**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное бюджетное образовательное учреждение высшего образования "Пермский государственный национальный исследовательский университет"**

**Кафедра математического обеспечения вычислительных систем**

Авторы-составители: **Юрков Кирилл Александрович**
**Городилов Алексей Юрьевич**

Рабочая программа дисциплины

# ДОБЫЧА ЗНАНИЙ: ТЕОРЕТИЧЕСКИЕ ОСНОВЫ И ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА DATA MINING

Код УМК 96041

Утверждено
Протокол №5
от «09» июня 2020 г.

Пермь, 2020

## 1. Наименование дисциплины

Добыча знаний: теоретические основы и инструментальные средства Data mining

## 2. Место дисциплины в структуре образовательной программы

Дисциплина входит в вариативную часть Блока « М.1 » образовательной программы по направлениям подготовки (специальностям):

Направление: **01.04.02** Прикладная математика и информатика

направленность Математическое и программное обеспечение вычислительных систем

## 3. Планируемые результаты обучения по дисциплине

В результате освоения дисциплины **Добыча знаний: теоретические основы и инструментальные средства Data mining** у обучающегося должны быть сформированы следующие компетенции:

**01.04.02** Прикладная математика и информатика (направленность : Математическое и программное обеспечение вычислительных систем)

**ПК.4** Способен интегрировать разработанное системное программное обеспечение

**Индикаторы**

**ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях

## 4. Объем и содержание дисциплины

| | |
|---|---|
| **Направления подготовки** | 01.04.02 Прикладная математика и информатика (направленность: Математическое и программное обеспечение вычислительных систем) |
| **форма обучения** | очная |
| **№№ триместров, выделенных для изучения дисциплины** | 4 |
| **Объем дисциплины (з.е.)** | 3 |
| **Объем дисциплины (ак.час.)** | 108 |
| **Контактная работа с преподавателем (ак.час.), в том числе:** | 36 |
| **Проведение лекционных занятий** | 12 |
| **Проведение лабораторных работ, занятий по иностранному языку** | 24 |
| **Самостоятельная работа (ак.час.)** | 72 |
| **Формы текущего контроля** | Защищаемое контрольное мероприятие (3)<br>Итоговое контрольное мероприятие (1)<br>Письменное контрольное мероприятие (2) |
| **Формы промежуточной аттестации** | Зачет (4 триместр) |

## 5. Аннотированное описание содержания разделов и тем дисциплины

### Добыча знаний: теоретические основы и инструментальные средства Data mining

#### Раздел 1. Основные принципы анализа данных и добычи знаний

В данном разделе вводятся основные понятия добычи знаний. Также рассматривается история развития интеллектуального анализа данных.
Во второй теме рассматриваются задачи решаемые Data Mining, примеры практических решений.

#### Раздел 2. Предобработка данных как необходимый этап добычи знаний

Тема 3. Data Mining как проект. Жизненный цикл проекта по добыче знаний
Деятельность по добыче знаний как проектная деятельность. Способы оценки результатов проекта Data Mining. Основные этапы проекта Data Mining.
Тема 4. Основные подходы к предобработке данных
Потребность в предобработке данных. Сложность предобработки данных. Способы представления различных типов данных. Проблема отклонений.

#### Раздел 3. Основные алгоритмы Data Mining

Тема 5. Искусственные нейронные сети. Типология. Области применения
Биологические нейронные сети. Понятие искусственной нейронной сети (ИНС). Архитектура ИНС. Классификация ИНС. Области применения ИНС. Проблема выбора топологии нейронной сети под решаемую задачу. Применение ИНС для решения задач Data Mining.
Тема 6. Деревья решений. Алгоритмы построения. Области применения
Понятие дерева решения. Основные алгоритмы построения. Проблема отсечения ветвей. Применение деревьев решений для задач Data Mining.
Тема 7. Генетические алгоритмы. Области применения
Естественный отбор и генетическое наследование. Представление генетической информации. Генетические операторы. Применение генетических алгоритмов. Совместное использование генетических алгоритмов и искусственных нейронных сетей.
Тема 8. Метод опорных векторов. Области применения
Основные понятия метода опорных векторов. Основные алгоритмы. Проблема поиска «ядра». Сравнение с ИНС.
Тема 9. Алгоритмы поиска ассоциативных правил и временных шаблонов
Задача продуктовой корзины, ее обобщения, подходы к решению. Алгоритм Apriori.
Тема 10. Аппарат математической статистики для решения задач Data Mining
Логистическая регрессия и ROC-анализ. Применение аппарата проверки гипотез для решения задачи Data Mining.
Тема 11. Алгоритмы решения задач кластеризации
Классификация алгоритмов решения задачи кластеризации. Сравнение различных подходов к решению задачи кластеризации.

#### Искусственные нейронные сети

Рассмотрено понятие биологической нейронной сети и понятие искусственной нейронной сети (ИНС). Описана общая архитектура ИНС. Введена классификация ИНС. Рассмотрены основные области применения ИНС. Описана проблема выбора топологии нейронной сети под решаемую задачу.

#### Деревья решений

Введено понятие дерева решения (ДР). Описаны основные алгоритмы построения ДР. Описана проблема отсечения ветвей, предложены решения.

#### Генетические алгоритмы

Естественный отбор и генетическое наследование. Представление генетической информации. Генетические операторы. Применение генетических алгоритмов. Совместное использование генетических алгоритмов и искусственных нейронных сетей.

**Метод опорных векторов**

Основные понятия метода опорных векторов. Основные алгоритмы. Проблема поиска «ядра». Сравнение с ИНС.

**Алгоритмы поиска ассоциативных правил и временных шаблонов**

Задача продуктовой корзины, ее обобщения, подходы к решению. Алгоритм Apriori.

**Аппарат математической статистики для решения задач Data Mining**

Логистическая регрессия и ROC-анализ. Применение аппарата проверки гипотез для решения задачи Data Mining.

**Алгоритмы решения задач кластеризации**

Классификация алгоритмов решения задачи кластеризации. Сравнение различных подходов к решению задачи кластеризации.

**Раздел 4. Text Mining**

Тема 12. Основные понятия и история Text Mining

Глоссарий основных понятий. Причины появления и развития Text Mining.

Тема 13. Основные задачи Text Mining и методы их решения

Типы задач Text Mining. Проблема поиска информации. Подходы к классификации и кластеризации текстов

**Раздел 5. Современные инструментальные средства Data Mining**

Тема 13. Современные инструментальные средства Data Mining

Обзор современных инструментальных средств Data Mining. Тенденции развития инструментальных средств Data Mining.

Тема 14. Microsoft SQL Server Analysis Services

Подход Microsoft SQL Server к анализу данных. Возможности Analysis Services. DMX как пример языка манипуляции моделью Data Mining. PMML как пример языка описания модели Data Mining.

**Итоговое контрольное мероприятие**

Список вопросов:

1. Основные понятия и история развития систем Data Mining
2. Задачи добычи знаний и методы их решений. Примеры
3. Data Mining как проект. Жизненный цикл проекта по добыче знаний. Основные этапы проекта Data Mining.
4. Потребность в предобработке данных. Сложность предобработки данных. Способы представления различных типов данных.
5. Задача поиска отклонений. Описание. Способы решений.
6. Искусственные нейронные сети. Типология. Области применения
7. Деревья решений. Алгоритмы построения. Области применения
8. Генетические алгоритмы. Области применения
9. Метод опорных векторов. Области применения
10. Алгоритмы поиска ассоциативных правил и временных шаблонов. Алгоритм Apriori.
11. Логистическая регрессия и ROC-анализ. Применение аппарата проверки гипотез для решения задачи

Data Mining.
12. Алгоритмы решения задач кластеризации
13. Основные понятия и история Text Mining. Основные задачи Text Mining и
14. Современные инструментальные средства Data Mining. Microsoft SQL

## 6. Методические указания для обучающихся по освоению дисциплины

Освоение дисциплины требует систематического изучения всех тем в той последовательности, в какой они указаны в рабочей программе.

Основными видами учебной работы являются аудиторные занятия. Их цель - расширить базовые знания обучающихся по осваиваемой дисциплине и систему теоретических ориентиров для последующего более глубокого освоения программного материала в ходе самостоятельной работы. Обучающемуся важно помнить, что контактная работа с преподавателем эффективно помогает ему овладеть программным материалом благодаря расстановке необходимых акцентов и удержанию внимания интонационными модуляциями голоса, а также подключением аудио-визуального механизма восприятия информации.

Самостоятельная работа преследует следующие цели:

– закрепление и совершенствование теоретических знаний, полученных на лекционных занятиях;

- формирование навыков подготовки текстовой составляющей информации учебного и научного назначения для размещения в различных информационных системах;

- совершенствование навыков поиска научных публикаций и образовательных ресурсов, размещенных в сети Интернет;

- самоконтроль освоения программного материала.

Обучающемуся необходимо помнить, что результаты самостоятельной работы контролируются преподавателем во время проведения мероприятий текущего контроля и учитываются при промежуточной аттестации.

Обучающимся с ОВЗ и инвалидов предоставляется возможность выбора форм проведения мероприятий текущего контроля, альтернативных формам, предусмотренным рабочей программой дисциплины. Предусматривается возможность увеличения в пределах 1 академического часа времени, отводимого на выполнение контрольных мероприятий.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации.

При проведении текущего контроля применяются оценочные средства, обеспечивающие передачу информации, от обучающегося к преподавателю, с учетом психофизиологических особенностей здоровья обучающихся.

## 7. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

При самостоятельной работе обучающимся следует использовать:

– конспекты лекций;

– литературу из перечня основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля);

– текст лекций на электронных носителях;

– ресурсы информационно-телекоммуникационной сети "Интернет", необходимые для освоения дисциплины;

– лицензионное и свободно распространяемое программное обеспечение из перечня информационных технологий, используемых при осуществлении образовательного процесса по дисциплине;

– методические указания для обучающихся по освоению дисциплины.

## 8. Перечень основной и дополнительной учебной литературы

**Основная:**

1. Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. — 3-е изд. — Москва, Саратов : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. — 469 c. — ISBN 978-5-4497-0289-0. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/89404.html

**Дополнительная:**

1. Федин, Ф. О. Анализ данных. Часть 2. Инструменты Data Mining : учебное пособие / Ф. О. Федин, Ф. Ф. Федин. — Москва : Московский городской педагогический университет, 2012. — 308 c. — ISBN 2227-8397. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. http://www.iprbookshop.ru/26445

## 9. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

**https://intuit.ru/studies/courses/6/6/info** Учебный курс по Data Mining

## 10. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

Образовательный процесс по дисциплине **Добыча знаний: теоретические основы и инструментальные средства Data mining** предполагает использование следующего программного обеспечения и информационных справочных систем:
доступ в режиме on-line в Электронную библиотечную систему (ЭБС);
доступ в электронную информационно-образовательной среду университета.
Необходимое лицензионное и(или) свободно распространяемое программное обеспечение:
Microsoft Visual Studio
Пакет JetBrains
транслятор экрана VNC-viewer

При освоении материала и выполнения заданий по дисциплине рекомендуется использование материалов, размещенных в Личных кабинетах обучающихся ЕТИС ПГНИУ (**student.psu.ru**).

При организации дистанционной работы и проведении занятий в режиме онлайн могут использоваться:

система видеоконференцсвязи на основе платформы BigBlueButton (https://bigbluebutton.org/).

система LMS Moodle (http://e-learn.psu.ru/), которая поддерживает возможность использования текстовых материалов и презентаций, аудио- и видеоконтент, а так же тесты, проверяемые задания, задания для совместной работы.

система тестирования Indigo (https://indigotech.ru/).

## 11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Для лекционных занятий требуется аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) с соответствующим программным обеспечением, меловой (и) или маркерной доской.
Для проведения лабораторных занятий - меловая и (или) маркерная доска, компьютерный класс (аппаратное и программное обеспечение определено в паспортах компьютерных классов)
Для групповых (индивидуальных) консультаций - аудитория, оснащенная меловой (и) или маркерной доской.
Для проведения текущего контроля - аудитория, оснащенная меловой (и) или маркерной доской.
Самостоятельная работа студентов: аудитория, оснащенная компьютерной техникой с возможностью подключения к сети «Интернет», с обеспеченным доступом в электронную информационно-образовательную среду университета, помещения Научной библиотеки ПГНИУ.

Помещения научной библиотеки ПГНИУ для обеспечения самостоятельной работы обучающихся:

1. Научно-библиографический отдел, корп.1, ауд. 142. Оборудован 3 персональными компьютера с доступом к локальной и глобальной компьютерным сетям.

2. Читальный зал гуманитарной литературы, корп. 2, ауд. 418. Оборудован 7 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

3. Читальный зал естественной литературы, корп.6, ауд. 107а. Оборудован 5 персональными компьютерами с доступом к локальной и глобальной компьютерным сетям.

4. Отдел иностранной литературы, корп.2 ауд. 207. Оборудован 1 персональным компьютером с доступом к локальной и глобальной компьютерным сетям.

5. Библиотека юридического факультета, корп.9, ауд. 4. Оборудована 11 персональными компьютерами с доступом к  локальной и глобальной компьютерным сетям.

6. Читальный зал географического факультета, корп.8, ауд. 419. Оборудован 6 персональными компьютерами с доступом к  локальной и глобальной компьютерным сетям.

Все компьютеры, установленные в помещениях научной библиотеки, оснащены следующим программным обеспечением:

Операционная система ALT Linux;

Офисный пакет Libreoffice.

Справочно-правовая система «КонсультантПлюс»

**Фонды оценочных средств для аттестации по дисциплине**
**Добыча знаний: теоретические основы и инструментальные средства Data mining**

**Планируемые результаты обучения по дисциплине для формирования компетенции. Индикаторы и критерии их оценивания**

**ПК.4**

**Способен интегрировать разработанное системное программное обеспечение**

| Индикатор | Планируемые результаты обучения | Критерии оценивания результатов обучения |
|---|---|---|
| **ПК.4.2**<br>Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Знать:<br>- основные принципы анализа данных и добычи знаний;<br>- основные понятия, архитектуры и алгоритмы обучения искусственных нейронных сетей;<br>- назначение и структуры генетических алгоритмов.<br>Уметь:<br>- определять тип задачи добычи знаний;<br>- выбирать адекватный способ решения для задачи добычи знаний. | **Неудовлетворител**<br>Не сформированы знания:<br>- основных принципов анализа данных и добычи знаний;<br>- основных понятий, архитектуры и алгоритмов обучения искусственных нейронных сетей;<br>- назначения и структуры генетических алгоритмов.<br>Не умеет:<br>- определять тип задачи добычи знаний;<br>- выбирать адекватный способ решения для задачи добычи знаний.<br>**Удовлетворительн**<br>Сформированы поверхностные знания:<br>- основных принципов анализа данных и добычи знаний;<br>- основных понятий, архитектуры и алгоритмов обучения искусственных нейронных сетей;<br>- назначения и структуры генетических алгоритмов.<br>В целом умеет:<br>- определять тип задачи добычи знаний;<br>- выбирать адекватный способ решения для задачи добычи знаний.<br>**Хорошо**<br>Сформированы систематические, но содержащие отдельные пробелы знания:<br>- основных принципов анализа данных и добычи знаний;<br>- основных понятий, архитектуры и алгоритмов обучения искусственных нейронных сетей;<br>- назначения и структуры генетических алгоритмов.<br>Умеет на достаточном уровне: |

| Индикатор | Планируемые результаты обучения | Критерии оценивания результатов обучения |
|---|---|---|
| | | **Хорошо**<br>- определять тип задачи добычи знаний;<br>- выбирать адекватный способ решения для задачи добычи знаний.<br>**Отлично**<br>Сформированы систематические знания:<br>- основных принципов анализа данных и добычи знаний;<br>- основных понятий, архитектуры и алгоритмов обучения искусственных нейронных сетей;<br>- назначения и структуры генетических алгоритмов.<br>В совершенстве умеет:<br>- определять тип задачи добычи знаний;<br>- выбирать адекватный способ решения для задачи добычи знаний. |

## Оценочные средства текущего контроля и промежуточной аттестации

Схема доставки : Базовая

**Вид мероприятия промежуточной аттестации :** Зачет
**Способ проведения мероприятия промежуточной аттестации :** Оценка по дисциплине в рамках промежуточной аттестации определяется на основе баллов, набранных обучающимся на контрольных мероприятиях, проводимых в течение учебного периода.
**Максимальное количество баллов :** 100

**Конвертация баллов в отметки**

**«отлично» -** от 81 до 100
**«хорошо» -** от 61 до 80
**«удовлетворительно» -** от 46 до 60
**«неудовлетворительно» / «незачтено»** менее 46 балла

| Компетенция (индикатор) | Мероприятие текущего контроля | Контролируемые элементы результатов обучения |
|---|---|---|
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Раздел 2. Предобработка данных как необходимый этап добычи знаний **Письменное контрольное мероприятие** | Способность решать задачи предобработки данных |
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Искусственные нейронные сети **Письменное контрольное мероприятие** | Способность решать практические задачи с помощью нейронных сетей |
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Деревья решений **Защищаемое контрольное мероприятие** | Способность решать практические задачи с использованием деревьев решений |
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Алгоритмы поиска ассоциативных правил и временных шаблонов **Защищаемое контрольное мероприятие** | Способность на практике использовать алгоритмы поиска ассоциативных правил |

| Компетенция (индикатор) | Мероприятие текущего контроля | Контролируемые элементы результатов обучения |
|---|---|---|
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Аппарат математической статистики для решения задач Data Mining **Защищаемое контрольное мероприятие** | Способность на практике использовать методы математической статистики для решения задач Data Mining |
| **ПК.4.2** Внедряет разработанное программное обеспечение для высокопроизводительных вычислительных комплексов и систем, базирующихся на знаниях | Итоговое контрольное мероприятие **Итоговое контрольное мероприятие** | студент должен быть способен:- определять тип задачи добычи знаний;- выбирать адекватный способ решения для задачи добычи знаний;- разрабатывать решения задач добычи знаний с применением алгоритмов на основе искусственных нейронных сетей, деревьев решений, метода опорных векторов, генетических алгоритмов, алгоритмов кластеризации;- решать задач добычи знаний с применением алгоритмов на основе искусственных нейронных сетей, деревьев решений, метода опорных векторов, генетических алгоритмов, алгоритмов кластеризации; разрабатывать учебно-методические комплексы для электронного и мобильного обучения по темам, связанным с добычей знаний;работать в международных проектах по тематике Data Mining пользоваться материалами на иностранном языке при решении задачи Data Mining. Иметь представление о региональных и мироваых потребностях в специалистах по решению задач Data Mining |

## Спецификация мероприятий текущего контроля

### Раздел 2. Предобработка данных как необходимый этап добычи знаний

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **10**
Проходной балл: **5**

| Показатели оценивания | Баллы |
|---|---|
| Знание подходов к работе с данными с выбросами | 2.5 |
| Знание места предобработки данных в общем наборе работ проекта Data Mining | 2.5 |
| Знание подходов к работе с данными с шумом | 2.5 |
| Знание подходов к работе с данными с пропусками | 2.5 |

### Искусственные нейронные сети

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **15**
Проходной балл: **7**

| Показатели оценивания | Баллы |
|---|---|
| Знать алгоритм обучения многослойного персептрона | 5 |
| Знать алгоритм обучения сети Кохонена | 5 |
| Уметь решать задачи кластеризации с помощью нейронных сетей | 3 |
| Знать типологию нейронных сетей. Уметь выбирать нейронную сеть под задачу. | 2 |

### Деревья решений

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **10**
Проходной балл: **5**

| Показатели оценивания | Баллы |
|---|---|
| Знание по крайней мере 1ого алгоритма построения деревьев решений | 5 |
| Навык практической реализации деревьев решений | 3 |
| Знание места деревьев решений в задачах Data Mining | 2 |

### Алгоритмы поиска ассоциативных правил и временных шаблонов

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **15**
Проходной балл: **7**

| Показатели оценивания | Баллы |
|---|---|
| Практический навык реализации алгоритма Apriori | 8 |
| Знание места алгоритмов поиска ассоциативных правил | 7 |

### Аппарат математической статистики для решения задач Data Mining

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **10**
Проходной балл: **5**

| Показатели оценивания | Баллы |
|---|---|
| Понимание места методов математической статистики в задачах Data Mining | 5 |
| Умение применять регрессионный анализ | 3 |
| Умение применять корреляционный анализ | 2 |

## Итоговое контрольное мероприятие

Продолжительность проведения мероприятия промежуточной аттестации: **2 часа**
Условия проведения мероприятия: **в часы самостоятельной работы**
Максимальный балл, выставляемый за мероприятие промежуточной аттестации: **40**
Проходной балл: **17**

| Показатели оценивания | Баллы |
|---|---|
| Студент способен разрабатывать решения задач добычи знаний с применением алгоритмов на основе искусственных нейронных сетей, деревьев решений, метода опорных векторов, генетических алгоритмов, алгоритмов кластеризации | 23 |
| Студент должен быть способен:- определять тип задачи добычи знаний;- выбирать адекватный способ решения для задачи добычи знаний | 10 |
| Студент способен разрабатывать учебно-методические комплексы для электронного и мобильного обучения по темам, связанным с добычей знаний | 4 |
| Студент способен пользоваться материалами на иностранном языке при решении задачи Data Mining | 3 |