Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой

Борисов Дмитрий Николаевич

Кафедра информационных систем

03.05.2023

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.07 Интеллектуальный анализ данных

1. Код и наименование направления подготовки/специальности:

02.04.01 Математика и компьютерные науки

2. Профиль подготовки/специализация:

Компьютерное моделирование и искусственный интеллект

3. Квалификация (степень) выпускника:

Магистратура

4. Форма обучения:

Очная

5. Кафедра, отвечающая за реализацию дисциплины:

Кафедра информационных систем

6. Составители программы:

Сычев Александр Васильевич, кандидат физико-математических наук, доцент кафедры информационных систем

7. Рекомендована: *НМС ФКН 03.05.2023, протокол № 7*

8. Учебный год:

2023-2024

9.Цели и задачи учебной дисциплины:

Целью изучения дисциплины является практическое знакомство студентов с современными технологиями анализа многомерных данных, включая математические модели, алгоритмы и программные средства, используемые для решения основных задач анализа многомерных данных: классификации, кластеризации и других.

Задачи учебной дисциплины:

- изучение структуры систем поддержки принятия решений (СППР), ее отличий от систем оперативной обработки данных (OLTP), многомерной модели данных OLAP и принципов работы с реализацией данной модели (на примере OLAP Analysis Services);

- изучение концепции Data Mining и основных задач, реализующих эту концепцию: классификацию, кластериацию, поиск ассоциативных правил и других;

- изучение структуры процесса Data Mining и связанных с ним стандартов;

- практическое изучение задач Data Mining и методов их решения с помощью программных пакетов

RapidMiner, Matlab и других.

10. Место учебной дисциплины в структуре ООП:

Учебная дисциплина относится к части блока Б1, формируемой участниками образовательных отношений, курс по выбору

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями выпускников) и индикаторами их достижения:

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-1 Способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий.	ПК-1.1 Обладает фундаментальными знаниями, полученными в области математических и (или) естественных наук, программирования и информационных технологий	Знает основные задачи, решаемые в рамках направления информационных технологий Data Mining, их характеристики и основные алгоритмы их решения
ПК-8 Способен создавать и исследовать новые математические модели в естественных науках, промышленности и бизнесе, с учетом возможностей современных информационных технологий, программирования и компьютерной техники	ПК-8.1 Знает основные методы проектирования и производства программного продукта, принципы построения, структуры и приемы работы с инструментальными средствами, поддерживающими создание программных продуктов и программных комплексов, их сопровождения, администрирования и развития (эволюции)	Знает структуру процесса Data Mining, ключевые модули и операторы для конструирования процесса анализа данных в среде RapidMiner

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-9 Способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования.	ПК-9.1 Владеет современными методами разработки и реализации алгоритмов математических моделей на базе языков и пакетов прикладных программ моделирования	Имеет навыки реализации основных алгоритмов для различных моделей классификации в среде RapidMiner
ПК-9 Способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования.	ПК-9.2 Умеет разрабатывать и реализовывать алгоритмы математических моделей на базе языков и пакетов прикладных программ моделирования	Умеет реализовывать основные алгоритмы для различных моделей классификации в среде RapidMiner
ПК-9 Способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования.	ПК-9.3 Имеет практический опыт разработки и реализации алгоритмов на базе языков и пакетов прикладных программ моделирования	Имеет опыт реализации основных алгоритмов для различных моделей классификации в среде RapidMiner на примере нескольких типовых наборов данных

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-8 Способен создавать и исследовать новые математические модели в естественных науках, промышленности и бизнесе, с учетом возможностей современных информационных технологий, программирования и компьютерной техники	ПК-8.2 Умеет использовать методы проектирования и производства программного продукта, принципы построения, структуры и приемы работы с инструментальными средствами, поддерживающими создание программного продукта	Умеет конструировать процесс анализа данных для решения типовых задач Data Mining в среде RapidMiner, а также настраивать параметры модулей, из которых состоит процесс
ПК-8 Способен создавать и исследовать новые математические модели в естественных науках, промышленности и бизнесе, с учетом возможностей современных информационных технологий, программирования и компьютерной техники	ПК-8.3 Имеет практический опыт применения указанных выше методов и технологий	Имеет опыт конструирования процесса анализа данных для решения задач Data Mining в среде RapidMiner, а также настройки параметров модулей, из которых состоит процесс, на примере типовых наборов данных
ПК-1 Способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий.	ПК-1.2 Умеет находить, формулировать и решать стандартные задачи в собственной научно-исследовательской деятельности в области программирования и информационных технологий	Умеет решать стандартные задачи анализа многомерных данных на основе технологий Data Mining

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-1 Способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий.	ПК-1.3 Имеет практический опыт научно-исследовательской деятельности в области программирования и информационных технологий	Имеет опыт решения стандартных задач анализа многомерных данных на основе технологий Data Mining на примере типовых наборов данных

12. Объем дисциплины в зачетных единицах/час:

3/108

Форма промежуточной аттестации:

Экзамен

13. Трудоемкость по видам учебной работы

Вид учебной работы	Семестр 1	Всего
Аудиторные занятия	36	36
Лекционные занятия	18	18
Практические занятия		0
Лабораторные занятия	18	18
Самостоятельная работа	36	36
Курсовая работа		0
Промежуточная аттестация	36	36
Часы на контроль	36	36
Всего	108	108

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1	Введение в Data Mining.	Основные определения, предметная область, актуальность и приложения.	Онлайн курс на edu.vsu.ru

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
2	Системы поддержки принятия решений и хранилища данных	Системы поддержки принятия решений (СППР). Классы задач анализа данных в СППР. Обобщенная архитектура СППР. OLTP и СППР - сравнение. Понятие об интеллектуальном анализе данных и Data Mining. Концепция хранилища данных (ХД). Свойства ХД. Архитектура систем ХД. Структура СППР с физическим ХД. Структура СППР с виртуальным ХД. Витрина данных (ВД). Структура СППР с самостоятельными ВД. Архитектура ХД.	Онлайн курс на edu.vsu.ru
3	OLAP-системы	Многомерная модель данных. Основные элементы OLAP. Представление данных в виде гиперкуба. Операции над гиперкубом. Определение OLAP-систем. Двенадцать правил Кодда для OLAP. Дополнительные правила Кодда. Тест FASMI. Архитектура OLAP-систем: MOLAP, ROLAP, HOLAP.	Онлайн курс на edu.vsu.ru

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
4	Задачи Data Mining	Правила классификации. Деревья решений. Конструирование дерева решений. Критерий расщепления. Алгоритмы: байесовский, CART, C4.5. Алгоритмы классификации: метод "ближайшего соседа", метод построения математических функций, метод опорных векторов (SVM). Постановка задачи. Методы прогнозирования временных рядов. Методы поиска ассоциативных правил. Алгоритм Apriori и его разновидности. Понятие кластерного анализа. Характеристики кластеров. Методы кластерного анализа: иерархические и неиерархические. Иерархические методы кластеризации. Агломеративные и дивизимные методы. Дендрограммы. Метрики подобия объектов. Проверка качества кластеризации. Достоинства и недостатки иерархических алгоритмов. Алгоритм k-средних. Достоинства и недостатки алгоритма. Проверка качества кластеризации. Этапы кластерного анализа. Сложности и проблемы кластерного анализа. Сравнение иерархических и неиерархических методов кластеризации. Этапы визуального анализа данных. Характеристики средств визуализации данных. Типы методов визуализации. Визуализация Data Mining моделей. Методы визуализации. Параллельные координаты. "Лица Чернова". Рекомендации по использованию средств визуализации. Качество визуализации. Основные тенденции в области визуализации.	Онлайн курс на edu.vsu.ru
5	Стандарты Data Mining	Аспекты стандартизации Data Mining. Стандарты унификации интерфейсов. Стандарт CWM. Стандарт CRISP. Методология CRISP-DM. SEMMA методология. Стандарт PMML	Онлайн курс на edu.vsu.ru
6	Процесс Data Mining	Этапы процесса Data Mining. Анализ предметной области. Постановка задачи. Подготовка данных. "Грязные" данные. Очистка данных. Этапы очистки данных. Моделирование. Организационные факторы Data Mining. Человеческие факторы. Роли в Data Mining.	Онлайн курс на edu.vsu.ru

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
1	Введение в Data Mining	1			3	4
2	Системы поддержки принятия решений и хранилища данных	2			6	8
3	OLAP-системы	3		4	8	15
4	Задачи Data Mining	8		14	45	67
5	Стандарты Data Mining	1			5	6
6	Процесс Data Mining	3			5	8
		18	0	18	72	108

14. Методические указания для обучающихся по освоению дисциплины

1) При изучении дисциплины рекомендуется использовать следующие средства:

- рекомендуемую основную и дополнительную литературу;
- методические указания и пособия;
- контрольные задания для закрепления теоретического материала;
- электронные версии учебников и методических указаний для выполнения лабораторно-практических работ.

2) Для лучшего усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала.
3) При проведении лабораторных занятий обеспечивается практическая демонстрация материалов лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и технологий анализа многомерных данных, излагаемых в рамках лекций.
4) При переходе на дистанционный режим обучения для создания электронных курсов, чтения лекций онлайн и проведения лабораторно-практических занятий используются информационные ресурсы образовательного портала "Электронный университет ВГУ (https://edu.vsu.ru/course/view.php?id=2488), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

№ п/п	Источник
1	Зайцев, К. С. Применение методов Data Mining для поддержки процессов управления IT-услугами : учебное пособие / К. С. Зайцев. — Москва : НИЯУ МИФИ, 2009. — 96 с. — ISBN 978-5-7262-1150-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/75805
2	Макшанов, А. В. Большие данные. Big Data : учебник для вузов / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — Санкт-Петербург : Лань, 2021. — 188 с. — ISBN 978-5-8114-6810-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165835

б) дополнительная литература:

№ п/п	Источник
1	Юре, Л. . Анализ больших наборов данных [Электронный ресурс] / Юре Л. , Ананд Р. , Джеффри Д. У. — Москва : ДМК Пресс, 2016 .— 498 с. — <URL:https://e.lanbook.com/book/93571>
2	Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. — 2-е изд. — Москва : ИНТУИТ, 2016. — 470 с. — ISBN 978-5-94774-819-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/100582

в) информационные электронно-образовательные ресурсы:

№ п/п	Источник
1	Чубукова И.А. Data Mining (Электронный курс) / Интернет-Университет Информационных Технологий.2006. - [Электрон. ресурс]. Режим доступа: http://www.intuit.ru/department/database/datamining/
2	Электронный курс на образовательном портале «Электронный университет ВГУ». –https://edu.vsu.ru/course/view.php?id=2488

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Онлайн курс https://edu.vsu.ru/course/view.php?id=2488

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

Программные пакеты: RapidMiner, Matlab

18. Материально-техническое обеспечение дисциплины:

1. Лекционная аудитория, оборудованная мультимедийным проектором.

2. Компьютерный класс факультета для проведения лабораторных занятий.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
1	1-4, 6	ПК-1	ПК-1.1	КИМы
2	4	ПК-8	ПК-8.1	КИМы, Практические задания
3	4,5	ПК-9	ПК-9.1	Практические задания
4	4	ПК-9	ПК-9.2	Практические задания
5	4	ПК-9	ПК-9.3	Практические задания
6	4	ПК-8	ПК-8.2	Практические задания
7	4	ПК-8	ПК-8.3	Практические задания
8	2-4	ПК-1	ПК-1.2	КИМы, Практические задания
9	3,4	ПК-1	ПК-1.3	Практические задания

Промежуточная аттестация

Форма контроля - Экзамен

Оценочные средства для промежуточной аттестации

- *оценка «зачтено» выставляется в том случае, если студентом выполнено итоговое практическое задание, подготовлен отчет и при устном собеседовании студент дал частичный ответ на оба вопроса в КИМе;*
- *оценка «незачтено» выставляется в том случае, если студентом не выполнено итоговое практическое задание или при устном собеседовании студент дал неправильный ответ на вопросы в КИМе.*

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Тестовые задания - 1 балл за каждый правильный тест (максимум).

Компетенция ПК-1

1. Выберите правильное соответствие для архитектуры OLAP-систем:

a) Для реализации используют многомерные БД

b) Для реализации используют реляционные БД

c) Для реализации используют и многомерные, и реляционные БД

a) HOLAP
b) MOLAP
c) ROLAP

2. Выберите правильное соответствие для деревьев решений:
a) Внутренние узлы дерева решений
b) Конечные узлы дерева (листья)
c) Ветвь дерева, идущая от внутреннего узла
d) Объединенная информация об атрибутах расщепления и предикатах расщепления в узле

a) атрибуты расщепления
b) критерий расщепления
c) значения зависимой категориальной переменной (метки класса)
d) предикат расщепления

3. Выберите правильное соответствие для задач Data Mining:
a) Состоит в определении класса объекта по его характеристикам. Множество классов известно заранее.
b) Позволяет определить по известным характеристикам объекта значение некоторого его параметра. Значением параметра в отличие от задачи классификации является не конечное множество классов, а множество действительных чисел.
c) Сводится к нахождению частных зависимостей между объектами или событиями. Зависимости представляются в виде правил.
d) Заключается в поиске независимых групп и их характеристик во всем множестве анализируемых данных.
e) На основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.
f) Предсказание непрерывных значений признака объекта.
g) Создается графический образ анализируемых данных.

a) Поиск ассоциативных правил
b) Задачи прогнозирования
c) Задача регрессии
d) Задача оценивания
e) Задача кластеризации
f) Задача классификации
g) Задача визуализации

4. Выберите правильное соответствие для методов и задач Data Mining:
a) Метод K-средних.
b) Метод опорных векторов.
c) Метод "ближайших соседей".
d) Метод деревьев решений.
e) Метод Apriori.
d) "Лица Чернова".

a) Поиск ассоциативных правил
b) Задача кластеризации
c) Задача классификации
d) Задача классификации
e) Задача классификации
f) Задача визуализации

5. Выберите правильное соответствие для оптимизации построения дерева решений:
a) Определяется целесообразность разбиения узла.

b) Построение заканчивается, если достигнута заданная глубина.
c) Ветвления продолжаются до того момента, пока все конечные узлы дерева не будут чистыми или будут содержать не более чем заданное число объектов.

a) Ограничение глубины дерева
b) Задание минимального количества примеров
c) "Ранняя остановка" (prepruning)

6. Выберите правильное соответствие для основных элементов OLAP:
a) Числовая величина которая располагается в ячейках гиперкуба
b) Множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя.
c) Атомарная структура куба, соответствующая полному набору конкретный значений измерений

a) Ячейка
b) Факт
c) Измерение

7. Выберите правильные утверждения, относящиеся к временным рядам и их прогнозированию:
a) Неслучайная функцию, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд.
b) Частота, с которой делается новый прогноз.
c) Число периодов в будущем, которые покрывает прогноз.
d) Периодически повторяющаяся компонента временного ряда. Через примерно равные промежутки времени форма кривой, которая описывает поведение зависимой переменной, повторяет свои характерные очертания.
e) Основная единица времени, на которую делается прогноз.

a) Тренд
b) Сезонная составляющая временного ряда
c) Период прогнозирования
d) Интервал прогнозирования
e) Горизонт прогнозирования

8. Выберите правильное соответствие:
a) Множество событий (операций), которые произошли одновременно.
b) Вероятность того, что из события A следует событие B.
c) Набор, у которого поддержка элементов выше определенного пользователем минимального значения поддержки
d) Процент транзакций из всего набора, которые содержат одновременно наборы элементов A и B.

a) Часто встречающийся набор
b) Транзакция
c) Поддержка правила
d) Достоверность

9. Выберите правильное соответствие для характеристик кластеров:
a) Среднее геометрическое место точек в пространстве переменных.
b) Максимальное расстояние точек от центра кластера.
c) Объект, который по мере сходства может быть отнесен к нескольким кластерам.

a) Радиус кластера
b) Спорный объект
c) Центр кластера

10. Выберите правильное соответствие для характеристик кластеров:
a) Характеризует время, которое требуется на создание модели и ее использование
b) Означает устойчивость к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями данных
c) Обеспечивает возможность понимания модели аналитиком
d) Предусматривает возможность работы при наличии в наборе данных шумов и выбросов

a) Интерпретируемость
b) Надежность
c) Робастность
d) Скорость

11. Выберите правильные утверждения, относящиеся к определению дендрограммы:
a) Используют, если необходимо отобразить соотношение частей и целого, т.е. для анализа состава или структуры явлений
b) Описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров
c) Позволяет отобразить тенденцию, передать изменения какого-либо признака во времени
d) Представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии
e) Представляет собой график отклонения значений, прогнозируемых при помощи модели, от реальных
f) Содержит n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров

12. Выберите правильные утверждения, относящиеся к задаче классификации:
a) Может быть многомерной (по двум и более признакам)
b) Может быть одномерной (по одному признаку)
c) Может быть только трехмерной (по трем признакам)
d) Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое
e) Относится к стратегии обучения с учителем
f) Решение задачи состоит из двух этапов: конструирования и модели и ее использования
g) Целью задачи является разбиение совокупности объектов на однородные группы.

13. Выберите утверждения, не относящиеся к задаче кластеризации:
a) В данной задаче классы изучаемого набора данных заранее не предопределены.
b) Набор исходных данных (или выборку данных) разбивают на два множества: обучающее и тестовое.
c) Относится к стратегии обучения с учителем.
d) Решение задачи состоит из двух этапов: конструирования и модели и ее использования.
e) Синонимами для названия задачи являются "обучение без учителя" и "таксономия".
f) Целью задачи является разбиение совокупности объектов на однородные группы.

14. Выберите правильные утверждения, относящиеся к задаче прогнозирования:
a) В данной задаче предсказывается класс зависимой переменной.
b) В данной задаче предсказываются числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему).
c) Задача заключается в установлении функциональной зависимости между зависимыми и независимыми переменными.
d) Основой для решения задачи является историческая информация, хранящаяся в базе данных в виде временных рядов.
e) Решение задачи направлено на определение тенденций динамики конкретного объекта или

события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем.

15. Выберите правильные утверждения, относящиеся к методу опорных векторов:
a) Для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.
b) При помощи данного метода не могут решаться задачи бинарной классификации.
c) Метод сильно подвержен проблеме переобучения.
d) Метод работает с любым количеством измерений.
e) Метод относится к группе граничных методов. Он определяет классы при помощи границ областей.
f) Для метода характерны устойчивые решения, нет проблем с локальными минимумами.

16. Выберите правильные утверждения, относящиеся к факторному анализу:
a) Это разложение графа на непересекающиеся по ребрам остовные подграфы специального вида.
b) Это метод, применяемый для изучения взаимосвязей между значениями переменных.
c) Это декомпозиция объекта (например, числа, полинома или матрицы) в произведение других объектов, или факторов, которые, будучи перемноженными, дают исходный объект.
d) Преследует цель сокращения числа переменных.
e) Преследует цель классификации переменных, т.е. определения структуры взаимосвязей между переменными.
f) Опирается на гипотезу о том, что анализируемые переменные являются косвенными проявлениями сравнительно небольшого числа неких скрытых факторов.

17. Опишите правильную последовательность действий для метода "ближайшего соседа"
a) Сопоставление имеющейся информации о задаче с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев
b) Сбор подробной информации о поставленной задаче
c) Проверка корректности каждого вновь полученного решения
d) Занесение детальной информации о новом прецеденте в базу прецедентов
e) Выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов
f) Адаптация выбранного решения к текущей проблеме, если это необходимо

Компетенция ПК-8

18. Выберите правильное соответствие для задач анализа данных в СППР:
a) СППР осуществляет поиск необходимых данных. Выполняются заранее определенные запросы
b) СППР производит группирование и обобщение данных в любом виде, необходимом аналитику
c) СППР осуществляет поиск функциональных и логических закономерностей в накопленных данных, построение моделей и правил, которые объясняют найденные закономерности и/или прогнозируют развитие некоторых процессов

a) Оперативно-аналитический
b) Информационно-поисковый
c) Интеллектуальный

19. Выберите правильное соответствие для средств визуализации данных:
a) Одномерные массивы, временные ряды
b) Точки двумерных графиков, географические координаты
c) Финансовые показатели, результаты экспериментов
d) Газетные статьи, web-документы
e) Структуры подчиненности в организации, электронная переписка, гиперссылки документов
f) Информационные потоки, отладочные операции

a) Тексты и гипертексты
b) Одномерные данные
c) Многомерные данные
d) Иерархические и связанные данные
e) Двумерные данные
f) Алгоритмы и программы

20. Выберите правильное соответствие для стандартов Data Mining:
a) CWM
b) JDM
c) CRISP-DM
d) SEMMA
e) PMML

a) Стандарт, описывающий основные фазы (задачи) процесса Data Mining
b) Стандарты интерфейсов для объектных языков программирования для обмена метаданными между различными программными продуктами и репозиториями, участвующими в создании корпоративных СППР
c) Стандарты по хранению и передаче моделей Data Mining
d) Стандарты интерфейсов для объектных языков программирования для обмена метаданными между различными программными продуктами и репозиториями, участвующими в создании корпоративных СППР
e) Стандарт, описывающий основные фазы (задачи) процесса Data Mining

21. Выберите, какие модели используются для представления полученных знаний в Data Mining:
a) Файлы
b) Тексты
c) Таблицы
d) Правила
e) Математические функции
f) Кластеры
g) Инструкции
h) Деревья решений

22. Укажите, какие из приведенных ниже задач относятся к основным задачам Data Mining?
a) Управление
b) Тестирование
c) Регрессия
d) Проектирование
e) Программирование
f) Прогнозирование и оценивание
g) Поиск ассоциативных правил
h) Контроль
i) Кластеризация
j) Визуализация

23. Укажите, какие из приведенных ниже стандартов имеют непосредственное отношение к Data Mining?
a) SOAP/XML
b) SEMMA
c) RDF
d) PMML
e) ITIL

f) HTTP
g) ECMA
h) CWM
i) CRISP-DM

24. Выберите утверждения, которые относятся с виртуальным хранилищам данных:
a) Характерна высокая степень избыточности данных в хранилище.
b) Недоступность хотя бы одного из оперативных источников данных может привести либо к невыполнению аналитического запроса, либо к неверным результатам.
c) Для выполнения аналитического запроса требуется объединение большого числа нормализованных таблиц.
d) Данные не копируются в единое хранилище, а извлекаются, преобразуются и интегрируются непосредственно при выполнении аналитических запросов в оперативной памяти компьютера.
e) Выполняется работа только с текущими, детализированными данными.
f) Время обработки запросов значительно увеличивается.
g) Возможно получение данных за долгий период времени.

25. Укажите правильную последовательность шагов процесса Data Mining:
a) Проверка и оценка моделей
b) Применение модели
c) Построение моделей
d) Постановка задачи
e) Подготовка данных
f) Коррекция и обновление модели
g) Выбор модели
h) Анализ предметной области

Компетенция ПК-9

26. Выберите правильное соответствие для методов кластеризации:

a) Аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации.
b) Методы демонстрируют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации.
c) Отказываются от определения числа кластеров и строят полное дерево вложенных кластеров.
d) Суть методов состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.
e) Методы основаны на итеративных методах дробления исходной совокупности.

a) Неиерархические методы
b) Неиерархические методы
c) Неиерархические методы
d) Иерархические методы
e) Иерархические методы

27. Выберите правильное соответствие для операций над OLAP-кубом:
a) Формирование подмножества многомерного массива данных, соответствующего единственному значению одного или нескольких элементов измерений, не входящих в это подмножество
b) Изменение расположения измерений, представленных в отчете или на отображаемой странице
c) Переход вверх по направлению от детального представления данных к агрегированному
d) Переход вниз по направлению от агрегированного представления данных к детальному

a) Срез

b) Консолидация
c) Детализация
d) Вращение

Вопросы с кратким (вычисляемым) ответом - 1 балл за каждый правильный тест (максимум)

Компетенция ПК-9

1. База данных содержит 8 транзакций на основе элементов {A, B, C, D, E, F}:

tid	items
1	ABC
2	BCD
3	CDE
4	BC
5	CD
6	ABCD
7	ABD
8	EF

укажите значение поддержки для ассоциативного правила A -> C.

2. База данных содержит 8 транзакций на основе элементов {A, B, C, D, E, F}:

tid	items
1	ABC
2	BCD
3	CDE
4	BC
5	CD
6	ABCD
7	ABD
8	EF

укажите значение достоверности для ассоциативного правила A -> C. Рассчитанное значение необходимо привести в % с округлением до целого значения.

3. База данных содержит 8 транзакций на основе элементов {A, B, C, D, E, F}:

tid	items
1	ABC
2	BCD
3	CDE
4	BC
5	CD
6	ABCD
7	ABD
8	EF

укажите значение поддержки для ассоциативного правила A -> B.

4. База данных содержит 8 транзакций на основе элементов {A, B, C, D, E, F}:

tid	items
1	ABC
2	BCD
3	CDE
4	BC
5	CD
6	ABCD

7 ABD
8 EF
укажите значение достоверности для ассоциативного правила A -> B. Рассчитанное значение необходимо привести в % с округлением до целого значения.

5. База данных содержит 8 транзакций на основе элементов {A, B, C, D, E, F}:
tid items
1 ABC
2 BCD
3 CDE
4 BC
5 CD
6 ABCD
7 ABD
8 EF
При минимальной поддержке, равной 2, сколько будет найдено часто встречающихся 3-элементных наборов элементов?

6. Набор данных включает атрибут "возраст". Значения атрибута приведены в порядке возрастания: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
Рассчитайте значения медианы M и среднего A (с округлением до целого значения) для приведенного массива значений.

7. Кластер содержит 3 объекта: A1(2,10); A2(2,5); A3(8,4).
Рассчитайте координаты центроида для данного кластера. Ответ необходимо привести в виде C(x,y). Значения x и y должны быть округлены до целых значений.

8. Кластер содержит 4 объекта: A1(3,8); A2(4,5); A3(8,5); A4(5,2).
Рассчитайте координаты центроида для данного кластера. Ответ необходимо привести в виде C(x,y). Значения x и y должны быть округлены до целых значений.

Вопросы с развернутым ответом

В зависимости от степени полноты ответа можно получить до 3 баллов. За первую часть вопроса можно получить до 1,5 баллов, за вторую - до 1,5 баллов.

Компетенция ПК-1

1. Кратко опишите основные этапы классификации на основе дерева решений. Чем сокращение дерева полезно при построении дерева решений?
2. Опишите метод "ближайшего соседа" для решения задачи классификации. Что представляет собой модель классификации в рамках данного метода?
3. Опишите, в чем заключается проблема переобучения? Какие подходы используются для устранения данной проблемы?

Компетенция ПК-8

4. Какие инструменты используются для визуализации многомерных (4+) данных? Кратко охарактеризуйте наиболее известные из них.
5. Опишите кратко основные этапы процесса решения задачи кластеризации.
6. Охарактеризуйте кратко этапы процесса Data Mining. Какие из этапов являются наиболее трудоемкими? Объясните почему?

Компетенция ПК-9

7. Опишите алгоритм поиска ассоциативных правил Apriori. Какие характеристики используются для оценки полученных правил?

8. Опишите алгоритм кластеризации K-средних (по шагам). Каким образом определяется ожидаемое количество кластеров?

20.2 Промежуточная аттестация

УТВЕРЖДАЮ

заведующий кафедрой *информационных систем*.

_____

__.__.20__

Направление подготовки / специальность 02,04,01 "Математика и компьютерные науки".

Дисциплина Интеллектуальный анализ данных.

Вид контроля зачет.

Контрольно-измерительный материал № 1

1. Хранилища данных (ХД). Свойства ХД. Архитектура систем ХД.
2. Этапы процесса Data Mining .

Преподаватель _____ Сычев А.В.