УДК 004.852
DOI: 10.25206/2310-4597-2025-1-38-41

Оценка эффективности метрик валидации в задачах классификации текстов

The effectiveness of validation metrics in text classification tasks

А. В. Коротких
Омский государственный технический университет, Омск, Российская Федерация

A. V. Korotkikh
Omsk state technical university, Omsk, Russian Federation

Аннотация. При решении задач классификации текстов исследователям необходимо оценивать эффективность работы алгоритмов на проверочных данных. В данной работе особое внимание уделено метрикам оценки при классификации текстов. В работе приведены наиболее часто используемые метрики, проведён сравнительный анализ метрик, даны рекомендации по их выбору. Представлены значения различных метрик при решении задачи оценки трудоёмкости задач разработки программного обеспечения, как задачи классификации текстов, а так же задачи классификации новостей. В качестве метрики позволяющей оценивать качество классификации на проверочной выборке предложено использование коэффициента корреляции Мэтьюса. Сделаны выводы о необходимости применения данной метрики взамен значения функции потерь на проверочной выборке.

***Ключевые слова*:** дисбаланс классов, матрица ошибок, коэффициент Мэтьюса, методы оценки трудоёмкости, машинное обучение, искусственный интеллект, классификация текстов на естественном языке, метрики валидации

Abstract. When solving text classification problems, researchers need to evaluate the effectiveness of algorithms based on validation data. In this paper, special attention is paid to validation methods for text classification. The paper presents the most frequently used metrics, provides a comparative analysis of metrics, and provides recommendations on their selection. The values of various metrics are presented when solving the software effort estimation problem as text classification task, as well as news classification tasks. The use of the Matthews correlation coefficient is proposed as a metric for evaluating the quality of classification in a test dataset. Conclusions are drawn about the need to use this metric instead of the value of the loss function in the validation dataset.

***Keywords*:** class imbalance, confusion matrix, Matthews correlation coefficient (MCC), effort estimation methods, machine learning, artificial intelligence, natural language text classification, validation metrics

Введение

Существует большое количество задач, решаемых с помощью алгоритмов классификации текстов: категоризация новостей, определение цели пользовательского запроса в чате, анализ эмоций и тональности, классификация жанров и стилей, медицинская диагностика по текстовым описаниям, юридическая категоризация, категоризация научных статей и пр. Всё чаще для решения этих задач применяются нейронные сети глубокого обучения.

Обучение параметров моделей производится методом стохастического градиентного спуска. Кроме того требуется настройка так называемых гиперпараметров модели. Для оценки качества работы модели после настройки гиперпараметров обучающая выборка делится на части, одна из них используется для обучения, а другая для проверки качества работы модели и настройки гиперпараметров. Для сравнения качества классификации могут использоваться следующие метрики: среднее арифметическое значение функции потерь на проверочной выборке, F1-мера, коэффициент корреляции Мэтьюса, доля верных ответов. В данной работе показано, что использование для подбора гиперпараметров метрик, отличных от среднего арифметического значения функции потерь, повышает качество классификации для задач с дисбалансом классов и большим уровнем шума, таких как, определение трудоёмкости задач разработки программного обеспечения по текстовым описаниям этих задач. Целью данной работы является систематизация метрик валидации для классификации текстов, анализ их применимости в зависимости от характеристик исходных данных.

Теория

В процессе обучения глубоких нейронных сетей отслеживают значения функции потерь. При этом, если на обучающей выборке функция потерь убывает, то для проверочной выборки убывание функции на определённых этапах сменяется ростом. В этом случае говорят о так называемом «переобучении». Если за критерий

остановки обучения взять минимальное среднее значение функции потерь на проверочной выборке, то на этом этапе обучение останавливают и производят оценку качества работы модели с выбранными значениями гиперпараметров.

В работе [1] дано следующее определение матрицы ошибок: пусть дано множество образцов $S=\{s_i:1\leq i\leq C\}$, множество классов $N=\{1,\ldots,n\}$ и две функции $t(s_i)$ и $p(s_i)$, такие что $t(s_i)$ равно истинному классу, а $p(s_i)$ равно предсказанному моделью классу. Тогда можно задать матрицу ошибок $C\in M(|N|x|N|)$, $C_{ij}=|\{s\in S: t(s)=i \text{ и } p(s)=j\}|$. Полученная матрица ошибок наиболее полно характеризует качество работы обучаемой модели, но её невозможно использовать для автоматического подбора гиперпараметров, возникают сложности анализа и при ручном подборе с большим количеством вариантов. Поэтому используются показатели каким либо образом характеризующие матрицу ошибок.

Наиболее простой мерой оценки результатов классификации (1) является отношение числа правильно классифицированных образцов m к общему их числу n, через значения элементов матрицы ошибок выражаемое следующим образом:

$$A=\frac{m}{n}, \text{где} m=\sum_{k=1}^{N} C_{kk}\ ,\ \ n=\sum_{i,j=1}^{N} C_{ij} \tag{1}$$

Но данная метрика очень плохо отражает качество предсказаний модели в случае несбалансированных классов. Если количество элементов одного из классов преобладает над количеством элементов других классов, то модель, присваивающая всё время метку этого класса, будет получать высокие значения данной метрики. При этом такая модель очевидно практически не применима.

Так же легко интерпретируемой метрикой является F1-мера, данная метрика является средним гармоническим значением показателей точности и полноты вычисляется следующим образом:

$$F1=2\frac{P\cdot R}{P+R},\ \text{где } P \text{ и } R: \tag{2}$$

$$P_k=\frac{C_{kk}}{\sum_{j=1,j\neq k}^{N} C_{kj}},\ \ P=\frac{\sum_{k=1}^{N} P_k}{N}.$$

$$R_k=\frac{C_{kk}}{\sum_{i=1,i\neq k}^{N} C_{ik}},\ \ R=\frac{\sum_{k=1}^{N} R_k}{N}.$$

Коэффициент корреляции Мэттьюса первоначально был предложен для бинарной классификации, в дальнейшем в работах [2] и [1] был обобщён для многоклассовой классификации:

$$MCC=\frac{\sum_{k,l,m=1}^{N} C_{kk}C_{ml}-C_{lk}C_{km}}{\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{lk}\right)\left(\sum_{f,g=1f\neq k}^{N} C_{gf}\right)\right]}\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{kl}\right)\left(\sum_{f,g=1f\neq k}^{N} C_{fg}\right)\right]}}. \tag{3}$$

Коээфициент корреляции Мэттьюса лежит в диапазоне [-1;1], 0 – при случайном угадывании, 1 – когда все предсказанные значения совпадают с истинными, и около -1, когда нет ни одного верно предсказанного значения. В работе [1] показано, что коэффициент корреляции Мэттьюса устойчив при изменении числа классов, может использоваться как для бинарной, так и для многоклассовой классификации, а так же для несбалансированных наборов данных.

Результаты

Для анализа применимости метрик валидации были выбраны два набора данных: новости BBC и трудоёмкости задач разработки программного обеспечения. Первый набор был взят, как хорошо классифицируемый моделями глубокого обучения набор, на втором наборе показатели точности классификации ниже. Количество классифицируемых текстов для набора «Новости BBC» по категориям представлено на рис. 1. На рис. 2 представлено количество текстов для набора «Трудоёмкость задач».

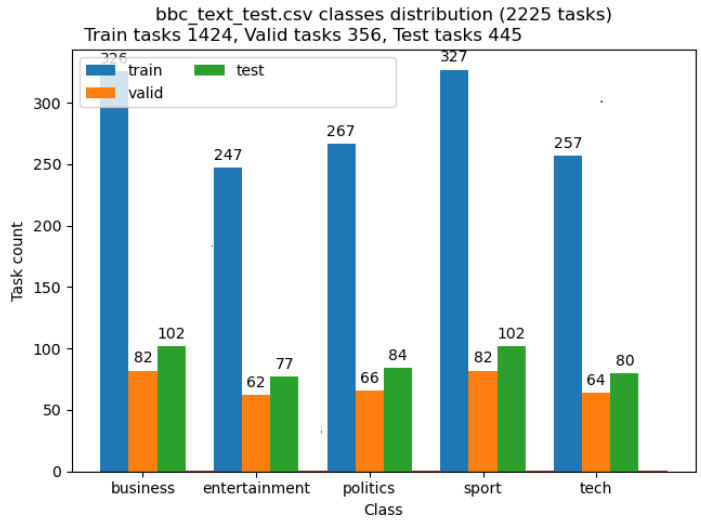

Рис. 1. Количество текстов по категориям
в наборе данных "Новости BBC"

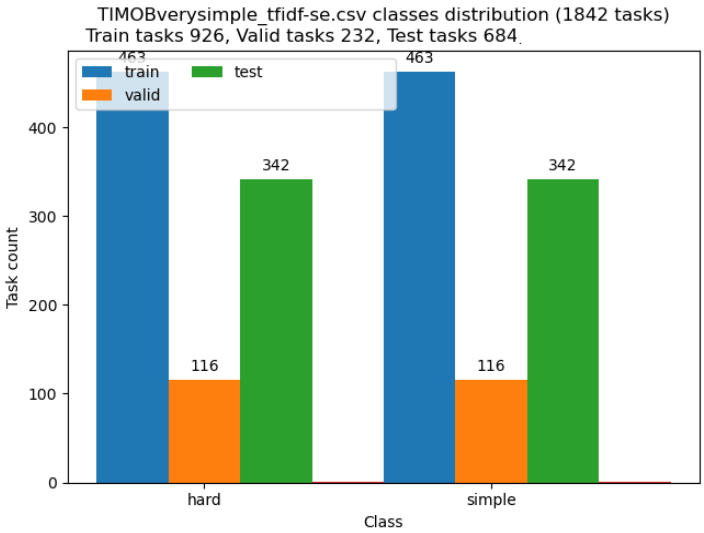

Рис. 2. Количество текстов по категориям
в наборе данных "Трудоёмкость задач"

Результаты экспериментов были получены с использованием следующих моделей классификации текстов:

1. на основе архитектуры GPT2 [3] к исходной модели добавлен классифицирующий линейный слой;

2. на основе архитектуры GPT2 [3] с добавлением промежуточного слоя с 768 входами, 768 выходами и функцией активации GELU[4], а затем классифицирующий линейный слой.

В таблице 1 приведены значения метрик валидации, полученные на различных этапах обучения, а так же значения метрик на контрольной выборке рассчитанные после тестирования соответствующих моделей. В первой строке для каждой модели представлены показатели для наименьшего значения функции потерь на проверочной выборке, в следующей строке – для наибольшего значения MCC (см. формулу 3) на проверочной выборке и в последней строке значения для последней эпохи обучения. Результаты эксперимента показывают, что в 2х случаях из 4х наилучшие значения показателей точности на контрольной метрике были получены при выборе модели по лучшим значениям MCC полученным на проверочной выборке. И во всех случаях модель с наименьшим значением функции потерь на проверочной выборке показывала худшие результаты на контрольной выборке.

Таблица 1

Значение метрик валидации

Набор данных	Модель	Значение функции потерь на проверочной выборке (№ эпохи)	MCC		F1		A	
			Проверочная	Контрольная	Проверочная	Контрольная	Проверочная	Контрольная
«Трудоёмкость задач»	1	**0,59 (3)**	0,3104	0,2546	0,6551	0,6270	0,6552	0,6272
		0,92 (5)	**0,4579**	**0,3498**	**0,6782**	**0,6442**	**0,6983**	**0,6594**
		0,68 (6)	0,3365	0,2606	0,6679	0,6299	0,6681	0,6301
		2,35(15)	0,2760	0,2401	0,6378	0,6196	0,6379	0,6199
	2	**0,58 (3)**	0,4363	0,3326	0,6920	0,6432	**0,7026**	0,6550
		0,66 (4)	**0,4579**	**0,3546**	0,6782	**0,6451**	0,6983	**0,6608**
		1,89 (14)	0,3980	0,2459	**0,6977**	0,6189	0,6983	0,6213
		1,92 (15)	0,3443	0,2226	0,6642	0,6028	0,6681	0,6082
«Новости BBC»	1	**0,20** (5)	**0,9719**	0,9720	**0,9763**	0,9774	**0,9775**	0,9775
		0,34 (10)	0,9649	**0,9832**	0,9715	**0,9862**	0,9719	**0,9865**
	2	**0,14 (2)**	0,9684	0,9747	0,9742	0,9796	0,9747	0,9798
		0,17 (3)	**0,9789**	0,9747	**0,9834**	0,9796	**0,9831**	0,9798
		0,35 (10)	0,9647	**0,9775**	0,9714	**0,9823**	0,9719	**0,9820**

Выводы и заключение

Результаты эксперимента показали, что подход с остановкой обучения при минимальном значении функции потерь на проверочной выборке даёт не самые лучшие значения метрик MCC, F1 и A для проверочной и контрольной выборок. В двух случаях из четырёх лучшие показатели на контрольной выборке были получены при выборе модели по лучшему значению MCC на проверочной выборке, ещё в двух случаях лучшей оказалась модель обученная на большем количестве эпох обучения. Таким образом, можно сделать предположение, что необходимо продолжать обучение при наличии возможности и в дальнейшем из обученных моделей выбирать лучшую по значению MCC. Вывод о применимости данного правила делать рано, необходимо провести исследования на большем количестве наборов данных и алгоритмов классификации.

Список источников

1. Jurman G., Riccadonna S., Furlanello C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction // PLoS ONE. 2012. Vol. 7, no. 8. P. e41882. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0041882 (date accessed: 03.05.2025).

2. Gorodkin J. Comparing two *K*-category assignments by a *K*-category correlation coefficient // Computational Biology and Chemistry. 2004. Vol. 28, no. 5. P. 367–374.

3. Language Models are Unsupervised Multitask Learners / A. Radford, J. Wu, R. Child [et al.]. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (date accessed: 03.05.2025).

4. Hendrycks D., Gimpel K. Gaussian Error Linear Units (GELUs). 2023. URL: http://arxiv.org/abs/1606.08415 (date accessed: 28.04.2025).