ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ВОПРОСЫ
ДЛЯ ПОДГОТОВКИ К ЭКЗАМЕНАМ

дисциплины Интеллектуальный анализ данных
для направления 09.04.04 Программная инженерия
уровень образования магистратура
профиль подготовки Искусственный интеллект и инженерия данных
форма обучения очная
кафедра-разработчик Системное программирование

Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 09.04.04 Программная инженерия, утверждённым приказом Минобрнауки от 19.09.2017 № 932

Разработчик программы,
доктор физ.-мат. наук, доцент,
профессор кафедры СП _____ М.Л. Цымблер
(ученая степень, ученое звание, должность) (подпись)

Зав. кафедрой Системное программирование
доктор физ.-мат. наук, проф. _____ Л.Б. Соколинский
(ученая степень, ученое звание) (подпись)

Челябинск

Введение в дисциплину. Феномен Больших данных. Понятие интеллектуального анализа данных. Технологический цикл анализа данных. Основные задачи интеллектуального анализа данных: поиск шаблонов, классификация, кластеризация, поиск аномалий.

Поиск шаблонов. Понятия транзакции, частого набора, шаблона, поддержки, достоверности. Основные алгоритмы поиска частых наборов: Apriori, Eclat, FP-Growth. Выбор полезных шаблонов на основе мер support, confidence, lift и др. Компактное представление частых наборов: максимально частые и замкнутые наборы, иерархии наборов. Фрагментация и сэмплинг для поиска частых наборов

Классификация. Процесс классификации: обучение модели, оценка модели, применение модели. Деревья решений. Меры оценки доли примесей в узле дерева решений: индекс Джини, энтропия; алгоритмы классификации ID3, C4.5, CART. Байесовская классификация. Классификация по ближайшим соседям. Оценка качества классификации: меры Accuracy, Precision, Recall, F1. Ансамблевая классификация: бэггинг, бустинг, случайный лес.

Кластеризация. Задачи кластеризации данных и подходы к ее решению. Разделительная кластеризация: алгоритмы k-means, k-medoids и др. Иерархическая кластеризация: дендрограммы, агломеративный и дивизимный подход. Меры схожести кластеров: Single linkage, Complete linkage, Group average и др. Плотностная кластеризация: алгоритм DBSCAN. Нечеткая кластеризация: алгоритм Fuzzy C-Means. Меры качества кластеризации: критерий Хопкинса, кросс-валидация, метод локтя, силуэтный коэффициент и др.

Поиск аномалий. Понятия аномалии (выброса), шума, новизны в данных. Виды аномалий: точечные, глобальные, контекстные, смешанные. Статистические методы поиска аномалий: z-значимость, правило трех сигм, гистограммы. Поиск аномалий на основе расстояния. Поиск аномалий на основе плотности: метод вложенных циклов, метод решеток. Поиск аномалий с помощью разделительной и плотностной кластеризации. Поиск аномалий на основе классификации: метод One Class SVM, метод изолирующего леса.

Основная литература:

Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN-13: 978-0-13-312890-1

Дополнительная литература:

1. Aggarwal C.C. Data Mining: The Textbook. Springer, 2015. 746 p. ISBN 978-3-319-14141-1.
2. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2012. 740 p. ISBN 978-0123814791