

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

	Рабочая программа дисциплины (модуля)
по дисциплине:	Интерпретируемые методы классификации и порождения знаний
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составил: С.О. Кузнецов, д-р физ.-мат. наук, профессор

Программа обсуждена на заседании кафедры алгоритмов и технологий программирования 13.05.2024

Аннотация

В курсе излагаются основные результаты из теории упорядоченных множеств и алгебраических решеток, которые могут эффективно использоваться в современных методах анализа (майнинга) данных. Особое внимание уделяется разделу современной прикладной теории решеток, основанному на анализе формальных понятий. Излагаются методы построения решеток понятий как основ формальных таксономий и онтологий предметных областей. На едином алгебраическом языке излагается теория импликативных зависимостей, точных и приближенных, а также связанных с ними функциональных зависимостей. С единых теоретико-порядковых позиций излагается ряд методов машинного обучения и майнинга данных. Важной сквозной темой курса является вычислительная сложность рассматриваемых задач и алгоритмов их решения. Обсуждаются разнообразные практические приложения рассматриваемых методов.

1. Цели и задачи

Цель дисциплины

Данный курс позволит студентам овладеть математическими основами важнейшей области разработки данных (Data mining) - построения иерархий классов объектов, импликаций, ассоциативных правил и зависимостей других типов на признаках. Студенты получат навыки автоматического построения иерархической модели предметной области, и находить зависимости в данных, а также анализировать алгоритмическую сложность такого рода задач и строить эффективные алгоритмы порождения иерархий классов объектов и систем зависимостей на множествах признаков объектов.

Задачи дисциплины

Овладеть математическими основами важнейшей области анализа и разработки данных (Data mining) - построения иерархий классов объектов и построения зависимостей (ассоциативных правил) на признаках. Студенты научатся строить иерархическую модель предметной области, находить зависимости в данных, анализировать алгоритмическую сложность такого рода задач, а также строить эффективные алгоритмы порождения иерархий классов объектов и зависимостей на множествах признаков объектов.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области математики, естественных наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
	ОПК-4.4 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные концепции и профессиональные результаты, системные методологии в профессиональной области;
- современное состояние и принципиальные возможности языков и систем программирования.

уметь:

- использовать новые знания и применять их в профессиональной деятельности;
- использовать современные теории, методы, системы и средства прикладной математики и информационных технологий для решения научно-исследовательских и прикладных задач.

владеть:

- строить иерархическую модель предметной области, находить зависимости в данных, анализировать алгоритмическую сложность такого рода задач;
- строить эффективные алгоритмы порождения иерархий классов объектов и зависимостей на множествах признаков объектов.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение: обзор курса. Отношения и графы.	5			5
2	Частично-упорядоченные множества и графы.	4			4
3	Решетки и полурешетки.	5			5
4	Анализ формальных понятий (АФП).	4			4
5	Модели представления знаний, машинного обучения, разработки данных на языке соответствий Галуа и решеток понятий.	5			5
6	Алгоритмические проблемы построения решеток замкнутых множеств и базисов импликаций.	5			5
7	Кластеризация и устойчивость понятий.	2			2
Итого часов		30			30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 1 (Осенний)

1. Введение: обзор курса. Отношения и графы.

Бинарные отношения. Графы, подграфы, части, циклы, клики, деревья, двудольные графы. Графы бинарных отношений. Свойства бинарных отношений (рефлексивность, симметричность, асимметричность, антисимметричность, транзитивность, связность, ацикличность, полнота) и их теоретико-графовое выражение. Важные виды бинарных отношений: эквивалентность, толерантность, частичный порядок.

Дополнительное отношение, обратное (дуальное) отношение, кодуальное отношение, симметрическое дополнение.

Свойства бинарных отношений: рефлексивность, транзитивность, симметричность, асимметричность, антисимметричность. Иллюстрация свойств на графе отношения.

Важные виды отношений: эквивалентность (классы эквивалентности), толерантность (классы толерантности).

2. Частично-упорядоченные множества и графы.

Частичный порядок, строгий Топологическая сортировка порядок, квазипорядок, линейный порядок, отношение покрытия (доминирования), ориентированный граф порядка, диаграмма (Хассе) порядка.

Частичный порядок как транзитивное замыкание отношения покрытия (также через произведение матриц), квазипорядок, отношение несравнимости, частичный порядок на элементах фактор-множества по отношению эквивалентности в квазипорядке; (квази)порядок на помеченных (раскрашенных) графах. Размерность Примеры порядков в математике и приложениях: порядок на мультимножествах, порядок на разбиениях упорядоченного множества (порядковая и мультипликативная).

3. Решетки и полурешетки.

Инфимум, супремум, полурешетки, квазирешетки, два определения решеток. Диаграммы полурешеток решеток. Виды решеток (полные, модулярные, матроиды, дистрибутивные, булевы) и их диаграммы. (Порядковые) фильтры и идеалы решеток. Пополнения частичных порядков до решеток (пополнение Дедекинда-Макнила) и дистрибутивных решеток (Теорема Биркгофа). Соответствия Галуа и их свойства. Соответствие Галуа, основанное на бинарном отношении.

4. Анализ формальных понятий (АФП).

Оператор замыкания и система замыканий (семейство Мура). Замкнутые множества, решетка замкнутых множеств.

Формальный контекст, формальное понятие, частичный порядок на формальных понятиях, решетка формальных понятий. Супремум и инфимум-неразложимые элементы решетки. Основная теорема АФП (Р. Вилле) о представимости полной решетки решеткой формальных понятий. Многозначные контексты, шкалирование данных.

Системы импликаций, правила Армстронга, связь с функциональными зависимостями в базах данных. Базисы импликаций: прямой базис, минимальный базис (Дюкенна-Гига). Псевдосодержания: определения Дюкенна-Гига и Гантера. Размеры базисов.

5. Модели представления знаний, машинного обучения, разработки данных на языке соответствий Галуа и решеток понятий.

Пространство версий через соответствия Галуа. Пространства версий с полурешеточным упорядочением классификаторов. ДСМ-метод порождения гипотез, гипотезы как содержания решетки понятий положительного контекста. Импликации и ДСМ-гипотезы. Гипотезы и пространства версий. Деревья решений и их погружение в решетку полупроизведения шкал. Узорные структуры и их проекции, обучение на узорных структурах. Импликации и ассоциативные правила на узорных структурах. Примеры приложений в биоинформатике и анализе текстов.

Ассоциативные правила в разработке данных (Data mining), их поддержка (support) и степень уверенность (confidence). Ассоциативные правила и решетки формальных понятий. Базис Люксембургера для ассоциативных правил. Базис, основанный на основном дереве диаграммы решетки понятий.

Решетки понятий как средство для построения таксономий и меромоний (системы классов, связанных отношением «быть частью»).

Определения онтологий. Онтология как частично-упорядоченное множество с дополнительным отношением на элементах. Программные средства построения онтологий. Автоматическое построение онтологий по объектно-признаковым таблицам как решеток понятий.

6. Алгоритмические проблемы построения решеток замкнутых множеств и базисов импликаций.

Теоретические оценки временной сложности в худшем случае. Классы сложности P, NP, co-NP и #P. #P-полнота задач подсчета размера решетки замкнутых множеств и размера минимального базиса. NP-полнота некоторых задач о понятиях: Задача определения псевдозамкнутости и co-NP.

Алгоритмы построения решеток: Норриса, Гантера, Замыкай-по-Одному, Нурина и др. Алгоритмы построения минимального базиса импликаций и базиса ассоциативных правил. Программная система ConExp построения решеток понятий, базисов импликаций и ассоциативных правил.

7. Кластеризация и устойчивость понятий.

Классические методы кластеризации, основанные на отношении и метриках сходства. Определение кластера как замкнутого множества объектов с «большим» общим числом признаков. Устойчивость понятия как мера качества кластера. Уровневые и интегральный индексы устойчивости. Устойчивость и дисперсия. Устойчивость и импликации. Устойчивость и свойства решетки понятий. Соотношение между уровневыми индексами устойчивости. Динамика устойчивости при росте числа примеров. Трудновычислимость устойчивости. Алгоритм с полиномиальной задержкой для вычисления индексов устойчивости. Приближенное вычисление устойчивости. Устойчивость в анализе сообществ и медицинской информатике.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Современная прикладная алгебра [Текст] / Г. Биркгоф, Т. К. Барти ; пер. с англ. Ю. И. Манина - СПб. Лань, 2005
2. Бинарные отношения, графы и коллективные решения [Текст] / Ф. Т. Алескеров, Э. Л. Хабина, Д. А. Шварц - М. Физматлит, 2012
3. MATLAB 7 [Текст] : программирование, численные методы / Ю. Л. Кетков, А. Ю. Кетков, М. М. Шульд. — СПб. : БХВ-Петербург, 2005. — 737 с.

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. B. Ganter, G. Stumme, R. Wille, Eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, State-of-the Art Series (2005), vol. 3626, pp. 196-225.

2. Jonas Poelmans, Sergei O. Kuznetsov, Dmitry I. Ignatov, Guido Dedene, Formal Concept Analysis in knowledge processing: A survey on models and techniques. In: Expert Systems with Applications, Vol. 40. No. 16, pp. 6601-6623, 2013
3. Mikhail A. Babin, Sergei O. Kuznetsov, Computing premises of a minimal cover of functional dependencies is intractable. In: Discrete Applied Mathematics, Vol. 161(6), pp. 742-749, 2013.
4. Sergei O. Kuznetsov, Jonas Poelmans, Knowledge representation and processing with formal concept analysis. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol.3(3), pp. 200-215, 2013.
5. Sergei O. Kuznetsov, On Stability of a Formal Concept. Annals of Mathematics and Artificial Intelligence, Vol. 49, pp.101-115, 2007.
6. B. Ganter, S.A. Obiedkov, Conceptual Exploration. Springer, 2016.
7. S.O. Kuznetsov, T.P. Makhalova, On interestingness measures of formal concepts. Inf. Sci.442-443: 202-219 (2018).
8. S.O. Kuznetsov, Ordered Sets for Data Analysis, arXiv 2019, <https://arxiv.org/abs/1908.11341>

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

В процессе самостоятельной работы обучающихся предполагается использование таких программных средств, как Mathcad, Scilab и др.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к практическим занятиям, выполнение двух индивидуальных домашних заданий.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	<u>1</u>
квалификация:	магистр
Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен	
Разработчик:	С.О. Кузнецов, д-р физ.-мат. наук, профессор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области математики, естественных наук и информационно-коммуникационных технологий	ОПК-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знание информационно-коммуникационных технологий для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
	ОПК-4.4 Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Интерпретируемые методы классификации и порождения знаний» обучающийся должен:

знать:

- фундаментальные концепции и профессиональные результаты, системные методологии в профессиональной области;
- современное состояние и принципиальные возможности языков и систем программирования.

уметь:

- использовать новые знания и применять их в профессиональной деятельности;
- использовать современные теории, методы, системы и средства прикладной математики и информационных технологий для решения научно-исследовательских и прикладных задач.

владеть:

- строить иерархическую модель предметной области, находить зависимости в данных, анализировать алгоритмическую сложность такого рода задач;
- строить эффективные алгоритмы порождения иерархий классов объектов и зависимостей на множествах признаков объектов.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Провести настройку «ленивой» схемы классификации на основе анализа ближайших соседей классифицируемого объекта по размеру пересечений с примерами разных классов и их поддержек в разных классах, с учетом значимости признаков и других возможных параметров модели.

2. Предложить определение «приближенного базиса импликаций», допускающее эффективный (полиномиальный по времени) алгоритм порождения.
3. Для конкретного массива данных и задачи классификации сравнить точность классификации с использованием различных методов бинаризации и операций сходства в узорной структуре.
4. Дополнить схему «ленивой классификации» элементом бустинга, учитывающим неоднородность обучающей выборки и значимости различных примеров. Провести экспериментальную настройку параметров модели с помощью скользящего контроля (кросс-валидации)
5. Для нескольких стандартных массивов данных провести сравнительный анализ точности классификации при выборе различных элементов пространств версий, от наиболее частных (замкнутых) до наиболее общих (минимальных генераторов).

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Примеры вопросов к экзамену:

1. Какова сложность построения транзитивного замыкания бинарного отношения, заданного множеством пар элементов (матрицей, графом)?
2. Обладает ли строгий порядок свойством антисимметричности?
3. Каких элементов не может содержать диаграмма частичного порядка?
4. Какова алгоритмическая сложность задачи топологической сортировки. Ответ обосновать.
5. Каких элементов не может содержать диаграмма (дистрибутивной, модулярной) решетки?
6. Доказать, что в произвольной решетке имеет место $(ab) \vee (cd) \geq (ac) \vee (bd)$ для произвольных a, b, c, d .
7. Какова алгоритмическая сложность проверки того, что частичный порядок, заданный отношением покрытия, является (полу)решеткой?
8. По произвольному контексту с помощью алгоритма ЗО построить множество всех понятий и нарисовать диаграмму решетки понятий.
9. По произвольному контексту построить «прямой базис» импликаций (т.е. основанный на «собственных посылках»).
10. По заданному контексту и заданным порогам поддержки и достоверности построить базис ассоциативных правил на основе отношения покрытия решетки понятий (базис Люксембургера).
11. По заданным положительным и отрицательным примерам, представленных в виде бинарных контекстов, построить ДСМ-гипотезы, частичные пространства версий, ассоциативные правила для положительных примеров, и провести классификацию примеров с неопределенным классом.
12. По заданным положительным и отрицательным примерам, представленным в виде узорных структур (на графах или кортежах интервалов), построить ДСМ-гипотезы, частичные пространства версий, ассоциативные правила для положительных примеров, и провести классификацию примеров с неопределенным классом.
13. Доказать эквивалентность двух определений решетки.
14. Построить решетку понятий с помощью алгоритма, имеющего полиномиальную задержку.
15. Определить ассоциативное правило для описаний, заданных множеством помеченных п графов.
16. Как устойчивость связана с поддержкой замкнутого множества признаков?

Пример экзаменационного билета:

Билет 1.

Вопрос 1. Какова алгоритмическая сложность построения транзитивного замыкания бинарного отношения, заданного множеством пар элементов (матрицей, графом)?

Вопрос 2. Каких элементов не может содержать диаграмма (дистрибутивной, модулярной) решетки?

Вопрос 3. По заданному контексту и заданным порогам поддержки и достоверности построить базис ассоциативных правил на основе отношения покрытия решетки понятий (базис Люксембургера).

Билет 2.

Вопрос 1. Какова алгоритмическая сложность проверки того, что частичный порядок, заданный отношением покрытия, является (полу)решеткой?

Вопрос 2. Доказать, что в произвольной решетке имеет место $(ab) \vee (cd) \geq (ac) \vee (bd)$ для произвольных a, b, c, d .

Вопрос 3. По заданным положительным и отрицательным примерам, представленным в виде узорных структур (на графах или кортежах интервалов), построить ДСМ-гипотезы, частичные пространства версий, ассоциативные правила для положительных примеров, и провести классификацию примеров с неопределенным классом.

Критерии оценивания

отлично

10 - всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

9 - систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений;

8 - глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, правильное обоснование принятых решений;

хорошо

7 - твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

6 - знает материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

5 - знает основной материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач неточности;

удовлетворительно

4 - фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

3 - характер знаний достаточен для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

неудовлетворительно

2 - не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет правильно использовать полученные знания при решении типовых практических задач.

1 - не знает формулировок основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

При проведении экзамена обучающемуся предоставляется 60 минут на подготовку. Опрос обучающегося по билету на экзамене не должен превышать двух астрономических часов. Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины и вычислительными средствами.