние значения магнитуд по продольным волнам, зарегистрированным длиннопериодной аппаратурой	ЦЧ	439	Не участвует. МД
22	MS	Средние значения	С	1912	

		магнитуд по поверхностным волнам			

3.2.2 Подготовка данных к анализу

QuakeML - это модель представления сейсмологических данных, ранее преобразованных в формат XML [83]. QuakeML предназначен для того, чтобы стандартизировать сейсмологический обмен данными, и применяется для решения широкого диапазона научных и технических проблем в сейсмологии.

«QuakeML» основан на стандарте обмена данными в сейсмологии, который активно развивается. Первая версия (0.51) выпущена в январе 2007 года, текущая версия (1.2) выпущена 14 февраля 2013 года. Она основана на общественном мнении о стандарте обмена данными в сейсмологии, выраженным следующими организациями: ETH, GFZ, USC, СУЭК, USGS, IRIS DMC, EMSC, ORFEUS, GNS, ZAMG, BRGM, Nanometrics и ISTI. В версии (1.2) рассматривается описание сейсмических событий, включая: приборы, приходы, амплитуды, величины, происхождения, механизмы очагов и момент тензоров.

«QuakeML» - файл имеет следующее строение:

- текст, заключенный между тегами <Module> и </Module> означает название службы и её версия, использованная для получения данных. Здесь же указан веб адрес, по которому можно обратиться к этой службе;
- текст, заключенный между тегами <ModuleURI> и </ModuleURI> содержит URI-адрес, по которому на прямую были получены данные. Этот адрес содержит в себе веб адрес службы, а также параметры запроса;
- текст, заключенный между тегами <SentDate> и </SentDate> - это дата, когда был проведен запрос к веб сервису;

- теги <Network> и </Network> заключают данные о произошедших сейсмических событиях, зафиксированные определенной сетью сейсмографов;
- текст, заключенный между тегами <StartDate> и </StartDate>, означает начало периода, за который был произведен поиск сейсмических событий в заданной области;
- текст, заключенный между тегами <EndDate> и </EndDate>, означает конец периода, за который был произведен поиск сейсмических событий в заданной области;
- текст, заключенный между тегами <Description> и </Description>, означает название сети сейсмографов;
- текст, заключенный между тегами <TotalNumberStations> и </TotalNumberStations>, означает общее количество сейсмографов в данной сети;
- текст, заключенный между тегами <SelectedNumberStations> и </SelectedNumberStations>, означает номер станции, зафиксировавшей сейсмическое событие;
- между тегами <Station> и </Station> содержится информация о названии города и страны, в котором произошло сейсмическое событие, а также координаты широты и долготы этого события;
- теги <StationEpoch> и </StationEpoch> означают начало и конец данных, прилагающихся к конкретному событию, зафиксированному определенной станцией;
- текст, заключенный между тегами <StartDate> и </StartDate>, означает начало периода, в который попало зафиксированное событие;
- текст, заключенный между тегами <EndDate> и </EndDate>, означает конец периода, в который попало зафиксированное событие;
- числовое значение, заключенный между тегами <Lat> и </Lat>, означает координаты широты, зафиксированного события;

- числовое значение, заключенный между тегами <Lon> и </Lon>, означает координаты долготы, зафиксированного события;
- числовое значение, заключенный между тегами <Elevation> и </Elevation>, означает высоту, над уровнем моря, на которой было зафиксировано сейсмическое событие;
- теги <Site>и </Site>, соответствуют сейсмическому событию и содержат в себе теги <Name> и </Name>;
- текст, заключенный между тегами <Name> и </Name>, означает название города и страны, в которых было зафиксировано сейсмическое событие;
- текст, заключенный между тегами <Agency> и </Agency> означает название службы, зафиксировавшей сейсмическое событие.

Один из плюсов «QuakeML» - это использование его международными службами, которые собирают данные с многих источников с потенциально противоречивыми мнениями по поводу того, как разделить наблюдаемое землетрясение на одиночные события.

В настоящее время «QuakeML» услуги установлены в нескольких учреждениях по всему миру, в том числе: EMSC, ORFEUS, ETH, Geoazur (Европа), NEIC, ANSS, СУЭК / SCSN (США), и GNS науки (Новая Зеландия).

Некоторые из этих учреждений уже предоставляют услуги «QuakeML» веб-каталога землетрясений [83].

К данным описанными признаками, представленными в таблице 13 и таблице 14 последовательно были применены несколько алгоритмов машинного обучения: Логистическая регрессия, Решающее дерево, Random Forest, Gradient Tree, Наивный Байес для нормального распределения, Наивный Байес для распределения Бернулли. Все алгоритмы используются внутри классов-оберток, которые разделяют общий интерфейс, реализованный в виде соглашения с оговоренным набором методов с обозначенной сигнатурой.

В течении эксперимента мониторинг выполнялся по следующим параметрам: правильность (accuracy), точность (precision), полнота (recall), ф-мера (f-score),

время. Ф-мера является основным параметром для сравнения. Остальные величины взяты для построения полной картины. Время оценивается только для получения общего представления о сравнении времени работы алгоритмов. Для оценки указанных параметров набор данных делится на обучающий и тестовый в сочетании 0,75 на 0,25. Результаты приведены в таблице 15.

Таблица 11 – Результаты отладочного эксперимента

№	Алгоритм	Правильность (accuracy)	Точность (precision)	Полнота (recall)	Ф-мера (f-score)	Время
1	K-ближайших соседей (таблица 13 – 1552 записи)	0.802	0.734	0.802	0.752	0.125
2	K-ближайших соседей (таблица 14 – 1911 записей)	0.789	0.729	0.789	0.748	0.125
3	Логистическая регрессия (таблица 13 – 1552 записи)	0.814	0.772	0.814	0.778	0.031
4	Логистическая регрессия (таблица 14 – 1911 записей)	0.81	0.665	0.81	0.73	0.016
5	Решающее дерево (таблица 13 – 1552 записи)	0.765	0.745	0.765	0.754	0.016
6	Решающее дерево (таблица 14 – 1911 записей)	0.81	0.745	0.81	0.75	~ 0.0

№	Алгоритм	Правильность (accuracy)	Точность (precision)	Полнота (recall)	Ф-мера (f-score)	Время
7	Random Forest (таблица 13 – 1552 записи)	0.822	0.782	0.822	0.781	0.234
1	Random Forest (таблица 14 – 1911 записей)	0.81	0.745	0.81	0.75	0.234
2	Gradient Tree (таблица 13 – 1552 записи)	0.82	0.778	0.82	0.779	0.281
3	Gradient Tree (таблица 14 – 1911 записей)	0.814	0.752	0.814	0.747	0.141
4	Наивный Байес для нормального распределения (таблица 13 – 1552 записи)	0.353	0.786	0.353	0.364	~ 0.0
5	Наивный Байес для нормального распределения (таблица 14 – 1911 записей)	0.236	0.734	0.236	0.173	0.016

Окончание таблицы 15

№	Алгоритм	Правиль-ность (accuracy)	Точность (precision)	Полнота (recall)	Ф-мера (f-score)	Время
6	Наивный Байес для распределения Бернулли (таблица 13 – 1552 записи)	0.804	0.762	0.804	0.773	0.016
7	Наивный Байес для распределения Бернулли (таблица 14 – 1911 записей)	0.812	0.751	0.812	0.752	~ 0.0

Результаты эксперимента по Ф-мере представлены ниже (рисунок 2).

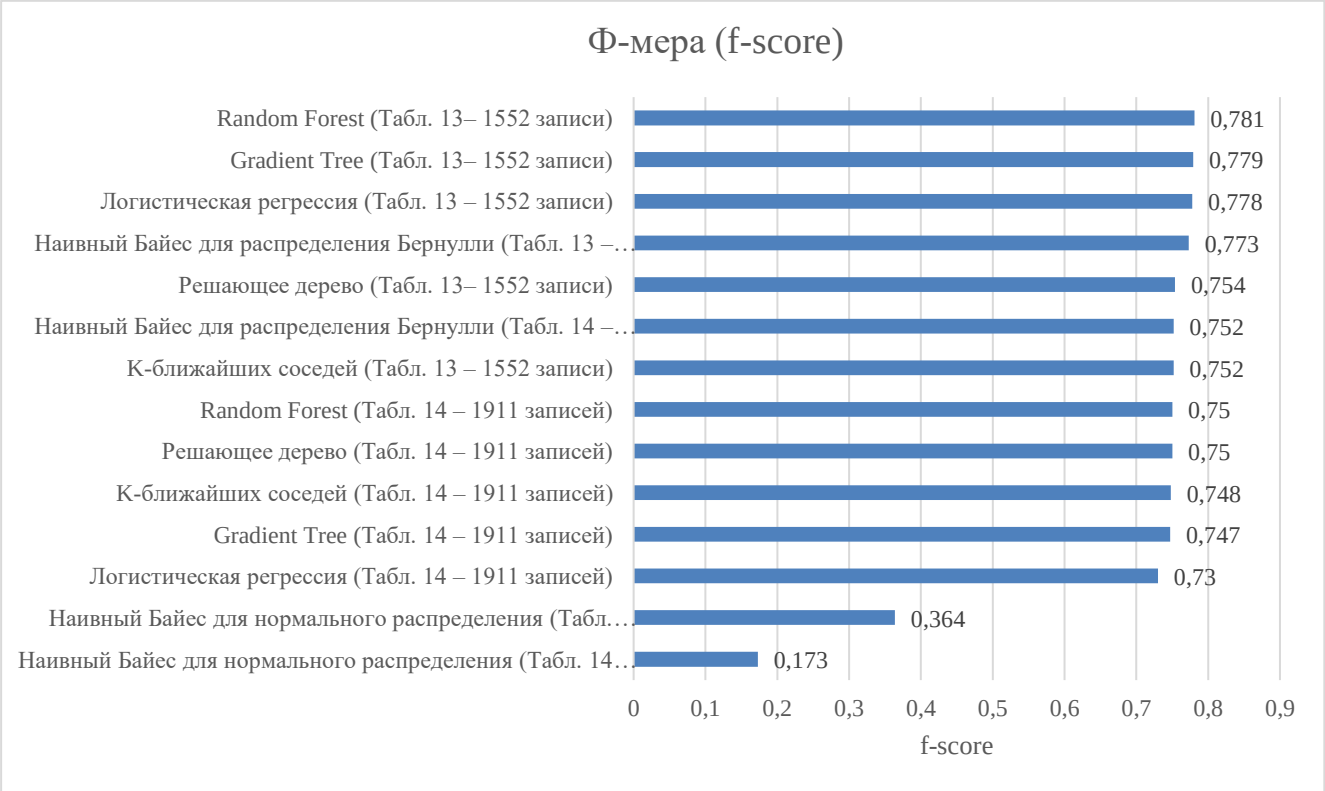

Рисунок 3 – Результаты эксперимента

Выводы по главе 3

В данной главе разработан проект реализации программного обеспечения сейсмологической аналитической системы: система разделена на две основные части:

– подсистема ввода, хранения и управления данными;

– подсистема анализа данных.

В качестве СУБД для хранения данных выбрана PostgreSQL, подсистема ввода, хранения и управления данными будет реализована на основе фреймворка Django работающего на основе Python 3. Подсистема анализа данных реализуется, с использованием различных библиотек для решения конкретных задач в контексте системы. Изучен на бор данных, который используется для разработки математического обеспечения для решения задачи, определённой в главе 1 данной работы. Данные о мониторинге злокачественных новообразований у детей и подростков состоят из 72 признаков, включающий в себя 1929 записей. Определён порядок подготовки данных к анализу.

В качестве метрики для оценки математического обеспечения определены: доля правильно классифицированных объектов (Accuracy), точность (Precision), полнота (Recall) и F-мера, последняя является основным параметром для сравнения

Часто параметры по умолчанию оказываются для алгоритмов, реализованных в современных мощных библиотеках машинного обучения достаточными для достижения качественного результата, что подтверждает тезис о возможности их использования различными специалистами, не имеющими глубоких знаний в алгоритмах машинного обучения. Но таким специалистам требуется предоставить доступ к инструментам, на что и нацелена описываемая система.

Алгоритмы, отобранные для реализации математического обеспечения системы, выдают довольно высокие результаты. Наиболее эффективными на примере решаемой задачи показали себя Random Forest (F-мера 0,781) и Gradient Tree (F-мера 0,779).

ГЛАВА 4 КОММЕРЦИАЛИЗАЦИЯ ПРОЕКТА

1.1 Актуальность коммерциализации

Различные задачи сейсмологической аналитики требуют использования разных методов анализа, выбор которых связан со значительными затратами времени специалистов в области анализа данных и не может быть сделан сейсмологами. Это подтверждает актуальность автоматизации процесса анализа с помощью специализированной системы, которая будет требовать от специалистов минимальных экспертных знаний в области интеллектуального анализа данных.

Описанная в главе 3 архитектура системы позволяет специалистам с минимальным погружением в специфику Data Science решать задачи сейсмологической аналитики.

Кроме того, структура хранения данных обеспечивает требования Постановления N 444 от 11 мая 1993 года Совета Министров Российской Федерации «О федеральной системе сейсмологических наблюдений и прогноза землетрясений». [28]. В настоящее время в России используются несколько компьютерных программ территориального популяционного регистра, разработанных на единой методологической основе, но продолжают функционировать и программы с недостаточным объемом вводимой информации, чаще всего это программы, созданные в 80-х годах прошлого века. Функциональность данных программ ограничена задачами хранения, накопления данных и формирования выборок данных.

1.2 Дорожная карта коммерциализации проекта

Для оценки возможностей развития и коммерциализации проекта рассмотрена дорожная карта коммерциализации проекта в проекции двух лет. В рамках

первого года рассматривается этап научно-исследовательской работы и апробирования системы в одной сейсмологической организации (таблица 16).

Таблица 10 – Дорожная карта коммерциализации проекта 2019 год. Этап 1 Научно-исследовательская работа и апробирование системы в одной сейсмологической организации

Направление	2019 год. Этап 1 Научно-исследовательская работа и апробирование системы в одной сейсмологической организации			
	1 квартал	II квартал	III квартал	IV квартал
Исследования и разработки	Исследование предметной области и текущего состояние системы	Исследование архитектурных решений в области сбора, хранения и анализа сейсмологических данных. Исследование эффективности алгоритмов анализа сейсмологических данных	Исследование эффективности алгоритмов анализа сейсмологических данных	Исследование эффективности алгоритмов анализа сейсмологических данных

Продолжение таблицы 16

Направление	2019 год. Этап 1 Научно-исследовательская работа и апробирование системы в одной сейсмологической организации			
	1 квартал	II квартал	III квартал	IV квартал
Создание продукта	Спецификация требований на основании проведённого исследования	Разработка структуры хранения данных и миграция данных	Разработка подсистем ввода, хранения и управления данными. Заполнение справочников. Разработка системы анализа.	Тестирование. Опытная эксплуатация системы
Общее организационное развитие и план по найму	Формирование плана кадрового развития	Подбор команды	Обучение команды	-
Защита интеллектуальной собственности и лицензирование	-	-	-	Подача заявок на регистрацию прав собственности на новый продукт

Окончание таблицы 16

Направление	2019 год. Этап 1 Научно-исследовательская работа и апробирование системы в одной сейсмологической организации			
	1 квартал	II квартал	III квартал	IV квартал
Маркетинг, внедрение продвижение	-	-	Представление новых возможностей продукта на выставках и конференциях	Представление новых возможностей продукта на выставках и конференциях

В рамках следующего года рассматривается этап развития функциональных возможностей аналитической системы (таблица 17).

Таблица 11 – Дорожная карта коммерциализации проекта 2020 год. Этап 2 Развитие функциональных возможностей

Направление	2020 год. Этап 2 Развитие функциональных возможностей			
	1 квартал	II квартал	III квартал	IV квартал
Исследования и разработки	Исследование повышения эффективности алгоритмов анализа сейсмологических данных и выявления потребности в новых функциях и анализе	Исследование актуальных задач анализа данных и исследование эффективности алгоритмов анализа сейсмологических данных	Исследование актуальных задач анализа данных и исследование эффективности алгоритмов анализа сейсмологических данных	Исследование актуальных задач анализа данных и исследование эффективности алгоритмов анализа сейсмологических данных

Продолжение таблицы 17

Направление	2020 год. Этап 2 Развитие функциональных возможностей			
	1 квартал	II квартал	III квартал	IV квартал
Создание продукта	Улучшения механизмов анализа данных	Разработка новых функциональных возможностей по решению сейсмологических аналитических задач	Тестирование	Опытная эксплуатация системы
Общее организационное развитие и план по найму	Подбор новых кадров и их обучение	-	-	-
Защита интеллектуальной собственности и лицензирование	-	-	-	Подача заявок на регистрацию прав собственности на новый продукт
Маркетинг, внедрение продвижение	-	Представление новых возможностей продукта на выставках и конференциях	Маркетинговая компания по продвижению продукта. Создание сайта продукта	Заключение договоров на внедрение системы

1.3 Цели и задачи

Целью является создание аналитической сейсмологической системы, назначение которой заключается в накоплении, хранении и анализе данных о мониторинге землетрясений.

Для достижения поставленной цели в рамках первого года развития проекта необходимо выполнить задачи в соответствии с планом работ (рисунок 3).

Название задачи	Длительность	Начало	Окончание
Определение требований к системе и подготовка проектной документации	28 дней	Пн 14.01.19	Ср 20.02.19
Подписание проектной документации	3 дней	Чт 21.02.19	Пн 25.02.19
Спецификация требований к системе и системный анализ	30 дней	Вт 26.02.19	Пн 08.04.19
Разработка структуры хранения данных	20 дней	Вт 09.04.19	Пн 06.05.19
Миграция данных	10 дней	Вт 07.05.19	Пн 20.05.19
Разработка подсистема ввода, хранения и управления данными	28 дней	Вт 21.05.19	Чт 27.06.19
Заполнение основных справочников	14 дней	Пт 28.06.19	Ср 17.07.19
Разработка подсистемы анализа данных	32 дней	Чт 18.07.19	Пт 30.08.19
Настройка и тестирование	21 дней	Пн 02.09.19	Пн 30.09.19
Подготовка эксплутационной документации	10 дней	Вт 01.10.19	Пн 14.10.19
Обучение персонала	21 дней	Вт 15.10.19	Вт 12.11.19
Опытная эксплуатация	14 дней	Ср 13.11.19	Пн 02.12.19
Внесение коррективов	14 дней	Вт 03.12.19	Пт 20.12.19
Приёмка работ	1 день	Пн 23.12.19	Пн 23.12.19
Ввод в эксплуатацию	7 дней	Вт 24.12.19	Ср 01.01.20

Рисунок 2 – Календарный план работ

1. Определение требований к системе и подготовка проектной документации.
2. Подписание проектной документации.
3. Спецификация требований к системе и системный анализ.
4. Разработка структуры хранения данных.
5. Миграция данных.

6. Разработка подсистемы ввода, хранения и управления данными в соответствии с разработанным в ходе магистерской работы программным обеспечением.

7. Заполнение основных справочников.

8. Разработка подсистемы анализа данных в соответствии с разработанными в ходе магистерской работы программным обеспечением и математическим обеспечением.

9. Настройка и тестирование.

10. Подготовка эксплуатационной документации.

11. Обучение персонала.

12. Опытная эксплуатация.

13. Внесение коррективов в систему по итогам опытной эксплуатации.

14. Приёмка работ.

15. Ввод в эксплуатацию.

Диаграмма Ганта составленная на основе календарного плана представлена ниже (Рисунок 3). Предполагаемый срок разработки и тестирования такого программного продукта – 1 год.

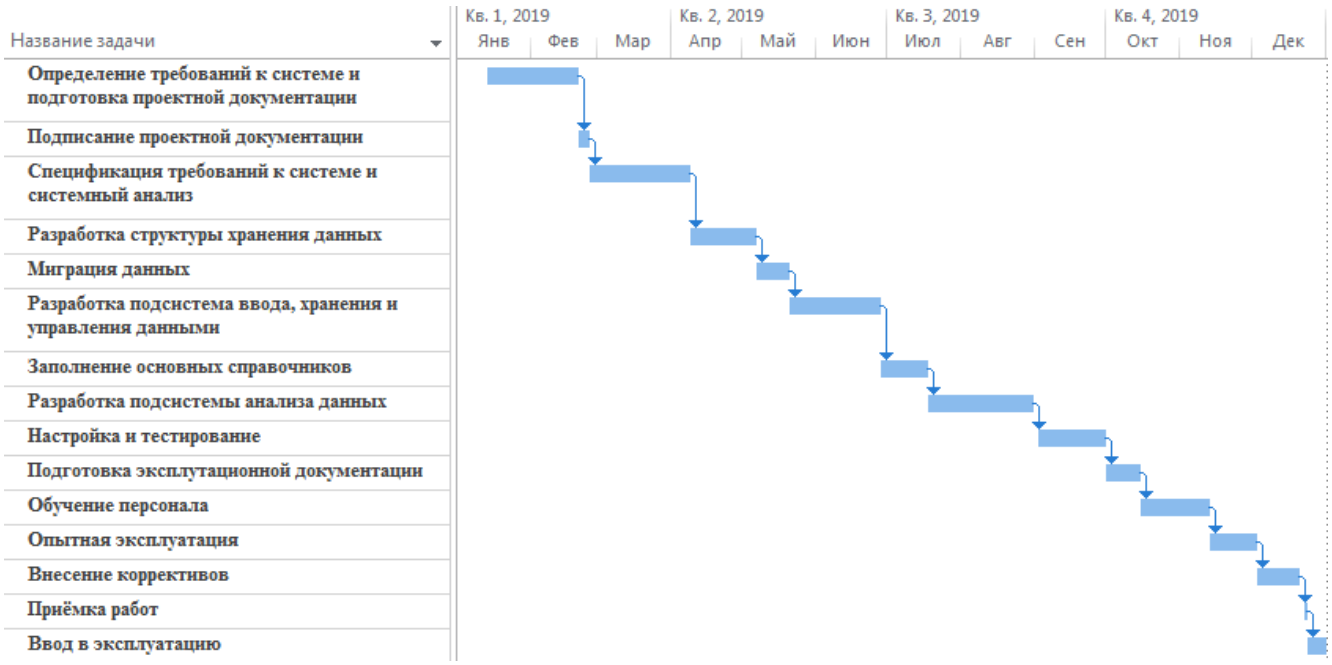

Рисунок 3 – Диаграмма Ганта

Выводы по главе 4

В данной главе составлена дорожная карта коммерциализации проекта на два года. Кроме того, составлен календарный план работ на первый год коммерциализации проекта. Предполагаемый срок разработки и тестирования такого программного продукта – 1 год.

Различные задачи сейсмологической аналитики требуют использования разных методов анализа, выбор которых связан со значительными затратами времени специалистов в области анализа данных и не может быть сделан сейсмологами, что подтверждает актуальность коммерциализации проекта.

ЗАКЛЮЧЕНИЕ

На основе полученных знаний во время обучения по направлению «Бизнес-информатика» и анализа научной и научно-исследовательской литературы и публикаций была проведена работа над разработкой математического и программного обеспечения сейсмологической аналитической системы.

В рамках проведённого исследования:

1. Определено понятие сейсмологической информационно-аналитической системы – комплекс аппаратных, программных средств, информационных ресурсов, методик, которые используются для обеспечения автоматизации аналитических работ для решения задачи сферы сейсмологии. Проведён анализ аналитического программного обеспечения.

2. Проведён анализ задач и обзор научных работ, посвящённых анализу данных в сфере сейсмологии, в ходе которого выделено четыре основных класса задач:

– задачи сейсмологической диагностики;

– задачи анализа изображений (топография и т.п.);

– задачи классификации и кластеризации;

– задачи предсказания (например, предсказание землятресений).

3. Определенна задача для проведения исследования, на примере решения которой разработан проект математического и программного обеспечения сейсмологической аналитической системы.

4. Проведено исследование существующих методов интеллектуального анализа данных для разработки математического программного обеспечения. Проведён анализ научных работ по использованию механизмов машинного обучения в сейсмологии и описаны примеры их использования.

5. Разработан проект реализации программного обеспечения сейсмологической аналитической системы: система разделена на две основные части:

– подсистема ввода, хранения и управления данными.

– подсистема анализа данных.

В качестве СУБД для хранения данных выбрана PostgreSQL, подсистема ввода, хранения и управления данными будет реализована на основе фреймворка Django работающего на основе Python 3. Подсистема анализа данных реализуется, с использованием различных библиотек для решения конкретных задач в контексте системы.

6. Определено математическое обеспечение системы: наиболее эффективными на примере решаемой задачи показали себя алгоритмы Random Forest (0,781) и Gradient Tree (0,779).

7. Составлена дорожная карта коммерциализации проекта на два года и составлен календарный план работ на первый год коммерциализации проекта. Предполагаемый срок разработки и тестирования такого программного продукта – 1 год.

Таким образом, решены все поставленные в данной работе задачи и цель магистерской работы можно считать достигнутой.

Направление дальнейшего исследования: повышение эффективности механизмов анализа данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Tomar D., Agarwal S. A survey on Data Mining approaches for Healthcare // International Journal of Bio-Science and Bio-Technology. – 2013. – Vol. 5 № 5. – P. 241-266.

2. Белов В.С. ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИЕ СИСТЕМЫ. Основы проектирования и применения: учебное пособие, руководство, практикум / Московский государственный университет экономики, статистики и информатики. — М., 2015. — 111 с.

3. Inmon W. H.Building the Data Warehouse, Third Edition John Wiley & Sons, Inc. New York, 2002 – 428 p.

4. Iqbal, M.I. Detection of vascular intersection in retina fundus image using modified cross point number and neural network technique / A.M. Aibinu, M. Nilsson, I.B. Tijani more authors // Int. Conf. Comput. Commun. Eng. - 2008. - P. 241-246.

5. Карасева Т.С. Решение задач сейсмологической диагностики методами интеллектуального анализа данных // Решетневские чтения. 2015. №19. URL: https://cyberleninka.ru/article/n/reshenie-zadach-meditsinskoy-diagnostiki-metodami-intellektualnogo-analiza-dannyh (дата обращения: 20.05.2019).

6. Langley P., Iba W., Thompson K. An analysis of Bayesian classifiers // Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, CA : AAAI, 1992. P. 223-228.

7. Потапов, А. Д. Землетрясения. Причины и последствия / А.Д. Потапов, И.Л. Ревелис. – М.: Высшая школа, 2009. – 248 с.

8. Потапов, В.П. Облачные вычисления в СО РАН – возможности применения и реализации / В.П. Потапов, О.Л. Пястунович, И.Е. Харлампенков // XII Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы» (DICR’2010), 30 ноября – 3 декабря 2010. – Новосибирск, 2010. – С. 17.

9. Потапов, В.П. Применение Internet-технологий для анализа и мониторинга

сейсмической ситуации горнодобывающего региона / В.П. Потапов, И.Е. Харлампенков // Вычислительные и информационные технологии для наук об окружающей среде: Избранные труды Международной молодежной школы и конференции CITES-2011, 3-13 июля 2011г., Томск. – Томск: Изд-во Томского ЦНТИ, 2011. – С. 173-175.

10. Beck, T. Robust model-based centerline extraction of vessels in CTA data / T. Beck, C. Biermann, D. Fritz, R. Dillmann // Proceedings of SPIE. – 2009. – Vol. 7259. – 72593O(9 pp). -doi:10.1117/12.810753.

11. Sinthanayothin, C. Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images / C. Sinthanayothin, J. Boyce, H. Cook, T. Williamson // British Journal of Ophthalmology. – 1999. – Vol. 83(8). – P. 902-910.

12. Abramoff, M. Web-based screening for diabetic retinopathy in a primary care population: The eye check project / M. Abramoff, M. Suttorp // Telemedicine and e-Health. – 2005. – Vol. 11(6). – P. 668-674.

13. Jan, J. Retinal image analysis aimed at blood vessel tree segmentation and early detection of neural-layer deterioration / J. Jan, J. Odstrcilik, J. Gazarek, R. Kolar // Computerized Medical Imaging and Graphics. - 2012. - Vol. 36(6). - P. 431-441.

14. Kheng, G.G. An automatic diabetic retinal image screening system book chapter in medical data mining and knowledge discovery / G.G. Kheng, H.S. Wynne, M. Li, H. Wang // Edited by Krzysztof Cios. - 2001. - Vol. 29. - P. 181-210.

15. Marin, D. A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features / D. Marin, A. Aquino, M.E. Gegundez-Arias, J.M. Bravo // IEEE Transactions on Medical Imaging. - 2011. - Vol. 30(1). -P. 146-158.

16. Newey, V.R. Online artery diameter measurement in ultrasound images using artificial neural networks / V.R. Newey, D.K. Nassiri // Ultrasound Med. Biol. - 2002. - Vol. 28(2). - P. 209-216.

17. Gregory, S. Nearest-neighbor methods in learning and vision: theory and practice / S. Gregory, D. Trevor, I. Piotr // Neural Information Processing / MIT Press,

2006.

18. Staal, J.J. Ridge based vessel segmentation in color images of the retina / J.J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken // IEEE Transactions on Medical Imaging. -2004. - Vol. 23(4). - P. 501-509.

19. Akita, K. A computer method of understanding ocular fundus images / K. Akita, H. Kuga // Pattern Recognition. - 1982. - Vol. 15. - P. 431-443.

20. Берестнева Ольга Григорьевна, Осадчая Ирина Александровна, Немеров Евгений Владимирович Методы исследования структуры сейсмологических данных // Вестник науки Сибири. 2012. №1 (2). URL: https://cyberleninka.ru/article/n/metody-issledovaniya-struktury-meditsinskih-dannyh (дата обращения: 20.05.2019).

21. Войтикова М.В., Войтович А.П., Хурса Р.В. Применение интеллектуального анализа данных для классификации гемодинамических состояний // АГ. 2015. №5 (43). URL: https://cyberleninka.ru/article/n/primenenie-intellektualnogo-analiza-dannyh-dlya-klassifikatsii-gemodinamicheskih-sostoyaniy-1 (дата обращения: 20.05.2019).

22. ANBARASI M., ANUPRIYA E., N.CH.S.N.IYENGAR, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376.

23. Rajkumar Asha, G.Sophia Reena, Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

24. Sellappan Palaniappan Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

25. Барсегян А.А., Куприянов М.С, Степаненко В.В.,. Холод И.И. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP : учеб. пособие. 2-е изд. / СПб.: 2007. 59 с.

26. M. Durairaj, V. Ranjani, Data Mining Applications In Healthcare Sector: A Study, International Journal of Engineering Science and Technology Vol. 2(10), 2013,

2277-8616.

27. RapidMiner сайт [электронный ресурс] – Режим доступа. – URL: https://rapidminer.com (дата обращения 02.07.2017);

28. Cao Z., Cao S., Xiong G., Guo L.Progress in Study of Encrypted Traffic Classification. In Proceedings of International standard conference on trustworthy computing and services, 2012, Beijing, China, pp. 78-86

29. Гайдышев И.П. Оценка качества бинарных классификаторов // Вестник ОмГУ. 2016. №1 (79). URL: https://cyberleninka.ru/article/n/otsenka-kachestva-binarnyh-klassifikatorov (дата обращения: 26.05.2019).

30. Богданов Л. Ю. Оценка эффективности бинарных классификаторов на основе логистической регрессии методом ROC-анализа // Вестник СГТУ. 2010. №2с. URL: https://cyberleninka.ru/article/n/otsenka-effektivnosti-binarnyh-klassifikatorov-na-osnove-logisticheskoy-regressii-metodom-roc-analiza (дата обращения: 26.05.2019)

31. Документация библиотеки Scikit Learn, раздел sklearn.decomposition.PCA, сайт [электронный ресурс] – Режим доступа. – URL: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (дата обращения 09.12.2017);

32. Павлова В. Ю. Основные вопросы статистического анализа в сейсмологических исследованиях // Клиническая онкогематология. 2009. №4. URL: https://cyberleninka.ru/article/n/osnovnye-voprosy-statisticheskogo-analiza-v-meditsinskih-issledovaniyah (дата обращения: 20.05.2019).

33. Савченко Л.М., Бежитский С.С. DataMining и области его применения // Актуальные проблемы авиации и космонавтики. 2015. №11. URL: https://cyberleninka.ru/article/n/datamining-i-oblasti-ego-primeneniya (дата обращения: 21.05.2019).