Вероятностные модели и методы оценки качества эталонных массивов текстов при классификации

© В. Г. Васильев
ООО «ЛАН-ПРОЕКТ»,
г. Москва
vvg_2000@mail.ru

Аннотация

В работе рассматриваются вероятностные модели ошибок экспертов при формировании эталонных массивов текстов, а также методы их вычисления. В рамках данных моделей находятся взаимосвязи между истинными и наблюдаемыми показателями качества, определяются размеры тестовых выборок и максимальные значения показателей качества. Приводятся примеры вычисления ошибок на материалах дорожек РОМИП.

1 Введение

При практическом построении средств автоматической классификации возникает большое количество различных проблем, связанных сложностью и недостатками исходных данных, ограниченностью существующих методов классификации и др. [7],[10],[12],[15].

При оценке качества классификации обычно производится сравнение результатов автоматической классификации с результатами ручной классификации, выполненной экспертами. При этом предполагается, что в эталонной ручной классификации ошибки отсутствуют. Однако на практике эксперты при оценивании документов также совершают ошибки, которые могут быть вызваны различными причинами: невнимательностью, случайными опечатками, неоднозначностью наименования рубрик, низкой квалификацией экспертов в рассматриваемой предметной области, большим количеством рубрик и др. В результате получаемые оценки качества являются искаженными и даже для полностью правильной классификации показатели качества могут отличаться от своих максимальных значений.

Развитие специальных сервисов в сети Интернет, которые обеспечивают привлечение к работе по формированию эталонных массивов текстов большого количества анонимных пользователей, дополнительно повышают актуальность исследований в области оценки качества получаемых таким образом массивов.

Оценку качества эталонных массивов проводить в следующих двух ситуациях.

1. Эталонный массив подготовлен ранее неизвестными экспертами – в данном случае у документа имеется только одна оценка, полученная неизвестным экспертом, и нет возможности управления работой экспертов.

2. Эталонный массив формируется известными экспертами – в данном случае для каждого документа имеется фиксированное число оценок, выполненных известными экспертами, и можно управлять процессом оценки документов.

В современной литературе основное внимание уделяется второй ситуации и оценка качества эталонного массива часто сводится к простой оценке степени согласованности мнений экспертов.

Для оценки согласованности мнений экспертов разработано много различных коэффициентов и показателей. При этом наибольшее распространение получили методы [1],[3],[5],[14], основанные на использовании различных вариантов статистики κ, которая имеет следующий вид:

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

где A_0 - базовая статистика, оценивающая согласованность мнений экспертов, A_e - оценка значения A_0 в случае выполнения случайной классификации объектов, A_{max} - оценка максимально возможного значения A_0.

Для проверки гипотезы о наличии статистически значимого отклонения меры согласованности от нулевого значения требуется знание распределения κ. На практике такое распределение удается аналитически описать только для простейших случаев. По этой причине для проверки соответствующих гипотез обычно используют бутстреп метод [9]. В некоторых работах [2] предлагают использовать следующие неформальные оценки значений κ. Если $\kappa < 0.4$, то считается, что мнения не согласованы, если $0.4 \le \kappa < 0.75$, то считается, что мнения слабо согласованы, и, если $\kappa > 0.8$, то мнения сильно согласованы. Однако такой подход является не совсем корректным, так как с ростом числа объектов статистически значимыми являются и меньшие отклонения κ от 0.

Основным же недостатком данных методов является то, что значения статистики κ напрямую не связаны со значениями показателей качества классификации.

Отдельные вопросы прямого влияния ошибок экспертов на качество классификации и информационного поиска также рассматривались в ряде работ. В частности, в [6] приводятся теоретические оценки влияния ошибок экспертов на величину ошибки классификации, ее дисперсию, размер тестовых выборок, в [12] проводится анализ вероятностей ошибок, допускаемых оценщиками в различных ситуациях, в [4] информация о вероятностях ошибок экспертов использовалась для улучшения функций ранжирования документов, а в [8] для оценки верхних границ для показателей качества информационного поиска. Основной проблемой, которая пока не получила эффективного решения, при этом является нахождение соответствующих оценок вероятностей ошибок экспертов.

Рассмотрим теперь формальное описание моделей ошибок экспертов, в рамках которых можно явным образом оценивать их вероятности и влияние на показатели качества классификации.

2 Вероятностные модели ошибок экспертов

2.1 Общее описание вероятностных моделей

Для анализа влияния ошибок в эталонном массиве на показатели качества классификации будем считать, что имеется объект (текст) x, который может быть одновременно отнесен к нескольким классам из множества $\Omega = \{\omega_1, \dots, \omega_k\}$.

Результаты классификации отдельного объекта x можно представить с помощью следующих векторов:

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

$\hat{c}^0 = (\hat{c}_1^0, \dots, \hat{c}_k^0)$, $\hat{c}_j^0 \in \{0,1\}$ – наблюдаемый экспертный вектор эталонной классификации объекта x (данный вектор может отличаться от истинного вектора из-за наличия ошибок);

$c^1 = (c_1^1, \dots, c_k^1)$, $c_j^1 \in \{0,1\}$ – наблюдаемый вектор оцениваемой (автоматической) классификации объекта x (данный вектор может отличаться от истинного вектора),

где $c_j^0, \hat{c}_j^0, c_j^1 = 1$, если объект x относится к классу ω_j, и $c_j^0, \hat{c}_j^0, c_j^1 = 0$, в противном случае, $j = 1, \dots, k$.

Соответственно результаты классификации n объектов $x_1, \dots, x_n$, которые распределены по k классам $\omega_1, \dots, \omega_k$, могут быть представлены с помощью следующих матриц размера $k \times n$:

$C^0 = \left(c_{ji}^0\right)_{k \times n}$ – ненаблюдаемая истинная матрица эталонной классификации, в которой нет ошибок;

$\hat{C}^0 = \left(\hat{c}_{ji}^0\right)_{k \times n}$ – наблюдаемая экспертная матрица эталонной классификации, в которой есть ошибки;

$C^1 = \left(c_{ji}^1\right)_{k \times n}$ – наблюдаемая матрица автоматической классификации, качество которой оценивается,

где $c_{ji}^0, \hat{c}_{ji}^0, c_{ji}^1 = 1$, если объект x_i относится к классу ω_j, и $c_{ji}^0, \hat{c}_{ji}^0, c_{ji}^1 = 0$, в противном случае, $i = 1, \dots, n$, $j = 1, \dots, k$.

Основные показатели качества классификации могут быть представлены в виде следующих вероятностей:

$P_j^0 = P(\, c_j^0 = 1 |\, c_j^1 = 1)$, $P_j^1 = P(\, \hat{c}_j^0 = 1 |\, c_j^1 = 1)$ – истинное и наблюдаемое значение точности;

$R_j^0 = P(\, c_j^1 = 1 |\, c_j^0 = 1)$, $R_j^1 = P(\, c_j^1 = 1 |\, \hat{c}_j^0 = 1)$ – истинное и наблюдаемое значение полноты;

$E_j^0 = P(\, c_j^0 \neq c_j^1)$, $E_j^1 = P(\, \hat{c}_j^0 \neq c_j^1)$ – истинное и наблюдаемое значение ошибки классификации;

$F_j^0 = \frac{2P\left(c_j^0 = 1, c_j^1 = 1\right)}{P\left(c_j^0 = 1\right) + P\left(c_j^1 = 1\right)}$, $F_j^1 = \frac{2P\left(\hat{c}_j^0 = 1, c_j^1 = 1\right)}{P\left(\hat{c}_j^0 = 1\right) + P\left(c_j^1 = 1\right)}$ – истинное и наблюдаемое значение F-меры.

При этом для обозначения вероятностей классов будем использовать следующие обозначения: $\pi_j^0 = P\left(c_j^0 = 1\right), \pi_j^1 = P\left(c_j^1 = 1\right), \hat{\pi}_j = P\left(\hat{c}_j^0 = 1\right)$.

Далее будем считать, что зафиксирован класс $\omega_j, j = 1, \dots, k$, и все показатели качества, а также элементы векторов и матриц классификации, для сокращения записи будем записывать без индекса j. Например, элементы $c_j^0, \hat{c}_j^0, c_j^1$ будем записывать $c^0, \hat{c}^0, c^1$.

Рассмотрим следующие модели ошибок экспертов:

- модель независимых ошибок – предполагается, что ошибки носят случайный характер и не зависят от значений истинного вектора эталонной классификации;

- модель условных ошибок – предполагается, что ошибки, совершаемые экспертом, зависят от значений истинного вектора эталонной классификации.

2.2 Модель независимых ошибок экспертов

В рамках данной модели взаимосвязь истинной и экспертной классификации можно представить в виде следующего соотношения:

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

где $z \sim Ber(\epsilon)$ – независимая случайная величина, $\epsilon \in [0,1]$ – вероятность успеха, т.е. $P(z = 1) = \epsilon$ и $P(z = 0) = 1 - \epsilon$. Заметим, что при $\mathrm{z} = 1$ справедливо $\hat{\mathrm{c}}^0 \neq \mathrm{c}^0$, а при $\mathrm{z} = 1$ справедливо $\hat{\mathrm{c}}^0 = \mathrm{c}^0$.

Можно показать, что вероятности ошибок первого и второго рода, а также ошибки классификации совпадают и равны ϵ. Также справедливо свойство о независимости ошибок экспертной и автоматической классификации.

Утверждение 1. Пусть $\epsilon \neq \frac{1}{2}$, тогда в рамках модели независимых ошибок справедливы следующие соотношения между истинными и наблюдаемыми значениями показателей качества классификации:

$E^0 = \frac{E^1-\epsilon}{1-2\epsilon}$, $E^1 = E^0(1-2\epsilon)+\epsilon$, где $E^1 \geq \epsilon$,

$P^0 = \frac{P^1-\epsilon}{1-2\epsilon}$, $P^1 = P^0(1-2\epsilon)+\epsilon$, где $P^0 \geq \epsilon$,

$R^0 = \frac{R^1\hat{\pi}-\epsilon\pi^1}{\hat{\pi}-\epsilon}$, $R^1 = \frac{\left((1-2\epsilon)\pi^0R^0+\epsilon\pi^1\right)}{(1-2\epsilon)\pi^0+\epsilon}$, где $R^1\hat{\pi} \geq \epsilon\pi^1$

$F^0 = \frac{\frac{1}{2}F^1(\pi^1+\hat{\pi})-\epsilon\pi^1}{(1-2\epsilon)\pi^1+(\hat{\pi}-\epsilon)}$, $F^1 = 2\frac{\epsilon\pi^1+(1-2\epsilon)(\pi^1+\pi^0)F^0}{\pi^1+\pi^0(1-2\epsilon)+\epsilon}$. ■

Таким образом, с использованием выражений, приведенных в утверждении 1, можно зная уровень ошибок экспертов восстанавливать истинные значения показателей качества классификации по наблюдаемым экспертным показателям.

Заметим, что при $\epsilon = \frac{1}{2}$ экспертные оценки показателей качества становятся не связанными с истинными значениями показателей качества, так как в данном случае $E^1 = 1/2$, $P^1 = 1/2$, $R^1 = \pi^1$, что не позволяет восстанавливать значения истинных показателей качества.

С использованием приведенных соотношений можно оценить диапазон изменения показателей качества при изменении уровня ошибок экспертов.

Следствие 1. При фиксированном значении $\epsilon \in (0,1)$ получаем, что $E^1 \in (\epsilon, 1-\epsilon)$, $P^1 \in (\epsilon, 1-\epsilon)$, $R^1 \in \left(\frac{\epsilon\pi^1}{(1-2\epsilon)\pi^0+\epsilon}, 1-\frac{\epsilon(1-\pi^1)}{(1-2\epsilon)\pi^0+\epsilon}\right)$.

На следующих рисунках приведены значения наблюдаемых показателей качества при фиксированных значениях истинных показателях качества и различных значениях ошибки.

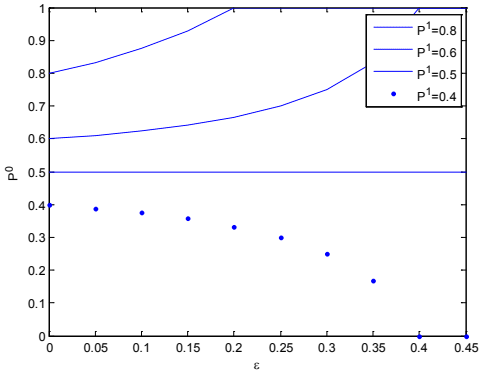

Рис. 1. График зависимости оценки истинной точности P^0 от ошибки эксперта при различных наблюдаемых значениях точности P^1

Из приведенного рисунка видно, что при наблюдаемых значениях точности меньше 0.5 при наличии ошибок истинные значения могут быть еще меньше. При наблюдаемой точности выше 0.5, напротив, истинные значения оказываются выше наблюдаемых значений.

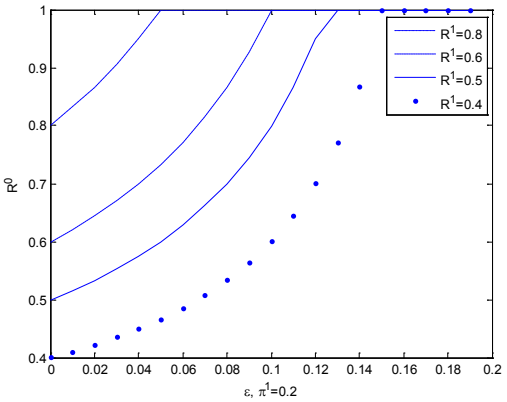

Рис. 2. График зависимости оценки истинной полноты R^0 от вероятности ошибки эксперта при различных фиксированных значениях наблюдаемой полноты R^1 и фиксированных значениях $\pi^1 = \hat{\pi} = 0.2$

Как можно заметить из приведенного рисунка даже при небольших значениях вероятности ошибки эксперта ϵ истинные и наблюдаемые значения полноты могут существенно отличаться.

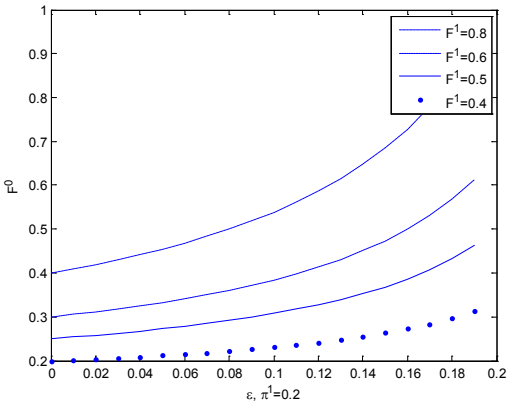

Рис. 3. График зависимости оценки истинной F-меры F^0 от вероятности ошибки эксперта ϵ при различных фиксированных значениях наблюдаемой F-меры F^1и фиксированных значениях $\pi^1 = \hat{\pi} = 0.2$

Как можно заметить из приведенного рисунка ошибка эксперта оказывает относительно меньшее влияние на значения F-меры, чем на значения полноты, но наблюдаемые значения все равно могут заметно отличаться.

2.3 Модель условных ошибок экспертов

В рамках данной модели предполагается, что вероятность ошибки эксперта зависит от того относится документ к рубрике или нет. Взаимосвязь истинной и экспертной классификации можно представить в виде следующего соотношения:

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

где $z^1 \sim Ber(\alpha)$ – независимая случайная величина, которая определяет ошибки первого рода, $z_2 \sim Ber(\beta)$ – независимая случайная величина, которая определяет ошибки второго рода.

Действительно, если $c^0 = 1$ и $z_1 = 1$, то $\hat{c}^0 = 0$, что соответствует ошибке первого рода. Если же $c^0 = 0$ и $z_2 = 1$, то $\hat{c}^0 = 1$, что соответствует ошибке второго рода.

Утверждение 2. В рамках модели независимых ошибок справедливы следующие соотношения между истинными и наблюдаемыми значениями показателей качества классификации:

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

$$R^1 = \frac{\pi(1-\alpha-\beta)R^0+\pi^1\beta}{\pi(1-\alpha-\beta)+\beta},\ R^0 = \frac{R^1\hat{\pi}-\beta\pi^1}{\hat{\pi}-\beta}. \blacksquare$$

Таким образом, если известны оценки вероятностей ошибок первого и второго рода для экспертов и наблюдаемые экспертные оценки точности и полноты, то можно вычислить истинные значения показателей точности и полноты.

Полученные соотношения между истинными и наблюдаемыми показателями качества позволяют оценить максимально возможные значения показателей качества, достижимые при определенном уровне ошибок.

Следствие 1. При фиксированных значениях $\alpha, \beta \in (0,1)$ получаем, что $P^1 \in (\beta, 1-\alpha)$, $R^1 \in \left(\frac{\pi^1\beta}{\pi(1-\alpha-\beta)+\beta}, 1-\frac{(1-\pi^1)\beta}{\pi(1-\alpha-\beta)+\beta}\right)$.

3 Оценка размеров эталонных массивов текстов

3.1 Оценка размеров эталонных массивов в рамках модели независимых ошибок

Для оценки размеров эталонных массивов текстов рассмотрим влияние, оказываемое ошибками экспертов на дисперсию выборочных оценок ошибки $\tilde{E}^1$, точности $\tilde{P}^1$ и полноты $\tilde{R}^1$, которые вычисляются следующим образом.

$$\tilde{E}^1 = \frac{1}{n}\sum_{i=1}^{n} I(\tilde{c}_i^0 \neq c_i^1),$$

$$\tilde{P}^1 = \frac{\sum_{i=1,c_i^1=1}^{n} I\,(\hat{c}_i^0 = 1)}{\sum_{i=1}^{n} I\,(c_i^1 = 1)} = \frac{\sum_{i=1}^{n} \hat{c}_i^0 c_i^1}{\sum_{i=1}^{n} c_i^1},$$

$$\tilde{R}^1 = \frac{\sum_{i=1,\hat{c}_i^0=1}^{n} I\,(\hat{c}_i^0 = 1)}{\sum_{i=1}^{n} I\,(\hat{c}_i^0 = 1)} = \frac{\sum_{i=1}^{n} \hat{c}_i^0 c_i^1}{\sum_{i=1}^{n} \hat{c}_i^0},$$

где $I(x) \in \{0,1\}$ – индикаторная функция, $\tilde{n}_0 = \sum_{i=1}^{n} \hat{c}_i^0$, $\tilde{n}_1 = \sum_{i=1}^{n} c_i^1$. Несложно показать, что данные статистики имеют следующие математические ожидания и дисперсии:

$$E\left(\tilde{E}^1\right) = E^1,\ D\left(\tilde{E}^1\right) = \frac{1}{n}E^1(1-E^1),$$

$$E\left(\tilde{P}^1\right) = P^1,\ D\left(\tilde{P}^1\right) = P^1(1-P^1)N^1,$$

$$E\left(\tilde{R}^1\right) = R^1,\ D\left(\tilde{R}^1\right) = R^1(1-R^1)\hat{N}^0,$$

где $N^1 = \sum_{s=1}^{n} \frac{1}{s}\binom{n}{s}(\pi^1)^s(1-\pi^1)^{n-s}$, $\hat{N}^0 = \sum_{s=1}^{n} \frac{1}{s}\binom{n}{s}(\hat{\pi})^s(1-\hat{\pi})^{n-s}. \blacksquare$

Из приведенного утверждения следует, что оценки показателей качества являются несмещенными, но при этом дисперсия является сложной функцией от размера выборки и вероятности успеха.

Для оценки показателей и зависимостей между истинными и наблюдаемыми показателями можно найти оценки истинных показателей с использованием следующих статистик.

$$\tilde{E}^0 = \tilde{P}(\mathrm{c}^0 \neq \mathrm{c}^1) = \frac{\tilde{E}^1 - \epsilon}{1-2\epsilon},$$

$$\tilde{P}^0 = \tilde{P}(\mathrm{c}^0 = 1 | c^1 = 1) = \frac{\tilde{P}^1 - \epsilon}{1-2\epsilon},$$

$$\tilde{R}^0 = \tilde{P}(c^1 = 1 | c^0 = 1) = \frac{\tilde{R}^1\tilde{\pi}^0 - \epsilon\tilde{\pi}^1}{\tilde{\pi}^0 - \epsilon},$$

Утверждение 3. Для статистик $\tilde{E}^0$, $\tilde{P}^0$ и $\tilde{R}^0$ справедливы следующие свойства для математических ожиданий и дисперсий:

$$E\left(\tilde{E}^0\right) = \frac{E^1-\epsilon}{1-2\epsilon},\ D\left(\tilde{E}^0\right) = \frac{1}{n}\left(\frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2} + E^0(1-E^0)\right),$$

$$E\left(\tilde{P}^0\right) = \frac{P^1-\epsilon}{1-2\epsilon},\quad D\left(\tilde{P}^0\right) = \left(\frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2} + P^0(1-P^0)\right)N^1,$$

$$E\left(\tilde{R}^0\right) = \frac{R^1\hat{\pi}-\epsilon\pi^1}{\hat{\pi}-\epsilon},\quad D\left(\tilde{R}^0\right) = \left(R^0(1-R^0) + \frac{\epsilon\left(R^0(\hat{\pi}-\epsilon)(1-2\pi^1)+\pi^1(\hat{\pi}-\epsilon\pi^1)\right)}{(\hat{\pi}-\epsilon)^2}\right)N^0,$$

где $N^1 = \sum_{s=1}^{n} \frac{1}{s}\binom{n}{s}(\pi^1)^s(1-\pi^1)^{n-s}$, $\hat{N}^0 = \sum_{s=1}^{n} \frac{1}{s}\binom{n}{s}(\hat{\pi})^s(1-\hat{\pi})^{n-s}$ ■

Заметим, что если бы имелась возможность напрямую подсчитать статистики для истинных значений показателей качества, то их дисперсии были бы равны следующим величинам:

$$D\left(\tilde{\tilde{E}}^0\right) = \frac{1}{n}E^0(1-E^0),$$

$$D\left(\tilde{\tilde{P}}^0\right) = P^0(1-P^0)N^1,$$

$$D\left(\tilde{\tilde{R}}^0\right) = R^0(1-R^0)N^0,$$

где $\tilde{\tilde{E}}^0 = \frac{1}{n}\sum_{i=1}^{n} I(c_i^0 \neq c_i^1)$, $\tilde{\tilde{P}}^0 = \frac{\sum_{i=1}^{n} I(c_i^0=1,c_i^1=1)}{\sum_{i=1}^{n} I(c_i^1=1)}$, $\tilde{\tilde{R}}^0 = \frac{\sum_{i=1}^{n} I(c_i^0=1,c_i^1=1)}{\sum_{i=1}^{n} I(c_i^0=1)}$.

Отсюда получаем, что справедливо следующее следствие из приведенного утверждения.

Следствие 1. Для обеспечения сохранения дисперсии оценок показателей на исходном уровне (соответствует ситуации, когда ошибки отсутствуют) требуется увеличение размера выборки в следующее число раз:

$l_E = \frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2} + 1$ – увеличение размера выборки для сохранения точности оценивания E^0,

$l_P = \frac{\epsilon(1-\epsilon)}{(1-2\epsilon)^2} + 1$ – увеличение размера выборки для сохранения точности оценивания показателя P^0,

$l_R = \frac{\epsilon\left(R^0(\hat{\pi}-\epsilon)(1-2\pi^1)+\pi^1(\hat{\pi}-\epsilon\pi^1)\right)}{(\hat{\pi}-\epsilon)^2} + 1$ – увеличение размера выборки для сохранения точности оценивания R^0.

На следующем рисунке показаны соответствующие зависимости для различных показателей качества.

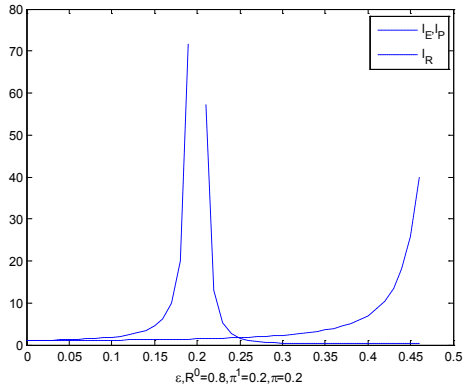

Рис. 4. Графики зависимости увеличения размера выборки для различных показателей качества от величины вероятности ошибки экспертной классификации

Из приведенного рисунка можно сделать вывод, что даже при относительно небольших значениях ошибки классификации может потребоваться существенное увеличение количества данных для обеспечения заданного уровня качества оценивания показателей. Более того чем меньше относительный размер класса, тем большее влияние оказывают случайные ошибки на показатели качества классификации. В частности, если размер класса составляет 20% от размера массива данных, то уже при уровне ошибки в 15% может потребоваться увеличение объема выборки в 10 раз.

3.2 Оценка размеров эталонных массивов в рамках модели условных ошибок экспертов

Для оценивания значений показателей качества воспользуемся статистиками $\tilde{P}^1$, $\tilde{R}^1$, которые были рассмотрены ранее. При этом можно использовать следующие статистики для оценки истинных значений показателей качества классификации:

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

Утверждение 4. Для статистик $\tilde{P}^0$ и $\tilde{R}^0$ справедливы следующие свойства для математических ожиданий и дисперсий:

$E\left(\tilde{P}^0\right) = \frac{P^1-\beta}{1-\alpha-\beta}$, $D\left(\tilde{P}^0\right) = \left(P^0(1-P^0) + \frac{\epsilon(1-\epsilon)}{(1-\alpha-\beta)^2} + P^0\frac{\alpha-\beta}{(1-\alpha-\beta)}\right)N^1$,

$E\left(\tilde{R}^0\right) = \frac{R^1\hat{\pi}-\beta\pi^1}{\hat{\pi}-\beta}$,

$D\left(\tilde{R}^0\right) = \left(R^0(1-R^0) + \frac{\beta\left(R^0(\hat{\pi}-\beta)(1-2\pi^1)+\pi^1(\hat{\pi}-\beta\pi^1)\right)}{(\hat{\pi}-\beta)^2}\right)N^0$,

где $N^1 = \sum_{s=1}^{n}\frac{1}{s}\binom{n}{s}(\pi^1)^s(1-\pi^1)^{n-s}$, $\widehat{N}^0 = \sum_{s=1}^{n}\frac{1}{s}\binom{n}{s}(\hat{\pi})^s(1-\hat{\pi})^{n-s}$ ■

Отсюда получаем, что справедливо следующее следствие из приведенного утверждения.

Следствие 1. Для обеспечения сохранения дисперсии оценок показателей на исходном уровне (соответствует ситуации, когда ошибки отсутствуют) требуется увеличение размера выборки в следующее число раз:

$l_P = \frac{\epsilon(1-\epsilon)}{(1-\alpha-\beta)^2} + P^0\frac{\alpha-\beta}{(1-\alpha-\beta)} + 1$ – увеличение размера выборки для сохранения точности оценивания показателя P^0,

$l_R = \frac{\beta\left(R^0(\hat{\pi}-\beta)(1-2\pi^1)+\pi^1(\hat{\pi}-\beta\pi^1)\right)}{(\hat{\pi}-\beta)^2} + 1$ – увеличение размера выборки для сохранения точности оценивания R^0.

4 Оценка вероятностей ошибок экспертов

4.1 Общее описание подхода

Для возможности практического использования выявленных зависимостей между истинными и наблюдаемыми значениями показателей качества классификации необходимо знать значения вероятностей ошибок экспертов. Однако их оценка является достаточно сложной задачей по следующим причинам:

1. Истинные матрицы эталонных классификаций являются неизвестными, что не позволяет вычислить ошибки экспертов напрямую;

2. В большинстве случаев доступной является только одна матрица экспертной классификации, что не позволяет оценивать качество работы одних экспертов по отношению к другим экспертам.

В ситуации, когда доступна только одна эталонная экспертная классификация массива документов, можно воспользоваться методами кластерного анализа для выявления «почти дубликатов» документов. Такой подход, в частности, подробно рассматривается и применяется в работах [10] и [4]. При отсутствии ошибок у документов, которые являются «почти дубликатами», должны быть одинаковые векторы классификации. При наличии же ошибок данные вектора будут отличаться.

Результаты выявления «почти дубликатов» или повторного оценивания объектов (документов) экспертами из множества $x_1, \dots, x_n$ можно представить в виде набора кластеров $\Psi = (\psi_1, \dots, \psi_s)$, где $\psi_l = \{x_{l1}, \dots, x_{lm_l}\}$, m_l – число элементов в кластере, s – число кластеров (число документов с повторной оценкой экспертами).

Пусть, как и ранее, зафиксирован некоторый класс ω_j, $j = 1, \dots, k$. Тогда каждому кластеру ψ_l, $l = 1, \dots, s$, можно поставить в соответствие $c_l^0 \in \{0,1\}$ – истинный признак относимости к классу ω_j и вектор наблюдаемых экспертных оценок $c_l = (c_{l1}, \dots, c_{lm_l})$, где $c_{lt} \in \{0,1\}$.

Рассмотрим теперь более подробно оценивание вероятностей ошибок экспертов в рамках модели независимых ошибок и в рамках модели условных ошибок.

4.2 Оценка вероятностей ошибок в рамках модели независимых ошибок

В рамках модели независимых ошибок справедливы следующие равенства:

$$c_{lt} = c_l^0 + z_{lt} - 2c_l^0 z_{lt},$$

где $z_{lt} \in \{0,1\}$ – независимая случайная величина, $P(z_{lt} = 1) = \epsilon, t = 1, .., m_l$.

Для нахождения оценки вероятности ошибки рассмотрим два подхода:

- прямая максимизация функции правдоподобия,

- использование ЕМ-алгоритма.

Нахождение оценки вероятности ошибки путем максимизации специальной функции правдоподобия. В данном случае для решения поставленной задачи рассмотрим для каждого кластера величину $u_l \in \{0,1\}, l = 1, \dots, s$, которая принимает значение равное 1, если $c_{l1} = c_{l2} = \dots = c_{lm_l}$, и 0, в противном случае. Тогда для $l = 1, \dots, s$ справедливо следующее равенство

$$P(u_l = 1) = P(c_{l1} = \dots = c_{lm_l}) = (1-\epsilon)^{m_l} + \epsilon^{m_l}.$$

Из приведенного утверждения следует, что $P(u_l = 1)$ является функцией от вероятности ошибки ϵ, но при этом для вычисления значений u_l не требуется знание истинных значений c_l^0, $l = 1, \dots, s$. Это свойство позволяет для нахождения ϵ воспользоваться методом максимального правдоподобия. В данном случае оценка ϵ является решением следующей оптимизационной задачи:

$$\epsilon^* = \arg\max_{\epsilon} L(u_1, \dots u_s|\epsilon),$$

где $L(u_1, \dots u_s|\epsilon) = \sum_{l=1}^{s} \log(P(u_l = 1)^{u_l}(1 - P(u_l = 1))^{1-u_l})$ - логарифм функции правдоподобия.

Можно показать, что максимум $L(u_1, \dots u_s|\epsilon)$ находится как решение следующего уравнения:

$$\sum_{l=1}^{s} m_l(\epsilon^{m_l-1} - (1-\epsilon)^{m_l-1})\left(\frac{u_l}{(1-\epsilon)^{m_l} + \epsilon^{m_l}} - \frac{(1-u_l)}{1-(1-\epsilon)^{m_l} - \epsilon^{m_l}}\right) = 0.$$

Прямое решение данного уравнения является достаточно сложной задачей. По этой причине для его решения можно воспользоваться численными методами. В тоже время в частном случае, когда $m_l = 2, l = 1, \dots, s$, можно найти точное решение данного уравнения.

Утверждение 5. При $m_l = 2, l = 1, \dots, s$ и $\epsilon < \frac{1}{2}$ максимум функции правдоподобия $L(u_1, \dots u_s|\epsilon)$ достигается при

$$\epsilon^* = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{2}{s}\sum_{l=1}^{s} u_l - 1}.$$

Таким образом, в ситуации, когда для каждого документа имеется только две оценки, возможно явное нахождение оценки вероятности ошибки эксперта. В общем же случае, требуется применение итерационных методов.

Нахождение оценки вероятности ошибки с использованием ЕМ-алгоритма. В данном случае для нахождения оценки вероятности ошибки рассмотрим расширенную функцию правдоподобия $L(c_1, c_1^0, \dots, c_s, c_s^0|\epsilon, \pi)$, в которую входят наблюдаемые признаки $c_l = (c_{l1}, \dots, c_{lm_l})$ и не наблюдаемые признаки $c_l^0, l = 1, \dots, s$, где $\pi = P(c_l^0 = 1)$, ϵ – вероятность ошибки эксперта.

С учетом приведенных обозначений в рамках модели независимых ошибок справедливо следующее равенство

$$logL(c_{11}, \dots, c_{1m_l}, c_1^0, \dots, c_{s1}, \dots, c_{sm_s}, c_s^0|\epsilon) = \sum_{l=1}^{s}\Big(c_l^0 \log\pi + (1-c_l^0)\log(1-\pi) + \sum_{t=1}^{m_l}\big(c_l^0(c_{lt}\log(1-\epsilon) + (1-c_{lt})\log\epsilon) + (1-c_l^0)(c_{lt}\log\epsilon + (1-c_{lt})\log(1-\epsilon))\big)\Big).$$

В соответствии с общей схемой построения ЕМ-алгоритма требуется решение следующих двух задач:

- найти условное математическое ожидание расширенной функции правдоподобия при фиксированных неизвестных параметрах (Е-шаг);

- найти максимум условного математического ожидания расширенной функции правдоподобия по неизвестным параметрам (М-шаг).

В данном случае в рамках Е-шага требуется найти следующее условное математическое ожидание:

$$E(logL(c_1, c_1^0, \dots, c_s, c_s^0, \epsilon, \pi)|c_1, .., c_s, , \epsilon, \pi).$$

Его вычисление сводится к нахождению следующих апостериорных вероятностей

$$g_l = P(c_l^0 = 1|c_{l1}, \dots, c_{lm_l}, \epsilon, \pi), l = 1, \dots, s,$$

которые можно вычислить следующим образом:

$$g_l = \left(1 + \left(\frac{1}{\pi} - 1\right)\left(\frac{1}{\epsilon} - 1\right)^{m_l - 2n_l}\right)^{-1},$$

где $n_l = \sum_{t=1}^{m_l} c_{lt}$ - число единиц в векторе результатов экспертной классификации.

Отсюда получаем выражение для нахождения математического ожидания логарифма функции правдоподобия:

$$E(logL(c_1, c_1^0, \dots, c_s, c_s^0, \epsilon, \pi)|c_1, .., c_s, \epsilon, \pi) = \sum_{l=1}^{s}\Big(g_l \log\pi + (1-g_l)\log(1-\pi) + \sum_{t=1}^{m_l}\big(g_l(c_{lt}\log(1-\epsilon) + (1-c_{lt})\log\epsilon) + (1-g_l)(c_{lt}\log\epsilon + (1-c_{lt})\log(1-\epsilon))\big)\Big).$$

Найдем теперь в рамках М-шага решение следующей задачи:

$$(\epsilon^*, \pi^*) = \arg\max_{\epsilon,\pi} E(logL(c_1, c_1^0, \dots, c_s, c_s^0, \epsilon, \pi)|c_1, .., c_s, \epsilon, \pi).$$

Можно показать, что максимум будет достигаться при

$$\pi^* = \frac{1}{s}\sum_{l=1}^{s} g_l,$$

$$\epsilon^* = \frac{\sum_{l=1}^{s}\sum_{t=1}^{m_l}(g_l + c_{lt} - 2g_l c_{lt})}{\sum_{l=1}^{s} m_l} = \frac{\sum_{l=1}^{s}(m_l g_l + n_l - 2g_l n_l)}{\sum_{l=1}^{s} m_l}$$

Для задания начальных значений параметров ϵ и π можно положить ϵ равной небольшому числу больше 0, например, $\epsilon = 0.01$, а $\pi = \frac{1}{2}$. В качестве критерия завершения работы алгоритма можно использовать два условия: число итераций равно t_{max} – положительное целое число, разница между новым и старым значениями ϵ меньше $\Delta_\epsilon \in (0,1)$.

Теперь можно описать ЕМ-алгоритм оценивания вероятности ошибки экспертов полностью. В качестве входных параметров у него выступают следующие: t_{max}, Δ_ϵ, $c_l = (c_{l1}, \dots, c_{lm_l})$, $l = 1, \dots, s$, - вектора экспертных классификаций.

ЕМ-алгоритм оценивания вероятности ошибки экспертов

1. Инициализация. Положить $t = 0$, $\epsilon^{(t)} = 0.01$, $\pi^{(t)} = \frac{1}{2}$, $n_l = \sum_{j=1}^{m_l} c_{lj}$, $l = 1, \dots, s$.

2. Е-Шаг. Вычислить для $l = 1, \dots, s$ апостериорные вероятности $g_l^{(t)}$ с использованием следующего выражения:

$$g_l^{(t)} = \left(1 + \left(\frac{1}{\pi^{(t)}} - 1\right)\left(\frac{1}{\epsilon^{(t)}} - 1\right)^{m_l - 2n_l}\right)^{-1}.$$

3. М-Шаг. Вычислить оценки параметров $\epsilon^{(t+1)}$ и $\pi^{(t+1)}$ с использованием следующих выражений:

$$\epsilon^{(t+1)} = \frac{\sum_{l=1}^{s}\left(g_l^{(t)}(m_l - 2n_l) + n_l\right)}{\sum_{l=1}^{s} m_l},$$

$$\pi^{(t+1)} = \frac{1}{s}\sum_{l=1}^{s} g_l^{(t)}.$$

4. Критерий завершения работы. Положить $t = t + 1$. Если $t > t_{max}$ или $\left|\epsilon^{(t+1)} - \epsilon^{(t)}\right| < \Delta_\epsilon$, то завершить работу алгоритма, в противном случае, перейти к шагу 2. ■

4.3 Оценка вероятностей ошибок в рамках модели условных ошибок

В рамках модели независимых ошибок справедливы следующие равенства:

$$c_{lt} = c_l^0(1 - z_{lt}^1) + (1 - c_l^0)z_{lt}^2,$$

где $z_{lt}^1, z_{lt}^2 \in \{0,1\}$ – независимые случайные величины, $P(z_{lt}^1 = 1) = \alpha$, $P(z_{lt}^2 = 1) = \beta$, $t = 1, .., m_l$.

Для нахождения оценок значений параметров α и β воспользуемся методом максимального правдоподобия и построим соответствующий ЕМ-алгоритм.

Рассмотрим расширенную функцию правдоподобия $L(c_1, c_1^0, \dots, c_s, c_s^0|\alpha, \beta, \pi)$, в которую входят наблюдаемые признаки $c_l = (c_{l1}, \dots, c_{lm_l})$ и не наблюдаемые признаки $c_l^0, l = 1, \dots, s$, где $\pi = P(c_l^0 = 1)$.

С учетом приведенных обозначений в рамках модели независимых ошибок справедливо следующее равенство

$$logL(c_1, c_1^0, \dots, c_s, c_s^0|\alpha, \beta, \pi) = \sum_{l=1}^{s}\Big(c_l^0 \log\pi + (1 - c_l^0)\log(1 - \pi) + \sum_{t=1}^{m_l}\big(c_l^0(c_{lt}\log(1 - \epsilon) + (1 - c_{lt})\log\epsilon) + (1 - c_l^0)(c_{lt}\log\epsilon + (1 - c_{lt})\log(1 - \epsilon))\big)\Big).$$

В данном случае в рамках Е-шага требуется найти следующее условное математическое ожидание:

$$E(logL(c_1, c_1^0, \dots, c_s, c_s^0|\alpha, \beta, \pi)|c_1, .., c_s, \alpha, \beta, \pi).$$

Несложно заметить, что его вычисление сводится к нахождению следующих апостериорных вероятностей

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

которые можно записать в следующем виде:

$$g_l = \left(1 + \left(\frac{1}{\pi} - 1\right)\left(\frac{\beta}{1 - \alpha}\right)^{n_l}\left(\frac{1 - \beta}{\alpha}\right)^{m_l - n_l}\right)^{-1},$$

где $n_l = \sum_{t=1}^{m_l} c_{lt}$.

Отсюда получаем выражение для нахождения математического ожидания логарифма функции правдоподобия:

$$E(logL(c_1, c_1^0, \dots, c_s, c_s^0|\alpha, \beta, \pi)|c_1, .., c_s, \alpha, \beta, \pi) = \sum_{l=1}^{s}\Big(g_l\log\pi + (1 - g_l)\log(1 - \pi) + \sum_{t=1}^{m_l}\big(g_l(c_{lt}\log(1 - \alpha) + (1 - c_{lt})\log\alpha) + (1 - g_l)(c_{lt}\log\beta + (1 - c_{lt})\log(1 - \beta))\big)\Big).$$

Найдем теперь в рамках М-шага решение следующей задачи:

$$(\alpha^*, \beta^*, \pi^*) = \arg\max_{\alpha,\beta,\pi} E(logL(c_1, c_1^0, \dots, c_s, c_s^0|\alpha, \beta, \pi)|\mathrm{c}, \alpha, \beta, \pi).$$

Можно показать, что максимум будет достигаться при следующих значениях параметров:

$$\pi^* = \frac{1}{s}\sum_{l=1}^{s} g_l,$$

$$\alpha^* = 1 - \frac{\sum_{l=1}^{s} g_l n_l}{\sum_{l=1}^{s} g_l m_l},$$

$$\beta^* = \frac{\sum_{l=1}^{s}(1 - g_l)n_l}{\sum_{l=1}^{s}(1 - g_l)m_l}.$$

Для задания начальных значений параметров α, β и π можно положить $\alpha = \beta = 0.01$, а $\pi = \frac{1}{2}$. В качестве критерия завершения работы алгоритма можно использовать два условия: число итераций равно t_{max} – положительное целое число, разница

между новым и старым значениями α, β меньше $\Delta_\epsilon \in (0,1)$.

Теперь можно описать ЕМ-алгоритм оценивания вероятности ошибки экспертов полностью. В качестве входных параметров у него выступают следующие: t_{max}, Δ_ϵ, $c_l = (c_{l1}, \dots, c_{lm_l})$, $l = 1, \dots, s$, - вектора экспертных классификаций.

ЕМ-алгоритм оценивания условных вероятностей ошибок

1. Инициализация. Положить $t = 0$, $\alpha^{(t)} = \beta^{(t)} = 0.01$, $\pi^{(t)} = \frac{1}{2}$, $n_l = \sum_{j=1}^{m_l} c_{lj}$, $l = 1, \dots, s$.

2. Е-Шаг. Вычислить для $l = 1, \dots, s$ апостериорные вероятности $g_l^{(t)}$ с использованием следующего выражения:

$$g_l^{(t)} = \left(1 + \left(\frac{1}{\pi} - 1\right)\left(\frac{\beta}{1-\alpha}\right)^{n_l}\left(\frac{1-\beta}{\alpha}\right)^{m_l - n_l}\right)^{-1}.$$

3. М-Шаг. Вычислить оценки параметров $\alpha^{(t+1)}, \beta^{(t+1)}$ и $\pi^{(t+1)}$ с использованием следующих выражений:

$$\alpha^{(t+1)} = 1 - \frac{\sum_{l=1}^{s} g_l^{(t)} n_l}{\sum_{l=1}^{s} g_l^{(t)} m_l},$$

$$\beta^{(t+1)} = \frac{\sum_{l=1}^{s} (1 - g_l^{(t)}) n_l}{\sum_{l=1}^{s} (1 - g_l^{(t)}) m_l},$$

$$\pi^{(t+1)} = \frac{1}{s}\sum_{l=1}^{s} g_l^{(t)}.$$

4. Критерий завершения работы. Положить $t = t + 1$. Если $t > t_{max}$ или $\max\left(\left|\alpha^{(t+1)} - \alpha^{(t)}\right|, \left|\beta^{(t+1)} - \beta^{(t)}\right|\right) < \Delta_\epsilon$, то завершить работу алгоритма, в противном случае, перейти к шагу 2. ■

5 Примеры оценивания ошибок экспертов

5.1 Примеры оценивания вероятностей ошибок при наличии нескольких экспертных классификаций

Пример оценивания вероятности ошибки экспертов в рамках модели независимых ошибок. Для иллюстрации оценивания вероятностей ошибок на практике рассмотрим задачу построения классификаторов для оценивания мнений пользователей, которая предлагалась в рамках семинара РОМИП-2012. В рамках РОМИП для оценки качества работы систем вручную были сформированы 3 эталонных массива текстов (массив с оценками книг, массив с оценками фильмов, массив с оценками камер), в которых каждый текст был оценен двумя экспертами по 2-х бальной шкале, 3-х бальной шкале, 5 бальной шкале. В следующей таблице приведены оценки вероятностей ошибок экспертов и оценки вероятностей классов, полученные с помощью ЕМ-алгоритма для массива с оценками книг.

Таблица 1. Вероятности ошибок экспертов для массива с оценками книг

	2 класса	**3 класса**	**5 классов**
Ошибки	0.017 0.017	0.011 0.094 0.081	0.013 0.026 0.070 0.150 0.094
Вероятности классов	0.083 0.918	0.0519 0.294 0.626	0.000 0.030 0.147 0.290 0.357

Приведенный пример показывает, что на практике величины ошибок могут быть достаточно большими и существенно отличаться для различных классов. Знание вероятностей ошибок позволяет получить оценки истинных значений показателей качества классификации, оценить объем исходных данных, необходимых для получения требуемой точности оценивания показателей качества.

Пример оценивания условных ошибок экспертов в рамках дорожки по классификации тональности оценок пользователей РОМИП-2012. Для иллюстрации оценивания вероятностей ошибок рассмотрим опять массив с оценками книг, который был сформирован в рамках РОМИП-2012. В следующей таблице приведены оценки вероятности ошибок первого и второго рода экспертов для различного числа классов, а также оценки вероятности ошибок, которые были получены в рамках модели независимых ошибок.

Таблица 2. Вероятности ошибок экспертов для массива с оценками книг

	$\boldsymbol{\pi}$	$\boldsymbol{\epsilon}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
2-класса	0.103 0.897	0.0166 0.0166	0.120 0.0061	0.006 0.121
3-класса	0.063 0.302 0.572	0.0110 0.0944 0.0811	0.0998 0.1055 0.0437	0.006 0.090 0.137
5-классов	0.006 0.042 0.155 0.261 0.323	0.013 0.026 0.070 0.150 0.094	0.983 0.174 0.095 0.113 0.054	0.013 0.021 0.066 0.164 0.115

Приведенные данные показывают, что ошибки могут принимать достаточно большие значения и при этом заметно отличаться для различных классов. Это приводит к тому, что оценки качества, получаемые на таком массиве, могут существенно отличаться от истинных значений.

С учетом найденных ошибок можно оценить максимально достижимые значения показателей точности и полноты с использованием следующих соотношений:

$$P^1 \in (\beta, 1-\alpha),$$

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

Максимальные и минимальные значения показателей для массива с оценками книг приведены в следующей таблице.

Таблица 3. Оценки максимальных и минимальных значений для точности и полноты при классификации на 2 класса

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

Полученные результаты показывают, что при классификации отрицательных отзывов ошибки могут быть значительно выше, чем при классификации положительных отзывов.

5.2 Пример оценивания вероятностей ошибок при наличии одной экспертной классификации

Рассмотрим теперь пример оценивания вероятностей ошибок экспертов в ситуации, когда имеется только одна матрица эталонной классификации. В качестве массива текстов возьмем материалы дорожки классификации нормативно-правовых документов, которая проводилась в рамках РОМИП-2009. Обучающее множество содержит 29943 документа, которые распределены по 721 классу.

Для получения оценивания вероятности ошибок экспертов проведем выявление «дубликатов» документов. При этом будем считать, что документы являются дубликатами, если мера косинусной близости между векторами документов будет больше 0.9. Непосредственный просмотр документов, мера близости между которыми больше данного порога, показал, что они действительно являются почти дубликатами.

В результате оценивания ошибок экспертов первого и второго рода представлены в форме гистограмм распределения значения ошибок по рубрикам на следующих двух рисунках (такая форма выбрана из-за большого числа рубрик).

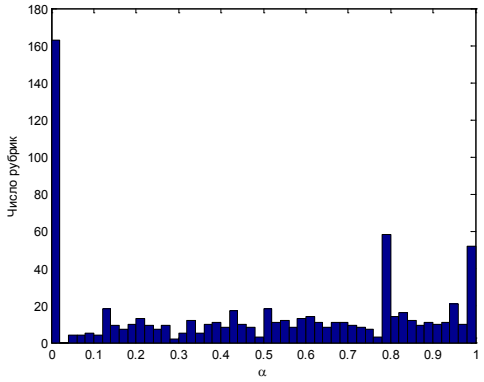

Рис. 5. Гистограмма распределения ошибок первого рода по рубрикам

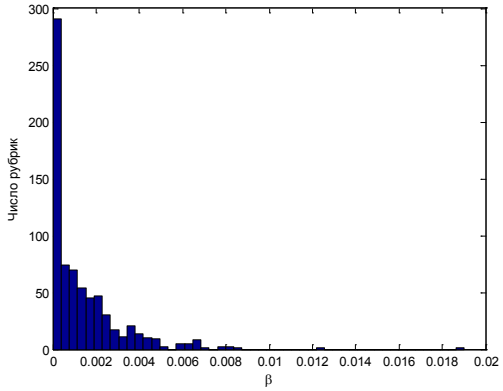

Рис. 6. Гистограмма распределения ошибок второго рода по рубрикам

Анализ полученных результатов показывает, что ошибки первого рода принимают достаточно большие значения и значительно больше ошибок второго рода, что соответствует известному эмпирическому наблюдению, что эксперты при ручной классификации чаще пропускают рубрики, чем добавляют неправильные.

На следующем рисунке также приведены максимальные значения показателей точности и полноты для рубрик, которые вычислены с использованием полученных оценок для ошибок первого и второго рода.

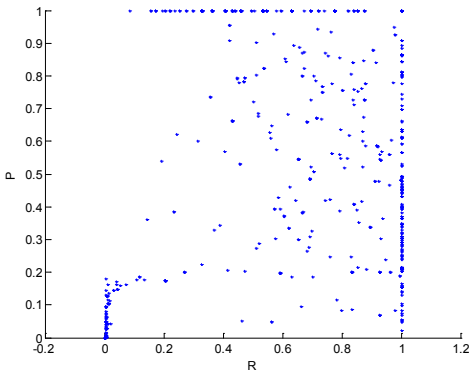

Рис. 7. Значения максимальных значений показателей точности и полноты для рубрик, полученные в рамках модели условных ошибок

Средние значения максимальных значений показателей точности и полноты по всем рубрикам равны следующим значениям:

$$maxP = 0.53,$$
$$maxR = 0.63.$$

Сравнение полученных максимальных значений показателей качества с теми, которые были достигнуты участниками дорожки (максимальное значение точности – 35%, максимальное значение полноты было равно 45%), объясняет получение участниками низких значений показателей качества.

6 Заключение

В работе рассмотрены две модели ошибок экспертов, а также предложен подход к оцениванию их вероятностей, основанный на использовании EM-алгоритма.

Разработанные модели и методы позволяют решать следующие прикладные задачи:

- вычислять значения ошибок экспертов как при наличии нескольких, так и при наличии только одной эталонной экспертной классификации;

- восстанавливать истинные значения показателей качества классификации по наблюдаемым экспертным оценкам значений соответствующих показателей;

- вычислять максимально возможные значения показателей качества классификации при данном уровне ошибок экспертов;

- оценивать величину дисперсии показателей качества и определять размер тестовых выборок, необходимый для обеспечения требуемой точности их оценивания, в зависимости от уровня ошибок экспертов;

- определять рубрики, которые требуют более внимательного оценивания.

Предложенный подход к оценке вероятностей ошибок экспертов является достаточно общим и его можно обобщить и для случая оценивания матриц условных вероятностей, рассматриваемых в работах [4] и [8].

Практическое использование предложенных моделей и методов показано на примерах оценивания ошибок экспертов и максимальных значений показателей качества на материалах дорожек РОМИП-2009 и РОМИП-2012.

Литература

[1] Cohen J. A coefficient of agreement for nominal scales // Educ. Psychol. Measurement. - 1960: Vol. 20. - p. 37-46.

[2] Eye A., Mun E. Y. Analyzing Rater Agreement: Manifest Variable Methods: Taylor and Francis, 2006. – 190 p.

[3] Fleiss J. L. Measuring nominal scale agreement among many raters // Psychological Bulletin. - 1971: Vol. 76. - p. 378-382.

[4] Gulin A., Kuralenok I., Pavlov D. Winning The Transfer Learnin Track of Yahoo!'s Learning To Rank Challenge with YetiRank // Journal of Machine Learning Research, Vol. 14, 2011. - p. 63-76.

[5] Gwet K. L. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters: Advanced Analytics, LLC, 2010. – 294 p.

[6] Lam C. P., Stork D. G. Evaluating classifiers by means of test data with noisy labels // Proceedings of the International Joint Conference on Artificial Intelligence, AAAI Press, 2003. – p. 513-518.

[7] Lewis D. D., Sebastiani F. Report on the Workshop on Operational Text Classification systems (OTC-01) // SIGIR Forum. - 2001. - 2: Vol. 35. - p. 8-11.

[8] Metricov P., Pavlu V., Aslam J. A. Impact of Assessor Disagreement on Ranking Performance // SIGIR'12. - Portland, Oregon, USA, 2012. - p. 1091-1092.

[9] Reichenheim M. E. Confidence intervals for the kappa statistic // The Stata Journal. - 2004: Vol. 4. - p. 421-428.

[10] Scholer F., Turpin A., Sanderson M. Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements // SIGIR'11, Beijing, China, 2011. - p. 1063-1072.

[11] Sebastiani F. Machine learning in automated text categorization // ACM Comput. Surv. - 2002. - 1: Vol. 34. - p. 1-47.

[12] Webber W., Oard D. W., Scholer F. Assessor Error in Stratified Evaluation // CIKM'10, Toronto, Ontario, 2010. - p. 539-548.

[13] Агеев М. С., Добров Б. В., Лукашевич Н. В. Поддержка системы автоматического рубрицирования для сложных задач классификации текстов // Труды 6-ой Всероссийской научной конференции – RCDL2004, 2004. – 10 с.

[14] Заславский А. А., Пригарина Т. А. Оценка согласованности субъективных классификаций при заданных классах // Социология. - 1994: Vol. 3-4. - с. 84-109.

[15] Лукашевич Н. В. Тезаурусы в задачах информационного поиска. - М.: Издательство Московского Университета, 2011. - 512 с.

Probabilistic models and methods for classifier etalon datasets quality estimation

Vitaly G. Vasilyev

In this paper two probabilistic models of expert errors and special iterative methods of their estimation are proposed. By using this framework the expert errors, size of etalon datasets, maximum values for quality metrics can be calculated. Examples of real calculations are shown on materials of the ROMIP tracks.