Классификация записей из микроблогов с помощью Википедии

Абишев Т. М.
545 группа, математико-механический факультет, СПбГУ

Научный руководитель: Барашев Д. В.
Доцент, математико-механический факультет, СПбГУ

Рецензент: ? ?. ?.
?, ?, ?

СПбГУ, Санкт-Петербург, май 2012

Введение

- Микроблоги, как источник данных
  - Выборы
  - Кассовые сборы
- Классификация записей
  - Особенности

Постановка задачи

- Построить классификатор сообщений из микроблогов
- Который будет использовать
  - википедию
  - контекст

Решение (контекст)

- Использование других записи автора, как контекст для классификации
- Алгоритм
  - Кластеризуем сообщения автора
  - Классифицируем на основании «большинства» в кластере сообщения

Решение (википедия)

- Алгоритмы классификации требуют векторов, не способны принимать просто текст
- Использовать Википедию для преобразования текста в вектора
- Алгоритм
  - Нахождение релевантных страниц Википедии к тексту
  - Получение их надкатегорий, как координат пространства

Эксперименты (описание 1)

- Тестовые данные
  - Математика/физика/биология/химия/программирование
  - Новости/личное/предложения от компаний
- Оценка результата
  - Точность – {правильно классифицированные к классу C} / {классифицированные к классу C}
  - Полнота – {правильно классифицированные к классу C} / {сущностей в классе C}
  - F-мера – среднее гармоническое точности и полноты

Эксперименты (описание 2)

- Алгоритмы классификации
  - Наивный байесовский
  - SVM – метод опорных векторов
  - J48 – метод для построения дерева принятия решений
- Алгоритмы кластеризации
  - kmeans на 20 кластеров
  - kmeans на 100 кластеров
  - xmeans от 10 до 200 кластеров

Эксперименты (результаты)

- Наилучшие результаты (как с использование, так и без контекста) показал алгоритм SVM
- Улучшение в его случае с 0.67 до 0.75 по F-мере в одном случае, и с 0.915 до 0.927 в другом
- Наиболее сбалансированным алгоритмом кластеризации является xmeans
- Наибольше улучшение демонстрирует SVM, наименьшее – J48.

Результаты

- Создан алгоритм классификации записей из кикроблогов
  - Использующий контекст записи
  - Использующий Википедию, как источник дополнительных данных
- Алгоритм показал хорошие результаты и продемонстрировал улучшение в сравнении с простым подходом для классификации записей