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

СПбГУ, Санкт-Петербург, июнь 2012

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

Построение контекста

- Использование других записей автора как контекст для классификации
- Алгоритм
  - Кластеризуем записи автора
  - Для каждого кластера классифицируем все записи по отдельности
  - Помечаем кластер большинством голосов

Использование контекста

- Находим наиближайший кластер к классифицируемой записи
- Результатом является метка кластера

Выделение признаков из текста

- Для классификации необходимы признаки
- Традиционный подход – bag of words
- Короткая длина записей как проблема
- Использование Википедии как вариант решения проблемы

Выделение признаков на основе Википедии

- Нахождение релевантных тексту страниц в Википедии
- Над-категории релевантных страниц как признаки текста

Тестовые данные и критерии качества

- Размеченные тестовые данные
  - Математика/физика/биология/химия/программирование (тематическая)
  - Новости/личное/предложения от компаний (целевая)
  - Количество записей – 1500
- Оценка результатов
  - Точность, полнота
  - F-мера – среднее гармоническое точности и полноты

Участники забега

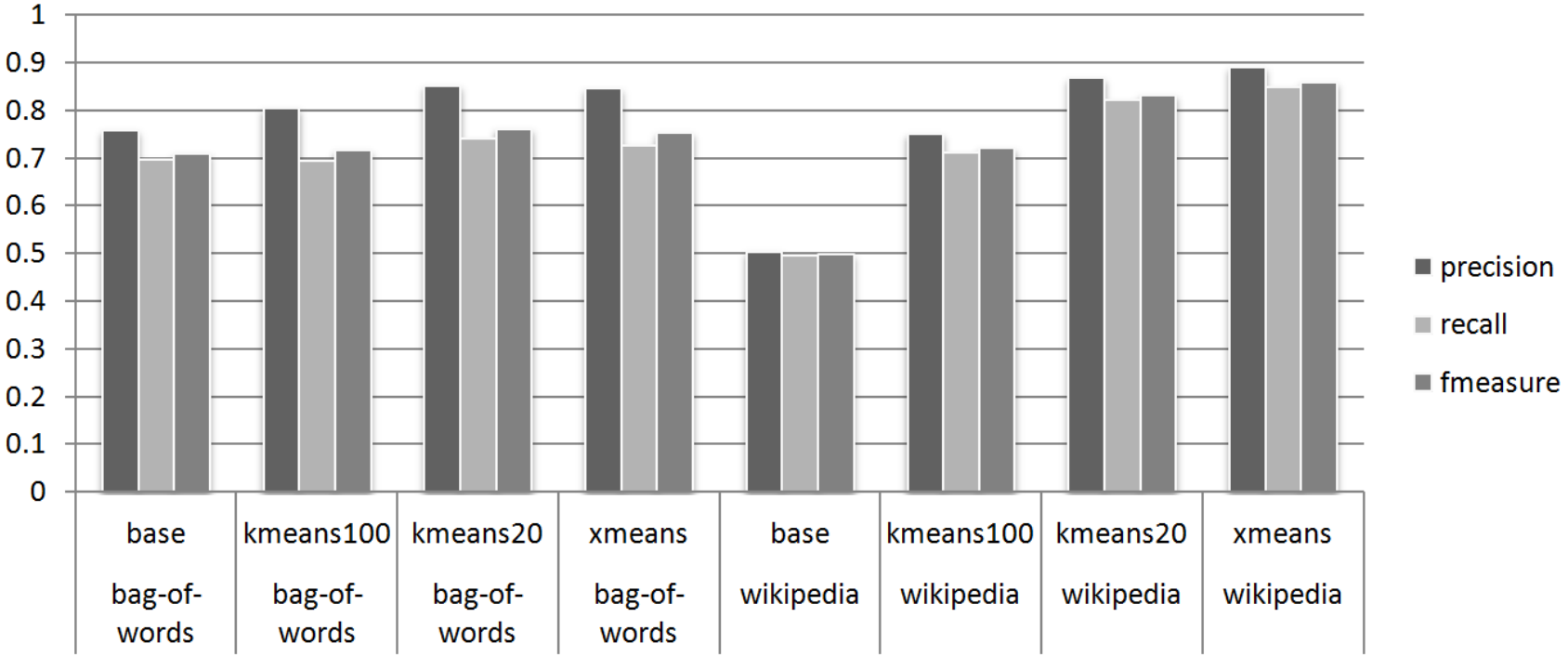
- Алгоритмы классификации
  - Наивный байесовский
  - SVM – метод опорных векторов
  - J48 – метод для построения дерева принятия решений
- Алгоритмы кластеризации
  - kmeans на 20 кластеров
  - kmeans на 100 кластеров
  - xmeans от 10 до 200 кластеров
- 12 вариантов классификаторов, 2 способа выделения признаков из текста

Результаты экспериментов

- Наилучшие результаты показал алгоритм SVM и xmeans
- Использование Википедии ухудшает базовую классификацию, но улучшает контекстную
- Использование контекста дает больший прирост для тематической выборки, чем для целевой
- Наименьшее улучшение – наивный байесовский алгоритм

Результаты экспериментов

- Результат для тематической выборки и алгоритма SVM



Результаты

- Создан алгоритм классификации записей из микроблогов
  - Использующий контекст записи
  - Использующий Википедию для извлечения признаков из текста
- Алгоритм показал хорошие результаты и продемонстрировал улучшение в сравнении с традиционным подходом для классификации записей