Оптимизация рекомендаций в задаче категоризации документов

Бергер Анна Игоревна
Научный руководитель: доцент, к.ф.-м.н.
Гуда Сергей Александрович

Южный Федеральный Университет
Институт математики, механики и компьютерных наук им. И. И. Воровича
Кафедра алгебры и дискретной математики

08.04.2016

1. Постановка задачи
2. Использованные подходы
3. Итерационные алгоритмы поиска максимума

Задача категоризации документов

- Множество документов $\mathfrak{D}$ - набор пар feat:value
- Множество категорий $\mathfrak{C}$ - набор документов
- Целевая функция $\Phi : \mathfrak{D} \to 2^{\mathfrak{C}}$ - неизвестна
- Метрика качества *Macro F1-score(MaF)*:

$$MaF = \frac{2 * MaP * MaR}{MaP + MaR} \tag{1}$$

$$MaP = \frac{\sum_{i=1}^{|\mathfrak{C}|} \frac{tp_{c_i}}{tp_{c_i} + fp_{c_i}}}{|\mathfrak{C}|}, MaR = \frac{\sum_{i=1}^{|\mathfrak{C}|} \frac{tp_{c_i}}{tp_{c_i} + fn_{c_i}}}{|\mathfrak{C}|}. \tag{2}$$

Цель работы

Ранее получена мера релевантности документа категории, основанная на определении информационного выигрыша.

Цель: поиск оптимального в смысле максимизации метрики качества *Macro F1-score(MaF)* числа предсказываемых документов для каждой категории

1 Постановка задачи

2 Использованные подходы

3 Итерационные алгоритмы поиска максимума

Способы поиска наилучшего количества предсказываемых документов

- Выбор абсолютного порогового значения ранжирующей функции
- Фиксирование количества предсказываемых документов для каждой категории
- Комбинация предыдущих методов
- Итерационные алгоритмы поиска максимума

1 Постановка задачи

2 Использованные подходы

3 Итерационные алгоритмы поиска максимума

Оптимизация глобального значения *Macro F1-score(MaF)*

Итерационный алгоритм поиска максимума основан на методе покоординатного спуска.

- Последовательный
- Стохастический

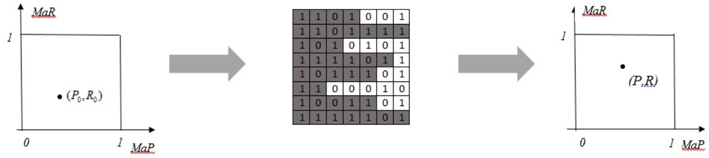
$$MaF = \frac{2 * MaP * MaR}{MaP + MaR} \quad (3)$$

$$MaP = \frac{\sum_{i=1}^{|\mathfrak{C}|} \frac{tp_{c_i}}{tp_{c_i} + fp_{c_i}}}{|\mathfrak{C}|} = MaP_0 + \frac{1}{|\mathfrak{C}|}\left(\frac{tp_{c_k}}{tp_{c_k} + fp_{c_k}} - \frac{tp_{c_{k\,old}}}{tp_{c_{k\,old}} + fp_{c_{k\,old}}}\right) \quad (4)$$

$$MaR = \frac{\sum_{i=1}^{|\mathfrak{C}|} \frac{tp_{c_i}}{tp_{c_i} + fn_{c_i}}}{|\mathfrak{C}|} = MaR_0 + \frac{1}{|\mathfrak{C}|} \frac{tp_{c_k} - tp_{c_{k\,old}}}{tp_{c_k} + fn_{c_k}}. \quad (5)$$

Оптимальное значение *MaF*

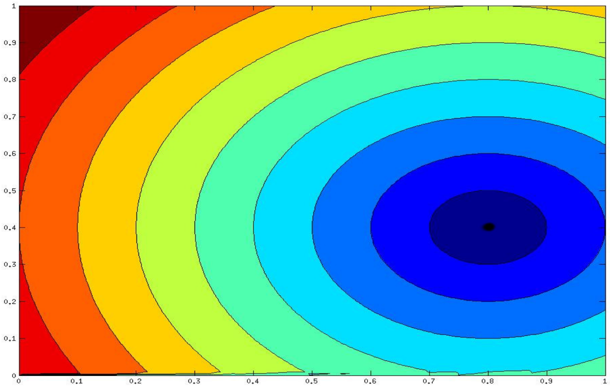
$$F : [0;1] \times [0;1] \to [0;1] \times [0;1]$$



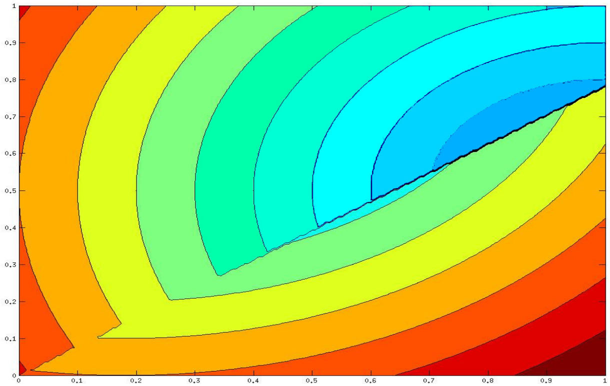
Утверждение 1

Если функция *MaF* достигает максимального значения в точке (P, R), то (P, R) является неподвижной точкой отображения F, задаваемого последовательным итерационным процессом.

Неподвижная точка



Парадокс: где неподвижная точка?



Существование неподвижной точки MaF

Утверждение 2

Пусть $B_0 = \{(P, R) | F(P_0, R_0) = (P, R)\}$. В выпуклой оболочке точек $(P, R) \in B_0$ существует неподвижная точка последовательного итерационного алгоритма.