Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Анализ тональности текста на основе машинного обучения

Студент: Малышев Иван Алексеевич ИУ7-51Б

Научный руководитель: Кузнецова Ольга Владимировна

Цель и задачи

Цель: исследование современных методов машинного обучения для задачи определения тональности в тексте на естественном языке.

Задачи:

- изучить существующие методы анализа тональности на основе машинного обучения;
- предложить критерии оценки качества методов;
- выбрать метод, который наиболее эффективно решает задачу анализа тональности, учитывая проблемы этой задачи.

Методы анализа тональности текста на основе машинного обучения

- Наивный байесовский классификатор
- Метод максимума энтропии
- Деревья решений
- Случайный лес
- Логическая регрессия
- Метод опорных векторов

Наивный байесовский классификатор

- Является вероятностным классификатором
- Основан на теореме Байеса с предположением о том, что все признаки являются независимыми

Преимущества:

- Простота реализации
- Большая скорость работы
- Малое количество данных необходимых для обучения

Недостатки:

- Низкое качество классификации

Метод максимума энтропии

- Также является вероятностным классификатором
- Основан на принципе максимальной энтропии
- Не предполагается независимость признаков

Преимущества:

- Простота реализации
- Малое количество данных необходимых для обучения

Недостатки:

- Низкое качество классификации

Деревья решений

- Представляют из себя древовидную структуру
- На ребрах («ветках») записаны атрибуты, , от которых зависит целевая функция
- В «листьях» записаны значения целевой функции
- В остальных узлах - атрибуты, по которым различаются случаи

Преимущества:

- Простота в интерпретации
- Отсутствие требования подготовки данных
- Возможность работать с большим объемом информации без подготовительных процедур

Недостатки:

- Проблема получения оптимального дерева решений
- Высокая зависимость от обучающих данных

Случайный лес

- Ансамбль решающих деревьев
- Деревья строятся до тех пор, пока в каждом листе не окажется очень мало объектов, то есть они сильно переобучены
- Затем все деревья объединяются и получается эффективный классификатор без недостатков дерева решений

Преимущества:

- Все преимущества и отсутствие недостатков дерева решений

Недостатки:

- Большие временные затраты на построение глубоких деревьев с большим числом признаков

Логическая регрессия

- Оценивает вероятность принадлежности объектов к классу путем сравнения с логической кривой по значениям множества признаков
- На практике часто рассматривается логическая регрессия с регуляризацией
- Регуляризация заключается в том, что модель начинает штрафовать за очень большие веса, что не дает модели переобучиться

Преимущества:

- Высокое качество классификации

Недостатки:

- Необходимость качественной предобработки признаков и их отбор

Метод опорных векторов

- Цель метода заключается в нахождении среди всех возможных гиперплоскостей пространства, отделяющих два класса обучающих примеров друг от друга, такой гиперплоскости, расстояния от которой до ближайших векторов обоих классов равны
- Часто применяется в задачах классификации текстов

Преимущества:

- Является одним из наиболее эффективных методов классификации
- Хорошо масштабируются
- Могут работать с большим количеством признаков
- Могут работать на очень больших выборках

Недостатки:

- Проблема неоднозначности построения гиперплоскости

Сравнение методов (1/2)

Метод	Обучающая выборка	Тестовая выборка
Логическая регрессия	0.93445	0.93445
Дерево решений	0.68204	0.65000
Случайный лес	0.90799	0.84000
Метод опорных векторов	0.89416	0.86167

Таблица результатов сравнения методов традиционного машинного обучения на основе данных из англоязычных корпусов текстов

Сравнение методов (2/2)

Метод	Точность (%)
Метод максимума энтропии	72.60
Случайный лес	88.39
Наивный байесовский классификатор	75.50
Метод опорных векторов	91.15

Таблица результатов сравнения методов традиционного машинного обучения на основе данных из отзывов о товарах с интернет-магазина

Проблемы анализа тональности и их решение

- Выделение имплицитной оценки
  - Решение: добавление списка правил с лингвистическими шаблонами для распознавания имплицитного мнения
- Выделение иронии и сарказма
  - Решение: создать еще один классификатор, обученный на шаблонах ироничных и саркастических высказываний, который передает свой результат работы классификатору тональности
- Дизамбигуляция
  - Решение: использование тезаурусов
- Разрешение референции
  - Решение: использование дополнительных правил и составленных списков кореферентных элементов

Заключение

- изучены существующие методы анализа тональности на основе машинного обучения;
- предложены критерии оценки качества методов;
- выбран метод, который наиболее эффективно решает задачу анализа тональности, учитывая проблемы этой задачи.