

Вероятностные модели и методы оценки качества эталонных массивов текстов при классификации

Васильев В.Г. (vvg_2000@mail.ru)

ИПИБ МГУ им. Ломоносова

Каппа-статистики

Общий вид статистик

$$\kappa = \frac{A_0 - A_e}{A_{max} - A_e}$$

Проблемы использования

1. Сложность интерпретации результатов
2. Нет прямой связи показателями качества классификации

Использование оценок ошибок экспертов

1. Lam C. P., Stork D. G. Evaluating classifiers by means of test data with noisy labels

Теоретические оценки влияния ошибок экспертов на величину ошибки классификации, ее дисперсию, размер тестовых выборок,

2. Webber W., Oard D. W., Scholer F. Assessor Error in Stratified Evaluation

Анализ видов и вероятностей ошибок, допускаемых оценщиками

3. Gulin A., Kuralenok I., Pavlov D. Winning The Transfer Learning Track of Yahoo!'s Learning To Rank Challenge with YetiRank

Улучшения функций ранжирования документов

4. Metricov P., Pavlu V., Aslam J. A. Impact of Assessor Disagreement on Ranking Performance

Оценка верхних границ для показателей качества информационного поиска

Модели ошибок экспертов

Основные элементы моделей

$\Omega = \{\omega_1, \dots, \omega_k\}$ - множество классов.

$c^0 = (c_1^0, \dots, c_k^0)$, $c_j^0 \in \{0,1\}$ – ненаблюдаемый истинный вектор эталонной классификации;

$\hat{c}^0 = (\hat{c}_1^0, \dots, \hat{c}_k^0)$, $\hat{c}_j^0 \in \{0,1\}$ – наблюдаемый экспертный вектор эталонной классификации;

$c^1 = (c_1^1, \dots, c_k^1)$, $c_j^1 \in \{0,1\}$ – наблюдаемый вектор оцениваемой (автоматической) классификации.

Основные модели

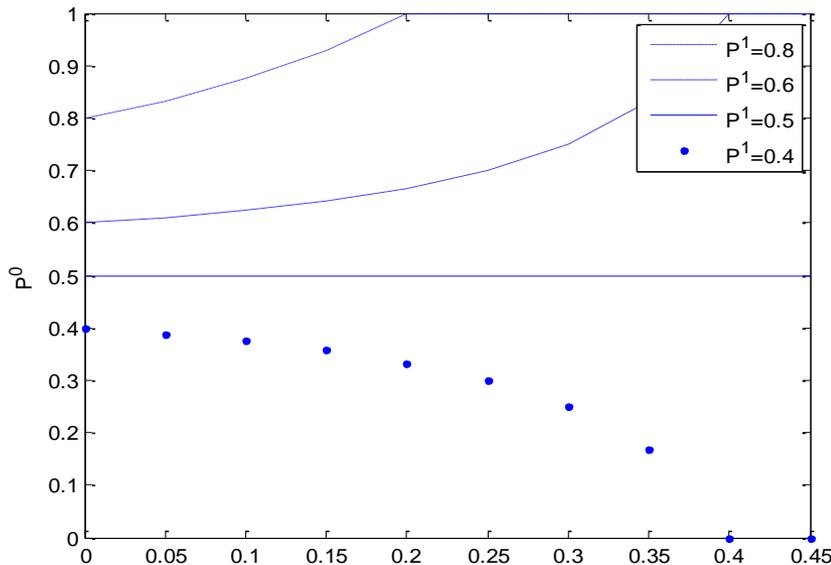
- Модель независимых ошибок;
- Модель условных ошибок.

Модель независимых ошибок

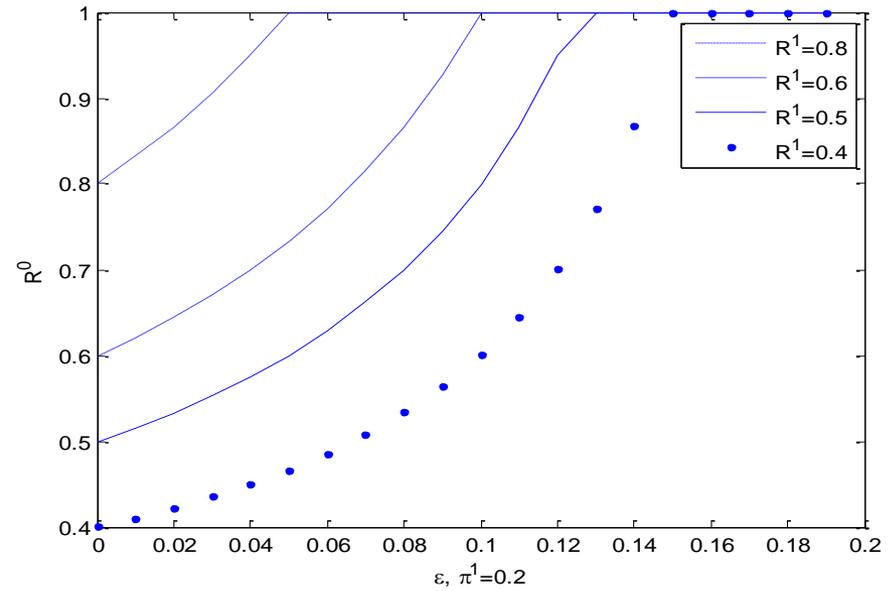
Взаимосвязь истинной и экспертной классификации:

$$\hat{c}^0 = c^0(1 - z) + (1 - c^0)z,$$

$z \sim \text{Ver}(\epsilon)$ – ошибка.



$$P^0 = \frac{P^1 - \epsilon}{1 - 2\epsilon}$$



$$R^0 = \frac{R^1 \hat{\pi} - \epsilon \pi^1}{\hat{\pi} - \epsilon}$$

Модель условных ошибок

Взаимосвязь истинной и экспертной классификации:

$$\hat{c}^0 = c^0(1 - z^1) + (1 - c^0)z^2,$$

$z^1 \sim \text{Ber}(\alpha)$, $z^2 \sim \text{Ber}(\beta)$ - ошибки первого и второго рода.

Оценка истинных значений показателей:

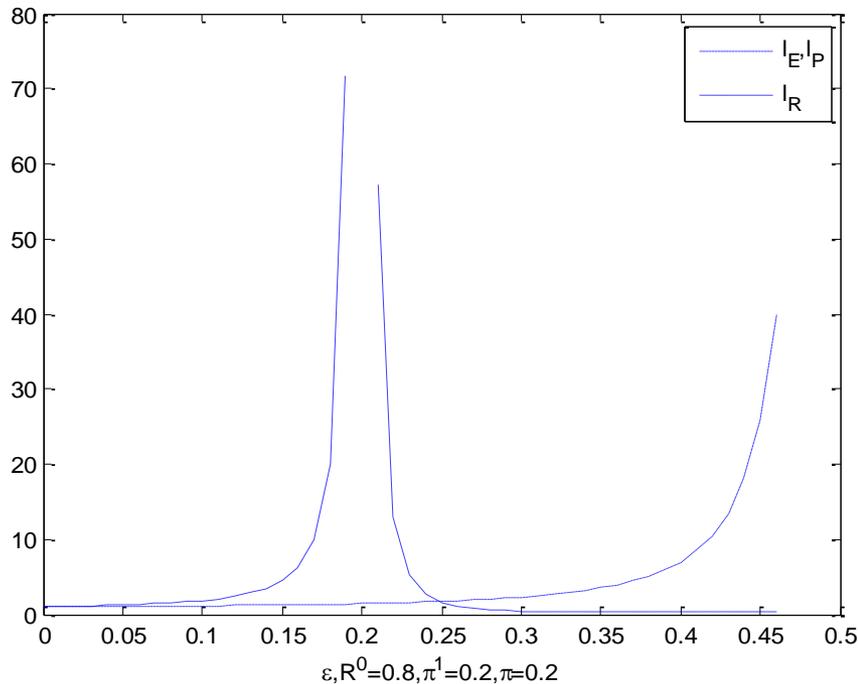
$$P^0 = \frac{P^1 - \beta}{1 - \alpha - \beta}, \quad R^0 = \frac{R^1 \hat{\pi} - \beta \pi^1}{\hat{\pi} - \beta}$$

Оценка интервалов значений наблюдаемых показателей:

$$P^1 \in (\beta, 1 - \alpha), \quad R^1 \in \left(\frac{\pi^1 \beta}{\pi(1 - \alpha - \beta) + \beta}, 1 - \frac{(1 - \pi^1) \beta}{\pi(1 - \alpha - \beta) + \beta} \right)$$

Оценка размеров обучающих выборок с учетом ошибок

Модель независимых ошибок



Модель условных ошибок

Увеличение размеров выборки для сохранения точности оценивания

$$l_P = \frac{\epsilon(1-\epsilon)}{(1-\alpha-\beta)^2} + P^0 \frac{\alpha-\beta}{(1-\alpha-\beta)} + 1$$

$$l_R = \frac{\beta(R^0(\hat{\pi}-\beta)(1-2\pi^1)+\pi^1(\hat{\pi}-\beta\pi^1))}{(\hat{\pi}-\beta)^2} + 1$$

Подход к оцениванию вероятностей ошибок

Представление массива документов

$$\Psi = (\psi_1, \dots, \psi_s),$$

$x_l = \{x_{l1}, \dots, x_{lm_l}\}$ - набор дубликатов.

c_l^0 – ненаблюдаемых истинная оценка.

$c_l = (c_{l1}, \dots, c_{lm_l})$ - вектор наблюдаемых оценок.

Расширенная функция правдоподобия

$$L(c_1, c_1^0, \dots, c_s, c_s^0 | \alpha, \beta, \pi)$$

Подходы к нахождению оценок:

- прямая максимизация функции правдоподобия,
- использование EM-алгоритма.

EM-алгоритм оценивания вероятностей ошибок

1. Инициализация.

$$t = 0, \alpha^{(t)} = \beta^{(t)} = 0.01, \pi^{(t)} = \frac{1}{2}.$$

2. E-Шаг. Вычисление апостериорных вероятностей

$$g_l = P(c_l^0 = 1 | c_{l1}, \dots, c_{lm_l}, \alpha, \beta, \pi).$$

3. M-Шаг. Вычисление оценок параметров

$$\alpha^{(t+1)}, \beta^{(t+1)} \text{ и } \pi^{(t+1)}$$

4. Проверка критерия завершения работы.

Оценивание вероятностей ошибок экспертов (РОМИП 2012 Массив отзывов о книгах)

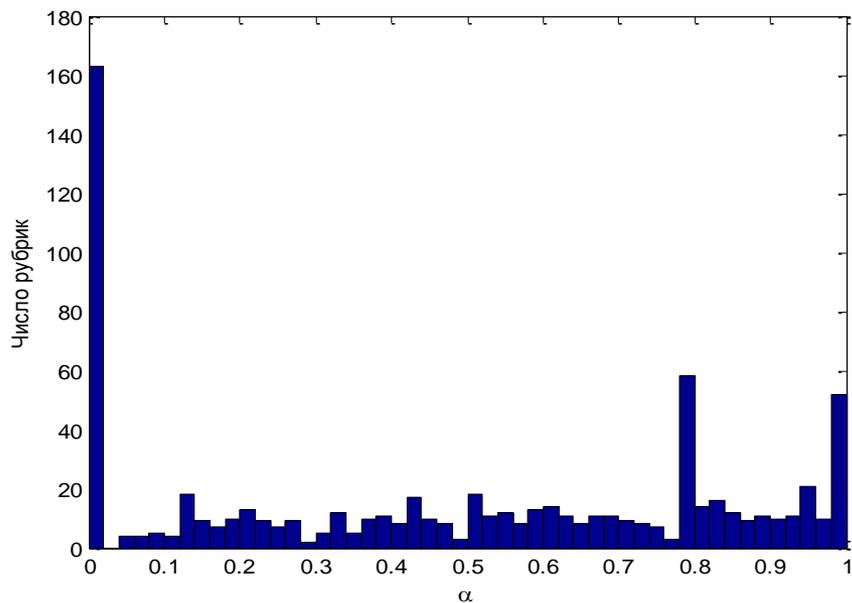
	π	ϵ	α	β
2-класса	10%	17%	12%	1%
	90%	2%	1%	12%
3-класса	6%	1%	10%	1%
	30%	9%	11%	9%
	57%	8%	4%	14%
5-классов	1%	1%	98%	1%
	4%	3%	17%	2%
	16%	7%	10%	7%
	26%	15%	11%	16%
	32%	9%	5%	12%

Оценивание вероятностей ошибок экспертов (РОМИП 2012 Массив отзывов о книгах)

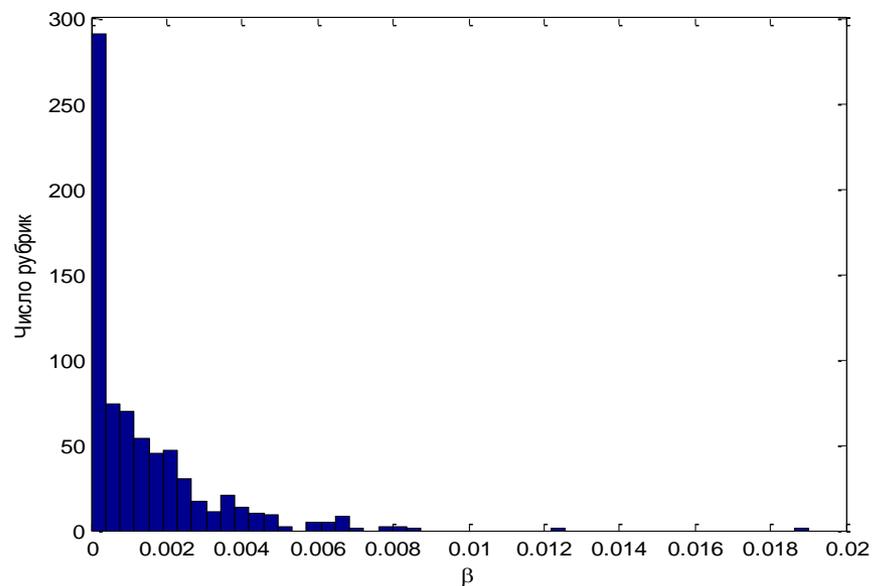
Оценки минимальных и максимальных
значения точности и полноты

	Класс 1 (отрицательные отзывы)	Класс 2 (положительные отзывы)
Точность	1%-88%	12%-99%
Полнота	1%-91%	11%-98%

Оценивание вероятностей ошибок при наличии одной экспертной классификации (РОМИП 2009 Дорожка классификации нормативно правовых документов)



Ошибка первого рода

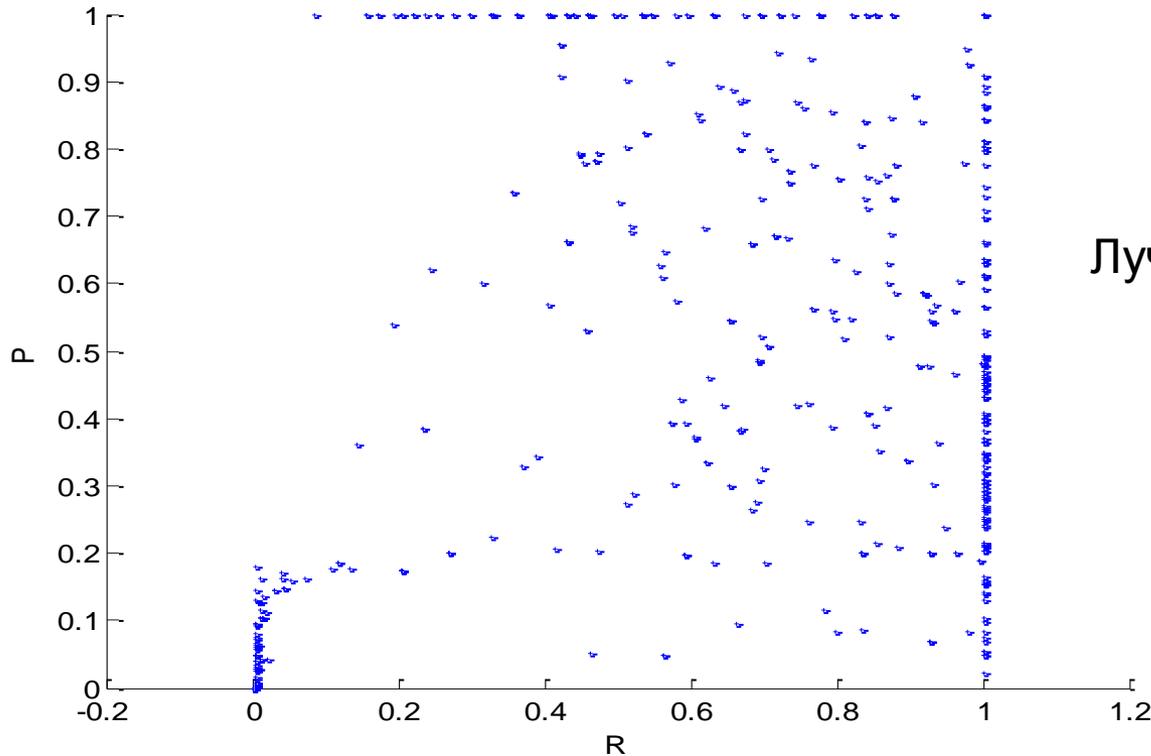


Ошибка второго рода

Оценивание вероятностей ошибок при наличии одной экспертной классификации (РОМИП 2009 Дорожка классификации нормативно правовых документов)

Оценки максимальных значений точности и полноты

Средние значения максимальных значений точности и полноты



$$P = 0.53$$

$$R = 0.63.$$

Лучшие результаты участников

$$P = 0.35$$

$$R = 0.45$$

Заключение

Прикладные задачи

1. Вычисление ошибок при наличии одной эталонной классификации
2. Вычисление истинных значения показателей качества
3. Оценка максимальных значения показателей качества
4. Оценка дисперсии показателей и размера обучающих массивов
5. Выявление проблемных рубрик

Перспективные задачи

1. Оценивание матриц условных вероятностей ошибок
2. Оценивание зависимостей между ошибками для разных рубрик
3. Использование оценок ошибок при обучении