

«УТВЕРЖДАЮ»

Проректор по научной работе

Драгунов В.К.

« 16 » июня 2015 г.

Программа аспирантуры

Направление 09.06.01 Информатика и вычислительная техника

Направленность (специальность) 05.13.17 Теоретические основы информатики

РАБОЧАЯ ПРОГРАММА

дисциплины по выбору

«Методы и средства анализа текста, распознавания образов, синтеза решающих правил»

Индекс дисциплины по учебному плану: Б1.В.ДВ.1.1

Всего: 108 часов

Семестр 1, в том числе

6 часов – контактная работа,
84 часа – самостоятельная работа,
18 часов – контроль

Программа составлена на основе федерального государственного образовательного стандарта высшего образования (уровень подготовки кадров высшей квалификации) по направлению подготовки 09.06.01 Информатика и вычислительная техника, утвержденного приказом Минобрнауки России от 30 июля 2014 г. № 875, и паспорта специальности, указанной в номенклатуре специальностей научных работников 05.13.17 Теоретические основы информатики, утвержденной приказом Минобрнауки России от 25 февраля 2009 г. № 59.

ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью изучения дисциплины является освоение математических методов распознавания образов и условий их применения в интеллектуальных системах, в частности, в системах обработки и анализа текстов и изображений, методов синтеза решающих правил в одно- и многокритериальных системах принятия решений, включая продукционные, а также методов логического вывода в интеллектуальных системах.

Задачами дисциплины являются:

- получение базовых знаний по задачам, детерминированным и статистическим методам анализа и классификации текстов и критериям качества соответствующих решающих правил;
- овладение методами теории распознавания образов – нейросетевыми, комбинаторно-логическими, статистическими, а также методами их комбинированного использования в задачах машинного обучения;
- формирование представления о методах синтеза решающих правил на основе традиционных методов одно- и многокритериальной оптимизации, а также на основе моделирования эмпирических знаний в условиях неопределенности и с использованием нечетких моделей;
- изучение современных методов индуктивного и дедуктивного вывода, в том числе в условиях неполноты информации;

– приобретение навыков использования изученных моделей и методов в практических задачах.

В процессе освоения дисциплины **формируются следующие компетенции:**

– способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);

– готовность использовать современные методы и технологии научной коммуникации на государственном и иностранном языках (УК-4);

– владение культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий (ОПК-2);

– способность к разработке новых методов исследования и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности (ОПК-3);

– способность представлять полученные результаты научно-исследовательской деятельности на высоком уровне и с учетом соблюдения авторских прав (ОПК-6);

– способность разрабатывать эффективные математические, логические, семиотические и лингвистические модели и методы взаимодействия информационных процессов для решения задач естествознания, техники, экономики и управления (ПК-1);

– способность разрабатывать научно-техническую документацию, оформлять научно-технические отчеты, обзоры, публикации по результатам выполненных исследований на государственном и иностранном языках (ПК-2);

– способность разрабатывать модели и методы анализа данных и обнаружения закономерностей в данных, алгоритмы анализа текста на естественном языке, методы распознавания образов и синтеза решающих правил (ПК-7).

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБРАЗОВАНИЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

В результате освоения дисциплины обучающийся должен демонстрировать **следующие результаты образования:**

знать:

- общую постановку задачи анализа текстовой информации и изображений (ОПК-2);
- методы и алгоритмы анализа текста, устной речи и изображений (ПК-7).
- методы машинного обучения и обнаружения новых знаний (ПК-1);
- методы распознавания образов (ПК-7);
- методы принятия решений в задачах преобразования информации в данные и знания (ПК-1);

уметь:

- применять методы и алгоритмы анализа текста, устной речи и изображений (ПК-7).
- разрабатывать и использовать методы машинного обучения и обнаружения новых знаний (ПК-2);
- применять методы распознавания образов (УК-4);
- применять методы принятия решений в задачах преобразования информации в данные и знания (ОПК-3).
- разрабатывать эффективные математические, логические, семиотические и лингвистические модели и методы взаимодействия информационных процессов для решения задач естествознания, техники, экономики и управления (ПК-1);

владеть:

- терминологией и навыками ведения профессиональной дискуссии по соответствующей тематике (ПК-2);
- навыками поиска информации о новых методах распознавания образов, анализа текстов, преобразования данных в знания (УК-1);

– навыками оценки эффективности математических, логических, семиотических и лингвистических моделей в задачах распознавания образов и анализа текстов (ОПК-6).

КРАТКОЕ СОДЕРЖАНИЕ РАЗДЕЛОВ ДИСЦИПЛИНЫ

Методы и средства анализа текста (22 часа самостоятельной работы)

Детерминированные методы анализа и идентификации текстов. Конечные автоматы и регулярные выражения. Недетерминированные конечные автоматы, их эквивалентность детерминированным автоматам. Распознавание образов, задаваемых регулярными выражениями. Распознавание подцепочек. Двусторонний детерминированный магазинный автомат. Линейный алгоритм распознавания вхождения цепочек. Позиционные деревья и их применение задачам идентификации цепочек.

Статистические методы классификации и анализа текстов. Методы статистического анализа литературных текстов. Энтропия и избыточность естественных языков. Методы К. Шеннона и А.Н. Колмогорова оценивания энтропии естественных языков. Статистические свойства распространенных европейских языков. Кинетические уравнения эволюции выборочных функций распределения текста по буквам. Последовательности букв в тексте как нестационарные ряды. Функционал близости текстов. Идентификация автора и жанра "условно неизвестного" литературного текста в библиотеке известных текстов. Сортировка текстов по категориям. Способы оценки качества классификации текстов. Бинарные классификаторы и агрегированные показатели. Коллекция Reuters-21578: состав и применение для оценки качества классификации текстов.

Распознавание образов (22 часа самостоятельной работы)

Кластеризация и распознавание с использованием метрик и полуметрик. Меры сходства объектов и их совокупностей. Некоторые алгоритмы кластеризации. Решающие функции и их свойства. Распознавание линейно разделимых образов. Метод потенциальных функций. Градиентные методы построения решающих функций. Предварительная обработка образов. Отбор

признаков и преобразование кластеров. Кластеризация с максимальной энергией. Кластеризация с минимальным потенциалом. Кластеризация с минимальным размером. Метод опорных векторов. Нейросетевое моделирование и распознавание. Обучение многослойного персептрона: метод обратного распространения ошибки. Устойчивость и сложность обучения.

Линейные комбинации распознавателей. Голосование распознавателей и распознаватели над распознавателями. Бустинг. Градиентный спуск в пространстве распознавателей. Оптимизация голосования распознавателей.

Комбинаторно-логические методы распознавания. Понятие теста. Линейные тестовые алгоритмы распознавания. Алгоритм Кудрявцева голосования по тестам. Распознавание частично-упорядоченных объектов. Функциональная интерпретация задачи распознавания. Алгоритмы разделения и распознавания. Общая схема. Примеры. Применение в молекулярной биологии. Сжатие генетического кода. Аналитические представления решающих правил. Схемы из функциональных элементов. Классификация и распознавание топологических форм. Топологические формы. Некоторые модели топологических форм. Принцип конечной топологии. Распознавание оптических образов текстов.

Статистические методы распознавания. Алгоритм максимальной апостериорной вероятности. Минимаксный алгоритм. Алгоритм Неймана — Пирсона. Последовательный алгоритм Вальда. Алгоритм максимального правдоподобия.

Распознавание образов и машинное зрение. Морфологический анализ. Принцип минимальной длины описания изображения. Машинное обучение в задачах анализа изображений.

Синтез решающих правил (20 часов самостоятельной работы)

Аксиоматические теории рационального поведения, многокритериальные решения при объективных моделях, методы оценки и сравнения многокритериальных альтернатив. Методы принятия решений в условиях неопределенности, моделируемой случайными процессами.

Экономные вычислительные методы принятия решений. Функции выбора, возможные подходы к оптимизации принятия решений по бинарному отношению. Эффективные методы линейного и выпуклого программирования в вычислительных схемах алгоритмов выбора. Многокритериальные модели принятия решений в условиях определенности. Метод линейной свертки и главного критерия. Многостадийные задачи принятия решений. Экспертные системы принятия решений. Продукционные экспертные системы. Представление и использование нечетких знаний. Моделирование формирования эмпирического знания.

Методы вывода в интеллектуальных системах

(20 часов самостоятельной работы)

Методы достоверного (дедуктивного) и правдоподобного (абдуктивного, индуктивного) выводов в интеллектуальных системах различного назначения. Методы дедуктивного вывода на графовых структурах. Классические и немонотонные модальные логики: логики убеждения и знания, немонотонные логики Мак-Дермотта и Дойла, автоэпистемические логики Мура, логики умолчания Рейтера. Основы теории аргументации. Базовые принципы построения систем обучения и принятия решений, задачи обучения "без учителя" и "с учителем". Индуктивные методы для случая с неполной информацией и методы теории приближенных множеств.

ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ РЕЗУЛЬТАТОВ ОБРАЗОВАНИЯ ПО ДИСЦИПЛИНЕ

Промежуточная аттестация по итогам освоения дисциплины:

1 семестр – дифференцированный зачет.

Вопросы для самоконтроля и проведения зачета

1. Как определяется конечный автомат и как описывается его функционирование?
2. Как применяется модель конечного автомата для анализа текстов.
3. Каково определение регулярного события и как оно связано с конечным автоматом?
4. Как определяется недетерминированный конечный автомат?

5. В чем заключается линейный алгоритм распознавания цепочек?
6. Что такое позиционное дерево и как оно применяется для распознавания цепочек?
7. Как оценивается энтропия естественного языка и как она связана с его избыточностью?
8. Каковы меры близости текстов при их классификации по категориям?
9. Каковы меры качества методов классификации текстов по категориям?
10. Каков состав текстовой коллекции Reuters-21578?
11. Каковы аксиомы метрики и полуметрики?
12. Каковы меры сходства объектов и их совокупностей?
13. Какие известны алгоритмы кластеризации?
14. Каково общее определение решающей функции в распознавании образов?
15. Как осуществляется распознавание линейно разделимых образов?
16. В чем состоит метод потенциальных функций.
17. Какова особенность градиентных методов построения решающих функций?
18. Каковы методы отбора признаков в теории распознавании образов?
19. В чем особенности кластеризации с максимальной энергией, с минимальным потенциалом, с минимальным размером.
20. В чем состоит метод опорных векторов?
21. Как применяется нейросетевое моделирование к распознаванию образов?
22. В чем состоит метод обратного распространения ошибки при обучении многослойного персептрона.
23. Что такое устойчивость и сложность обучения?
24. Как организуются композиции распознавателей?
25. В чем состоит метод градиентного бустинга?
26. В чем заключается тестовое распознавание?

27. Каковы принципы распознавания частично-упорядоченных объектов?
28. В чем состоит принцип конечной топологии распознавания топологических форм?
29. В чем состоит алгоритм максимальной апостериорной вероятности в теории статистического распознавания?
30. Опишите минимаксный алгоритм. Алгоритм Неймана — Пирсона.
31. В чем состоит последовательный алгоритм Вальда.
32. Опишите алгоритм максимального правдоподобия.
33. Каковы методы оценки и сравнения многокритериальных альтернатив?
34. В чем особенность методов принятия решений в условиях неопределенности?
35. Как строятся функции выбора при оптимизации принятия решений по бинарному отношению.
36. Как применяются методы линейного и выпуклого программирования в вычислительных схемах алгоритмов выбора?
37. Перечислите многокритериальные модели принятия решений в условиях определенности.
38. В чем заключаются методы линейной свертки и главного критерия.
39. В чем состоит особенность многостадийных задач принятия решений.
40. Что называют экспертными системами принятия решений?
41. Как представляются нечеткие знания?
42. В чем состоит особенность продукционных экспертных систем.
43. Каковы особенности методов достоверного (дедуктивного) и правдоподобного (абдуктивного, индуктивного) выводов в интеллектуальных системах?
44. Как определяются методы дедуктивного вывода на графовых структурах.

45. В чем различие и особенности классических и немонотонных модальных логик: логики убеждения и знания, немонотонные логики Мак-Дермотта и Дойла, автоэпистемические логики Мура, логики умолчания Рейтера?

46. Что представляет собой теория аргументации?

47. В чём состоят базовые принципы построения систем обучения и принятия решений.

48. Как различаются задачи обучения "без учителя" и "с учителем"?

49. В чем состоят индуктивные методы обучения для случая с неполной информацией.

50. Каковы особенности методов теории приближенных множеств?

Вопросы для проведения зачета

1. Способы задания и описания функционирования конечного автомата.

2. Модель конечного автомата для анализа текстов.

3. Теорема Клини о регулярных событиях.

1. Способы задания и описания функционирования недетерминированного конечного автомата.

2. Линейный алгоритм распознавания цепочек.

3. Позиционное дерево и его применение для распознавания цепочек.

4. Энтропия и избыточность естественного языка.

5. Меры близости текстов при их классификации по категориям.

6. Меры качества методов классификации текстов по категориям.

7. Текстовая коллекция Reuters-21578?

8. Аксиомы метрики и полуметрики.

9. Меры сходства объектов и их совокупностей.

10. Алгоритмы кластеризации.

11. Общее определение решающей функции в распознавании образов?

12. Распознавание линейно разделимых образов?

13. Метод потенциальных функций.

14. Градиентные методы построения решающих функций?

15. Методы отбора признаков в теории распознавании образов?

16. Особенности кластеризации с максимальной энергией, с минимальным потенциалом, с минимальным размером.
17. Метод опорных векторов.
18. Применение нейросетевого моделирования к распознаванию образов.
19. Метод обратного распространения ошибки при бучении многослойного персептрона.
20. Устойчивость и сложность обучения.
21. Композиция распознавателей.
22. Метод градиентного бустинга.
23. Тестовое распознавание.
24. Принципы распознавания частично-упорядоченных объектов?
25. Принцип конечной топологии распознавания топологических форм?
26. Алгоритм максимальной апостериорной вероятности в теории статистического распознавания.
27. Минимаксный алгоритм. Алгоритм Неймана — Пирсона.
28. Последовательный алгоритм Вальда.
29. Алгоритм максимального правдоподобия.
30. Методы оценки и сравнения многокритериальных альтернатив.
31. Особенность методов принятия решений в условиях неопределенности.
32. Синтез функции выбора при оптимизации принятия решений по бинарному отношению.
33. Применение методов линейного и выпуклого программирования в вычислительных схемах алгоритмов выбора?
34. Многокритериальные модели принятия решений в условиях определенности.
35. Методы линейной свертки и главного критерия.
36. Многостадийные задачи принятия решений.
37. Экспертные системы принятия решений.

38. Представление нечетких знаний?
39. Особенности продукционных экспертных систем.
40. Особенности методов достоверного (дедуктивного) и правдоподобного (абдуктивного, индуктивного) выводов в интеллектуальных системах.
41. Методы дедуктивного вывода на графовых структурах.
42. Классические и немонотонные модальные логики: логики убеждения и знания, немонотонные логики Мак-Дермотта и Дойла, автоэпистемические логики Мура, логики умолчания Рейтера.
43. Теория аргументации.
44. Базовые принципы построения систем обучения и принятия решений.
45. Задачи обучения "без учителя" и "с учителем".
46. Индуктивные методы обучения для случая с неполной информацией.
47. Методы теории приближенных множеств.

Критерии оценки за освоение дисциплины определены в Инструктивном письме И-23 от 14 мая 2012 г.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Основная литература:

1. Мерков А.Б. Распознавание образов. Введение в методы статистического обучения. – М.: Эдиториал УРСС, 2011. – 256 с.
2. Мендель А. В. Модели принятия решений: учебное пособие для вузов. – М.: ЮНИТИ-ДАНА, 2013. – 463 с.
3. Болотов А.А., Гашков С.Б., Фролов А.Б. Криптографические протоколы на эллиптических кривых: учебное пособие. – М.: Изд. дом МЭИ, 2007. – 84 с.
4. Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. Достоверный и правдоподобный вывод в интеллектуальных системах / Под.

ред. В.Н. Вагина, Д.А. Поспелова. Учебное пособие. – 2-е изд., испр. и доп. – М.: ФИЗМАТЛИТ, 2008. – 712 с.

5. Редько В.Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики. – 7-е изд. – М.: Эдиториал УРСС, 2011. – 224 с.

6. Финн В.К. Искусственный интеллект: Методология, применения, философия. Науч.ред. М.А. Михеенкова. –М.: Эдиториал УРСС, 2011. –448 с.

7. Фролов А.Б. Классификация и распознавание топологических форм: учебное пособие. Ред. В.Б. Кудрявцев. – М.: Изд. дом МЭИ, 2010. – 52 с.

Дополнительная литература:

8. Кудрявцев В.Б., Андреев А.Е., Гасанов Э.Э. Теория тестового распознавания. – М.: Физматлит. 2007. – 320 с.

9. Ларичев О.И. Теория и методы принятия решений, а также Хроника событий в Волшебных странах: учебник для вузов. – 3-е изд., перераб. и доп. – М.: Логос, 2008. – 392 с.

10. В.В. Подиновский, В.Д. Ногин. Парето-оптимальные решения многокритериальных задач. – 2-е изд., испр. и доп. – М.: Физматлит, 2007. – 256 с.

11. Толчеев В.О. Современные методы обработки и анализа текстовой информации. Учебное пособие. М.: Изд-во МЭИ, 2006. – 76 с.

12. Толчеев В.О. Основы теории классификации многомерных наблюдений: учебное пособие. – М.: Изд. дом МЭИ, 2012. – 124 с.

13. Петровский А.Б. Теория принятия решений. – М.: Академия. 2009. – 400 с.

14. Ширяев В.И., Ширяев Е.В. Принятие решений. Математические основы. Статические задачи. – М.: Эдиториал УРСС, 2009. – 208 с.

15. Яглом А.М., Яглом И.М. Вероятность и информация. – 4-е изд., стер. – М.: Эдиториал УРСС, 2006. – 512 с.