

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: Шалымов Д. С.

Доцент, математико-механический факультет, СПбГУ

Введение

- Микроблоги как источник данных
 - Выборы
 - Кассовые сборы
- Классификация записей
 - По тематике
 - Спам/не спам
 - Содержательные/не содержательные

Постановка задачи

- Построить классификатор записей из микроблогов
- Который будет использовать
 - Википедию
 - Контекст

Решение (контекст)

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

Решение (википедия)

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

Эксперименты (описание 1)

- Размеченные тестовые данные
 - Математика/физика/биология/химия/программирование
 - Новости/личное/предложения от компаний
- Оценка результатов
 - Точность – $\{\text{правильно классифицированные к классу } C\} / \{\text{классифицированные к классу } C\}$
 - Полнота – $\{\text{правильно классифицированные к классу } C\} / \{\text{сущностей в классе } C\}$
 - 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)

- Сравниваем «базовый» алгоритм классификации и построенный по моей схеме
- Наилучшие результаты 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 подходящим алгоритмом кластеризации является xmeans
- Наименьшее улучшение – наивный байесовский алгоритм.

Результаты

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