Search and Data Mining: Tools

Clustering

Anya Yarygina

Boris Novikov

Оценки

Активность	**Макс. Оценка**
Проект (каждый из 5)	20
Большой доклад	15
Короткий доклад	5
Устный экзамен (в сессию)	20

Итоговая оценка	**Баллы**
Отлично	> 81
Хорошо	61 - 80
Удовлетв.	40 – 60

Как добиться снижения оценки	
Некачественное выполнение задания (небрежный отчет, отсутствие выводов)	-5
Неполностью выполненное задание (некоторые пункты задания не выполнены)	-10
Задание не отправлено через 2 недели после формулировки	-5
Задание не отправлено через 4 недели после формулировки	-10

Проекты

Project #	Topic	Start Date	Due date
1	Data pre-processing	Feb 21	Mar 06
2	OLAP	Mar 07	Mar 20
3	Text search	Mar 21	Apr 03
4	Classification	Apr 04	Apr 17
5	Clustering	Apr 18	May 03

Отчет отправляется по электронной почте одним письмом на всех преподавателей с темой «Search and Mining Tools and Techniques» до 23:59 (UTC+04:00) дня указанного в таблице.

Структура практического занятия

- Постановка задачи
  - Формулировка задачи
  - Описание требований
  - Советы по выполнению задания
  - Вопросы по выполнению задания
- Обсуждение предыдущей задачи
  - «Идеальные» отчеты
  - Отсутствующие отчеты
  - Вопросы по отчетам

Задание 1: Подготовка данных

- Выбрать наборы данных для практических заданий из предложенных или согласовать любые другие и ознакомиться с ними
- Текстовый отчет с содержательным описанием 4 выбранных наборов данных для каждого практического задания

Задание 2: OLAP

- Выполнение задания включает
  - выбор и подготовку модели;
  - проектирование хранилища данных;
  - генерацию и загрузку данных;
  - проектирование запросов для получения нескольких отчетов;
  - построение отчетов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание формализации задачи (какие размерности определены, какие меры, атрибуты и т.д.)
  - Описание запросов, извлекающих данные из полученного куба (какие содержательные результаты предполагалось получить и какими запросами это достигается)

Задание 3: Text Search

- Выполнение задания включает
  - выбор коллекции текстовых документов и задачи поиска
  - выбор структуры документов
  - загрузка документов в индекс
  - эксперименты с разными типами текстовых запросов:
    - Текстовые запросы
    - AND, OR, LIKE
    - Запросы к разными полям документа
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Описание отображения формальной задачи на данные (документ, поля с типами)
  - Описание инструмента индексирования и поиска (название, параметры, подходы)
  - Описание процесса загрузки данных
  - Описание процесса поиска: осмысленные запросы разного типа и результаты
  - Замеры времени выполнения запросов, если загружается вся коллекция или только половина

Задание 4: Classification

- Выполнение задания включает:
  - Выбор коллекции тестовых данных
  - Выбор структуры данных
  - Запуск минимум трех существенно отличающихся алгоритма классификации
  - Проведение эксперимента с выбором атрибутов (любой алгоритм)
  - Проведение эксперимента с ансамблем (любой алгоритм bagging или boosting)
  - Анализ результатов всех экспериментов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Отображение формальной задачи на данные (объект, атрибуты с типами, класс)
  - Короткое описание инструмента (загрузка данных, интерфейсы, доступные алгоритмы)
  - Описание результатов классификации:
    - Алгоритм
    - Параметры
    - Качество: confusion matrix, точность
    - Выводы
  - Выводы по всей серии экспериментов

Задание 5: Clustering

- Выполнение задания включает:
  - Выбор коллекции тестовых данных
  - Выбор структуры данных
  - Запуск минимум одного алгоритма кластеризации с заданным числом кластеров (перебор в небольшом диапазоне значений 2-6)
  - Запуск минимум одного алгоритма кластеризации с автоматическим подбором числа кластеров
  - Запуск минимум двух существенно отличающихся алгоритма кластеризации с «правильным» числом кластеров
  - Анализ результатов всех экспериментов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Отображение формальной задачи на данные (объект, атрибуты, расстояние)
  - Короткое описание инструмента (загрузка данных, интерфейсы, доступные алгоритмы)
  - Описание результатов классификации:
    - Алгоритм
    - Параметры
    - Качество (метрики, если есть разбивка по классам; зрительный анализ на основе визуализации)
    - Выводы
  - Выводы по всей серии экспериментов

Задание 5: Инструменты

- Weka
- R
- RapidMiner
- Orange

Задание 5: Пример

- Weka GUI
- BIRCHGenerate data
  - 4 clusters
  - 2 attributes
  - 146 instances
- Visualize
- Clustering
  - EM
  - SimpleKMeans
  - HierarchicalClusterer
  - XMeans
- Classes to cluster evaluation
- Visualize cluster assignments

Задание 5: Clustering

- Выполнение задания включает:
  - Выбор коллекции тестовых данных
  - Выбор структуры данных
  - Запуск минимум одного алгоритма кластеризации с заданным числом кластеров (перебор в небольшом диапазоне значений 2-6)
  - Запуск минимум одного алгоритма кластеризации с автоматическим подбором числа кластеров
  - Запуск минимум двух существенно отличающихся алгоритма кластеризации с «правильным» числом кластеров
  - Анализ результатов всех экспериментов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Отображение формальной задачи на данные (объект, атрибуты, расстояние)
  - Короткое описание инструмента (загрузка данных, интерфейсы, доступные алгоритмы)
  - Описание результатов классификации:
    - Алгоритм
    - Параметры
    - Качество (метрики, если есть разбивка по классам; зрительный анализ на основе визуализации)
    - Выводы
  - Выводы по всей серии экспериментов

Задание 4: Classification

- Выполнение задания включает:
  - Выбор коллекции тестовых данных
  - Выбор структуры данных
  - Запуск минимум трех существенно отличающихся алгоритма классификации
  - Проведение эксперимента с выбором атрибутов (любой алгоритм)
  - Проведение эксперимента с ансамблем (любой алгоритм bagging или boosting)
  - Анализ результатов всех экспериментов
- Краткий текстовый отчет
  - Неформальное описание задачи и данных
  - Отображение формальной задачи на данные (объект, атрибуты с типами, класс)
  - Короткое описание инструмента (загрузка данных, интерфейсы, доступные алгоритмы)
  - Описание результатов классификации:
    - Алгоритм
    - Параметры
    - Качество: confusion matrix, точность
    - Выводы
  - Выводы по всей серии экспериментов