ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ПРАКТИЧЕСКИМ ЗАНЯТИЯМ

дисциплины 1.Ф.П0.12 Технологии аналитической обработки информации
для направления 09.03.04 Программная инженерия
уровень образования Бакалавриат
профиль подготовки Инженерия информационных и интеллектуальных систем
форма обучения очная
кафедра-разработчик Системное программирование

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

Оглавление

1 . Выполнение работ 3
2 . Поиск шаблонов 5
Задание 1. Поиск частых наборов 5
Задание 2. Поиск ассоциативных правил 5
3 . Классификация 7
Задание 3. Байесовская классификация 7
Задание 4. Классификация с помощью дерева решений 7
Задание 5. Ансамблевая классификация с помощью бэггинга 7
Задание 6. Ансамблевая классификация с помощью случайного леса 7
Задание 7. Ансамблевая классификация с помощью бустинга 7
4 . Кластеризация 9
Задание 8. Разделительная кластеризация 9
Задание 9. Плотностная кластеризация 9
Задание 10. Иерархическая кластеризация 10
Задание 11. Качество кластеризации 10
5 . Поиск аномалий 11
Задание 12. Поиск точечных аномалий 11
Задание 13. Поиск коллективных аномалий 11

1. Выполнение работ

Задание, выполняемое на практическом занятии, предполагает решение студентом небольшой учебно-исследовательской задачи по теме дисциплины, подготовку и защиту отчета о разработанном решении. Задача, как правило, заключается в выполнении интеллектуального анализа указанного набора данных (временного ряда) и визуализации полученных результатов.

Студенту необходимо создать на одном из свободно доступных сервисов (github, bitbucket и др.) публичный *репозиторий по дисциплине* для сохранения исходных текстов разработанных решения заданий и др. материалов, создаваемых в рамках практических занятий.

Набор данных предлагается студентом и согласовывается с преподавателем, при этом предпочтительны референсные наборы данных – размещенные в авторитетных свободно доступных интернет-репозиториях (например, UCI Machine Learning Repository https://archive.ics.uci.edu/) или/и упомянутые в научных статьях, опубликованных в авторитетных рецензируемых журналах.

Алгоритм интеллектуального анализа данных может быть реализован студентом с помощью сторонних библиотек или самостоятельно (предпочтительно). При разработке *программы* допустимо использовать любые языки программирования, библиотеки и инструментальные средства (если явно не указано обратное).

Исходные тексты программы и сопутствующие материалы задания (наборы данных, результаты работы и визуализации, отчет) необходимо сохранять в репозитории по дисциплине (отдельный каталог для каждого задания). Исходные тексты должны быть документированы (наличие спецификаций файлов и подпрограмм).

Дополнительные (бонусные) баллы: за качественное использование в реализации параллельных/распределенных алгоритмов.

Отчет о выполнении задания должен включать в себя следующие основные элементы:

- полные ФИО автора отчета, адрес электронной почты для связи;
- формулировка задания;
- библиографическая ссылка и краткие сведения о наборе данных;
- краткие сведения о средствах реализации (если применимо) и гиперссылка на каталог репозитория с исходными текстами и сопутствующими материалами;

– рисунки с результатами визуализации[1];
– краткие пояснения к полученным результатам.

Защита отчета предполагает устные ответы студента на вопросы преподавателя по реализации программы и полученным результатам.

[1] Рисунки должны иметь подписи. Графики и диаграммы на рисунках должны иметь легенду, подписи осей с указанием единиц измерения (если применимо).

2. Поиск шаблонов

Задание 1. Поиск частых наборов

Выполните поиск частых наборов объектов в трех различных наборах данных с помощью следующих алгоритмов (или их модификаций): Apriori, FP-Growth, ECLAT. Наборы данных должны существенно отличаться друг от друга по количеству транзакций и/или типичной длине транзакции (количеству объектов). Варьируйте пороговое значение поддержки (например: 1%, 3%, 5%, 10%, 15%, 20%). Проверьте идентичность результатов, полученных с помощью различных алгоритмов.

1. Подготовьте список частых наборов, в которых не более семи объектов (разумное количество). Проанализируйте и изложите содержательный смысл полученного результата.
2. Выполните визуализацию полученных результатов в виде следующих диаграмм:

- сравнение быстродействия алгоритмов на фиксированном наборе данных при изменяемом пороге поддержки;
- общее количество частых наборов объектов на фиксированном наборе данных при изменяемом пороге поддержки;
- максимальная длина частого набора объектов на фиксированном наборе данных при изменяемом пороге поддержки;
- количество частых наборов объектов различной длины на фиксированном наборе данных при изменяемом пороге поддержки.

Задание 2. Поиск ассоциативных правил

Выполните поиск ассоциативных правил для наборов данных из задания 1. Зафиксируйте значение пороговое значение поддержки (например, 10%), варьируйте пороговое значение достоверности (например, от 70% до 95% с шагом 5%). Получите список результирующих правил в удобочитаемом виде (антецедент➔консеквент).

1. Подготовьте список правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов (разумное количество). Проанализируйте и изложите содержательный смысл полученного результата.
2. Выполните визуализацию полученных результатов в виде следующих диаграмм:

- сравнение быстродействия поиска правил на фиксированном наборе данных при изменяемом пороге достоверности;
- общее количество найденных правил на фиксированном наборе данных при изменяемом пороге достоверности;
- максимальное количество объектов в правиле на фиксированном наборе данных при изменяемом пороге достоверности;
- количество правил, в которых антецедент и консеквент суммарно включают в себя не более семи объектов, на фиксированном наборе данных при изменяемом пороге достоверности.

3. Классификация

Задание 3. Байесовская классификация

Выполните классификацию набора данных с помощью Байесовской классификации, варьируя соотношение мощностей обучающей и тестовой выборок от 60%:40% до 90%:10% с шагом 5%.

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм.

Задание 4. Классификация с помощью дерева решений

Выполните классификацию набора данных из задания 3 с помощью построения дерева решений, фиксируя критерий выбора атрибута разбиения (information gain, gain ratio, index gini) и варьируя соотношение мощностей обучающей и тестовой выборок (от 60%:40% до 90%:10% с шагом 10%). Выполните визуализацию построенных деревьев решений.

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм.

Задание 5. Ансамблевая классификация с помощью бэггинга

Выполните классификацию набора данных из задания 3 с помощью бэггинга, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм. Нанесите на диаграммы соответствующие значения, полученные в заданиях 3, 4.

Задание 6. Ансамблевая классификация с помощью случайного леса

Выполните классификацию набора данных из задания 3 с помощью случайного леса, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм. Нанесите на диаграммы соответствующие значения, полученные в заданиях 3, 4, 5.

Задание 7. Ансамблевая классификация с помощью бустинга

Выполните классификацию набора данных из задания 3 с помощью бустинга, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм. Нанесите на диаграммы соответствующие значения, полученные в заданиях 3, 4, 5, 6.

4. Кластеризация

Задание 8. Разделительная кластеризация

1. Выполните кластеризацию набора 2-х или 3-мерных данных с помощью алгоритма k-Means (предполагается, что полученные кластеры будут выпуклыми), используя различные значения параметра k (из интервала 3..9).

   Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

2. Внесите шум в набор данных (случайным образом изменить определенную долю объектов набора: 1%, 3%, 5%, 10%; изменение может заключаться в добавлении/вычитании к/из одной/нескольких координат объекта случайного числа).

   Выполните кластеризацию зашумленного набора данных с помощью алгоритмов k-Means и k-Medoids (или PAM), используя различные значения параметра k (из интервала 3..9).

   Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

3. Выполните кластеризацию набора данных из задания 9 (с невыпуклыми кластерами) с помощью алгоритмов k-Means и k-Medoids (или PAM), используя различные значения параметра k (из интервала 3..9).

   Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

Задание 9. Плотностная кластеризация

1. Выполните кластеризацию набора 2-х или 3-мерных данных с помощью алгоритма DBSCAN (предполагается, что полученные кластеры не будут выпуклыми), используя различные значения параметров $MinPts$ (из интервала 3..9) и Eps.

   Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

2. Выполните кластеризацию зашумленного набора данных из задания 8 с помощью алгоритма DBSCAN, используя различные значения параметров $MinPts$ (из интервала 3..9) и Eps.

   Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

Задание 10. Иерархическая кластеризация

Выполните иерархическую кластеризацию набора данных, используя различные меры схожести: Single linkage, Complete linkage, Group average, расстояние Уорда (Ward).

Выполните визуализацию полученных результатов в виде дендрограмм.

Задание 11. Качество кластеризации

Для набора данных из задания 8 подберите оптимальное количество кластеров с помощью двух любых приемов из следующего множества: метод локтя, кросс-валидация, силуэтный коэффициент, визуализация матрицы схожести.

Постройте диаграммы, подтверждающие полученные результаты.

5. Поиск аномалий

Задание 12. Поиск точечных аномалий

Выполните поиск точечных аномалий (выбросов) в двух различных наборах одномерных данных с помощью двух любых приемов из следующего множества: метод максимального правдоподобия, оценка χ^2, построение гистограмм.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных точек.

Задание 13. Поиск коллективных аномалий

Выполните поиск коллективных аномалий (выбросов) в двух различных наборах 2-х или 3-мерных данных с помощью двух любых приемов из следующего множества: метод вложенных циклов, метод решеток, кластеризация.

Выполните визуализацию полученных результатов в виде точечных графиков, использующих два цвета для отражения нормальных/аномальных точек.