УДК 004.852 **Информационные технологии**

Горячкин Б. С., *кандидат технических наук, доцент; Московский государственный технический университет им. Н.Э. Баумана*

Чечнев А. А., *магистр 2 года обучения, Московский государственный технический университет им. Н.Э. Баумана*

АНАЛИЗ ЧУВСТВИТЕЛЬНОСТИ МЕТРИК БИНАРНОЙ КЛАССИФИКАЦИИ К ДИСБАЛАНСУ ДАННЫХ

Аннотация: Статья посвящена проблеме выбора адекватной метрики для оценивания моделей бинарной классификации по матрице ошибок.

Приведено описание большого количества сравниваемых метрик, оценивающих качество модели бинарной классификации. Проведен анализ работы метрик по сгенерированным матрицам ошибок с различными степенями сбалансированности и процентами ошибок. Выявлены метрики, которые не зависят от дисбаланса классов, и при приведении их к одному числовому диапазону совпадают. На основании этого сделан вывод о метрике, позволяющей наилучшим образом оценивать как сбалансированные, так и не сбалансированные данные.

Ключевые слова: Метрика, бинарная классификация, дисбаланс классов, сбалансированные данные, несбалансированные данные, матрица ошибок.

Annotation: The article is devoted to the problem of choosing an adequate metric for evaluating binary classification models based on the error matrix.

A large number of comparable metrics that evaluate the quality of the binary classification model are described. The analysis of the work of metrics on the generated error matrices with different degrees of balance and error percentages is carried out. The metrics that do not depend on the class imbalance are identified, and when they are reduced to the same numerical range, they coincide. Based on this, a conclusion is

made about the metric that allows you to best evaluate both balanced and unbalanced data.

Key words: Metric, binary classification, class imbalance, balanced data, unbalanced data, confusion matrix.

Введение

Выбор правильной метрики имеет решающее значение при оценке моделей машинного обучения. Для оценки моделей ML в различных приложениях могут быть использованы различные метрики, поэтому цель данной работы является обзор популярных метрик для лучшего их понимания и определения задач, для которых они могут быть использованы. В некоторых случаях использование одной метрики может не дать полной картины решаемой проблемы, в связи с этим появляется необходимость выявить наиболее объективную интегральную оценку качества.

Задачи классификации подразделяются на бинарные (в случае, когда целевой класс может принимать два значения) и множественные. В задаче классификации предпочтительно, чтобы обучающие примеры были относительно равномерно распределены среди классов [1], однако в реальности практически все наборы данных в задачах бинарной классификации не сбалансированы, что приводит к проблеме качественной оценки работы классификатора. Классическая метрика аккуратности (accuracy) может выдавать высокое значение качества, при этом в реальности модель работает хуже прогноза, всегда определяющего один класс.

Целью данной статьи является исследование применимости различных метрик для задач бинарной классификации и выбор наилучшей, учитывающей дисбаланс классов.

Были исследованы 16 метрик, применяемых для задач бинарной классификации и проанализирована их применимость для сбалансированных и несбалансированных классов целевой переменной.

Описание метрик бинарной классификации

Часто результаты работы классификаторов оцениваются по матрицам ошибок (confusion matrix) [2], представленной в табл.1. Столбцы матрицы описывают значения истинного класса, строки – что предсказал классификатор. Бинарная целевая переменная принимает целочисленное значение (0, 1), поэтому будем считать, что положительный класс означает, что значение целевой переменной принимает единицу, а отрицательный – значение целевой переменной принимает 0. Матрица ошибок для бинарной классификации содержит следующие составляющие (табл. 1).

Таблица 1. Матрица ошибок

Предсказанный класс	Истинный класс	
	Положительный класс	Отрицательный класс
Положительный класс	Истинно положительные (TP)	Ложно положительные (FP)
Отрицательный класс	Ложно отрицательные (FN)	Истинно отрицательные (TN)

Исследуемые в данной работе метрики [3; 4], оценивающие близость двух бинарных множеств (предсказанных и реальных) по матрице ошибок, представлены в табл. 2.

Таблица 2. Исследуемые метрики

	Название метрики	Математическая формулировка
1	False positive rate (FPR), ложноположительный коэффициент	$\frac{FP}{FP + TN}$
2	False negative rate (FNR), ложноотрицательный коэффициент	$\frac{FN}{FN + TP}$
3	Sensitivity или Recall или TPR, Чувствительность	$\frac{TP}{TP + FN}$
4	Specificity или TNR, специфичность	$\frac{TN}{TN + FP}$
5	Precision, точность	$\frac{TP}{TP + FP}$
6	Accuracy, правильность	$\frac{TP + TN}{n}$
7	Jaccard index, индекс Жаккара	$\frac{TP}{TP + FN + FP}$
8	F1-score, гармоническое среднее	$\frac{2 * precision * recall}{precision + recall}$

9	F2 - score	$\frac{(1+\beta)*precision*recall}{\beta^2*\text{precision}+\text{recall}}$ при $\beta = 2$
10	Geometric mean (GM), геометрическое среднее	$\sqrt{precision*recall}$
11	Area under ROC curve (AUC), площадь под ROC-кривой	$\frac{sensitivity+specificity}{2}$
12	Cohen's kappa, каппа Коэна	$\frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{n^2}$
13	MCC – коэффициент корреляции Мэтьюса	Формула (1)
14	Confusion entropy (CEN), энтропия ошибки	Формула (2)
15	Discriminatory power (DP), степень разделимости	$k*\left(log\left(\frac{sensitivity}{1-sensitivity}\right)+log\left(\frac{specificity}{1-specificity}\right)\right)$
16	Youden's index, индекс Юдена	$sensitivity + specificity - 1$

Первые семь метрик можно отнести к одному классу. Их объединяет то, что в них используются не все составляющие матрицы ошибок и соответственно они не могут адекватно оценить качество классификации в одиночку. Первые пять метрик описывают долю содержания составляющих матрицы ошибок относительно общего числа значений соответствующего класса. Например, метрика *precision* (точность) показывает долю истинных срабатываний положительного класса, относительно всех предсказаний данного класса [5].

Наиболее простой для вычислений и популярной оценкой классификаторов является функция *accuracy.* По сути формула считает количество корректных предсказаний относительно всех значений, иными словами – долю правильных ответов. Однако она имеет парадоксальное свойство: в случае несбалансированных данных классификаторы с меньшим значением *accuracy* могут давать лучший прогноз при решении прикладных задач, чем классификаторы, имеющие более высокие значения этого параметра [6].

Метрики 8 – 10 в формулах содержат три составляющие матрицы ошибок и объединяют в себе точность (*precision*) и полноту (*recall)*. Данные метрики не учитывают истинно отрицательные (TN) и ориентированы на мажоритарный класс.

F-score в общем виде записывается следующим образом:

$$Fscore = \frac{(1+\beta)*precision*recall}{\beta^2*\text{precision}+\text{recall}} \qquad (1)$$

Параметр β позволяет варьировать важность одного показателя, относительно другого [7]. На практике часто используются метрики *F1, F2*. Функция *F1* – это простое гармоническое среднее между *recall* и *precision*. *F2* – вариант функции *F1*, в котором значение *precision* имеет больший вес. Функция *GM* – геометрическое среднее этих же величин.

Функции *F1* и *F2* так же, как *accuracy*, *recall* и *precision*, точнее оценивают результаты классификации доминирующего множества при несбалансированных данных [8].

Часто результат работы алгоритма на фиксированной тестовой выборке визуализируют с помощью *ROC-кривой* (*ROC* - receiver operating characteristic, иногда говорят «кривая ошибок»), а качество оценивают, как площадь под этой кривой – *AUC* (*AUC* - area under the curve).

Проблемой этой метрики является отсутствие вычислительной функции, она строится численно как функция *FPR* от *FNR*, при фиксации одной из составляющих и вычислении значения другой и ее невозможность ее построения для задач с множественной классификацией. Однако для бинарных задач ее можно аппроксимировано построить по одной матрице ошибок. Для этого необходимо найти площадь под кривой, составленной точками (0, 0), (*FPR*, *FNR*), (1, 1) [9].

В исходном варианте в формуле вычисления степени разделимости классов (*DP*) [10] коэффициент k принимает значение 0,5513. В таком случае функция может принимать большие значения, поэтому заменим константу k на логарифм для удобного представления на графике.

$$DR^* = log(\frac{DP}{k}) \qquad (2)$$

Коэффициент корреляции Мэтьюса (*MCC)* используется для оценки множественной классификации и может быть применен для бинарных задач. В общем

случае метрика оценивает степень корреляции предсказанных и истинных значений. Данный показатель принимает значения в диапазоне [-1; 1], где единица означает идеальный прогноз, ноль – случайный, а -1 – полностью противоположный [11]. Метрика может быть рассчитана по матрице ошибок следующим образом:

$$MCC = \frac{tp * fn - fp * fn}{\sqrt{(tp + fp)(fp + fn)(tn + fp)(tn + fn)}} \quad (3)$$

Функция *CEN* также используется для многоклассовых задач. Для бинарного случая в статье [12] приведена следующая формула:

$$CEN = \frac{(fn+fp)*log_2(n^2-(tp-tn)^2)}{2n} - \frac{fn*log_2(fn)+fp*log_2(fp)}{n} \quad (4)$$

В данной формуле присутствуют логарифмы от FN и FN, в связи с этим при отсутствии ошибок хотя бы в одном из классов, метрика не выдаст никакого результата. При минимальном количестве ошибок *CEN* стремится к нулю, а при общей ошибке в 50% (Случайное предсказание) стремится к единице. Модифицируем формулу так, чтобы большие значения соответствовали лучшему результату:

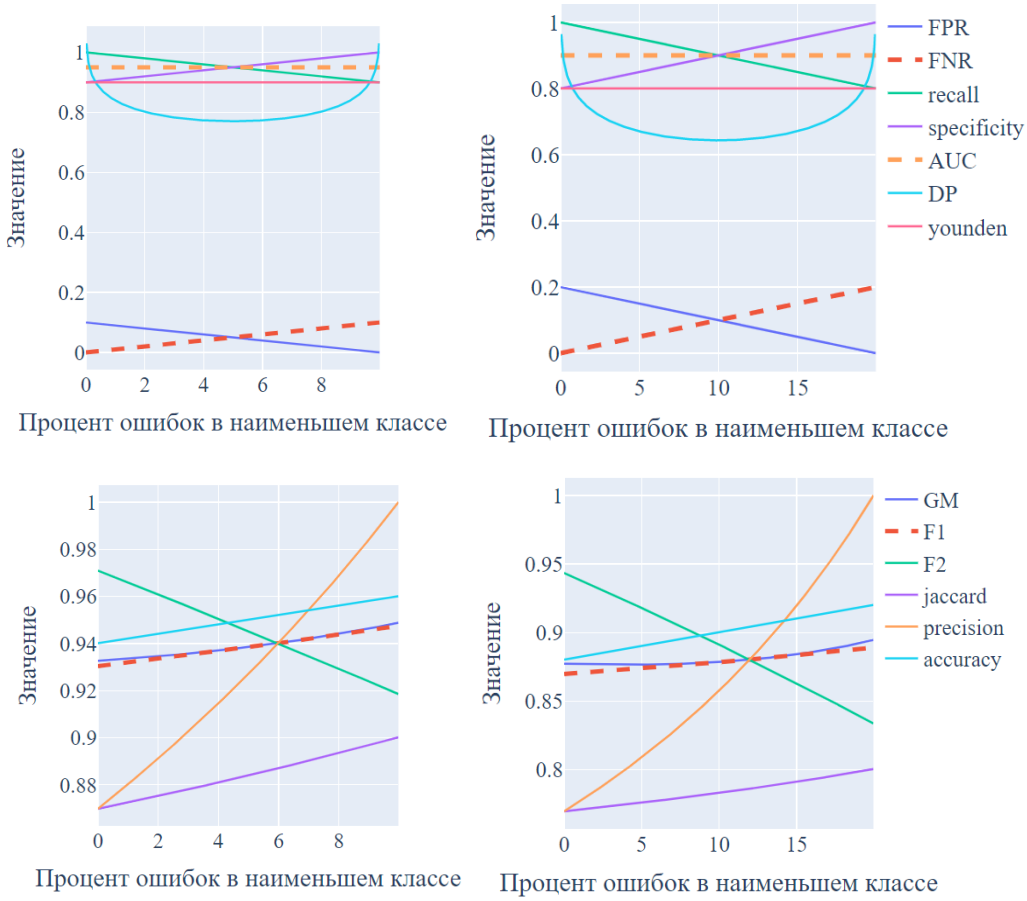
$$CEN^* = 1 - CEN \quad (5)$$

Анализ работы метрик

Для анализа работы метрик, представленных в табл. 2 будем генерировать матрицы ошибок с различными параметрами сбалансированности и распределениями ошибок по классам. Для обозначения степени несбалансированности данных будем использовать параметр k, который будет указывать долю примеров позитивного класса. Поскольку на практике идеально сбалансированных данных практически не бывает, будем считать, что, если пропорция примеров двух классов будет более чем 40/60 (k = [0.4; 0.6]), то будем такие классы считать сбалансированными. Также исследуем зависимость степени суммарной ошибки и распределение ошибки по классам.

❖ ***Анализ сбалансированных данных***

На данном этапе генерируются матрицы ошибок со сбалансированным количеством объектов каждого класса (k = 0,4). Суммарная ошибка была зафиксирована и распределялась среди двух классов. В качестве суммарной ошибки были выбраны значения E=10% и E=20%. На рис. 1 представлены значения метрик в зависимости от доли распределения ошибок в наименьшем (позитивном) классе. Таким образом значение функций от 0 показывает случай, когда все значения наименьшего позитивного класса были предсказаны верно, а в наибольшем классе E% предсказаны неверно.



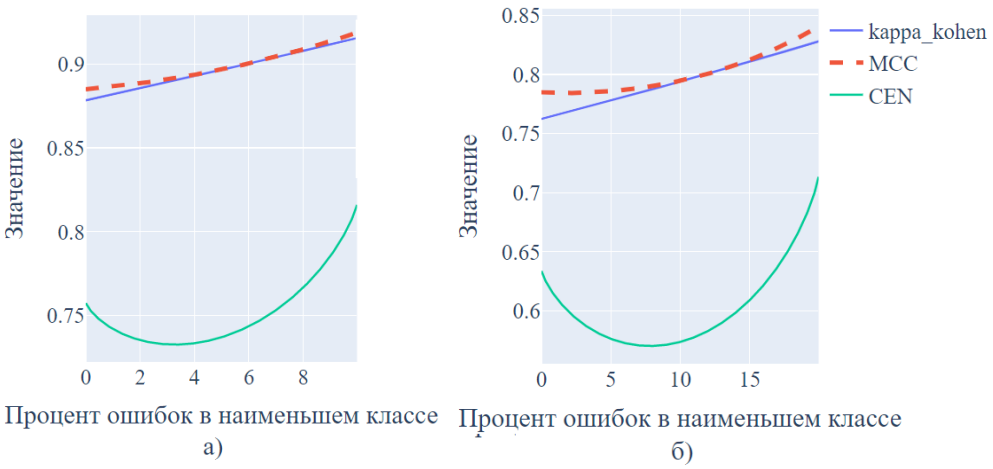


Рис. 1. Распределение ошибок по классам для сбалансированных данных. а) – суммарная ошибка 10%, б) – 20%

На представленных графиках можно увидеть, что метрики *FPR*, *FNR*, *sensitivity*, *specificity*, *AUC* и *индекс Юндена* линейно зависят от суммарного процента ошибок. *FPR*, *FNR*, *sensitivity, specificity* описывают характеристики определенного класса, а *AUC* и *Юнден* принимает константу и не зависят от распределения ошибок среди классов.

Функции *accuracy*, *precision*, *индекс Жаккара*, *F1*, *F2* и *GM* визуально линейны, однако имеют отклонения от прямых. Степень отклонения зависит от суммарной ошибки.

Каппа Кохен и *MCC* не линейны и близки по значениям при малых процентах суммарной ошибки. При ошибке в 20% *MCC* начинает загибаться вверх и на концах принимать большие значения, чем *каппа Кохена*.

Функция *CEN* при увеличении числа ошибок наименьшего класса на первом этапе возрастает, после чего убывает, что является неудобным, для сравнения полученных результатов. Аналогично функция *DP*, являющейся симметричной принимает наименьшее значение, при равном проценте ошибок в двух классах и принимает наибольшее значение в случае распределения ошибок в одном из классов. Таких показатели говорят о неудобстве использования данных метрик, для сравнения результатов модели.

❖ ***Анализ несбалансированных данных***

Да данном этапе суммарная ошибка была зафиксирована и равна 20%. Генерировались матрицы ошибок с дисбалансом k = 0.1 и k=0.01. Адекватно

оценивающая метрика должна быть неизменчива, относительно распределение ошибок по классам и оценивать суммарную ошибку.

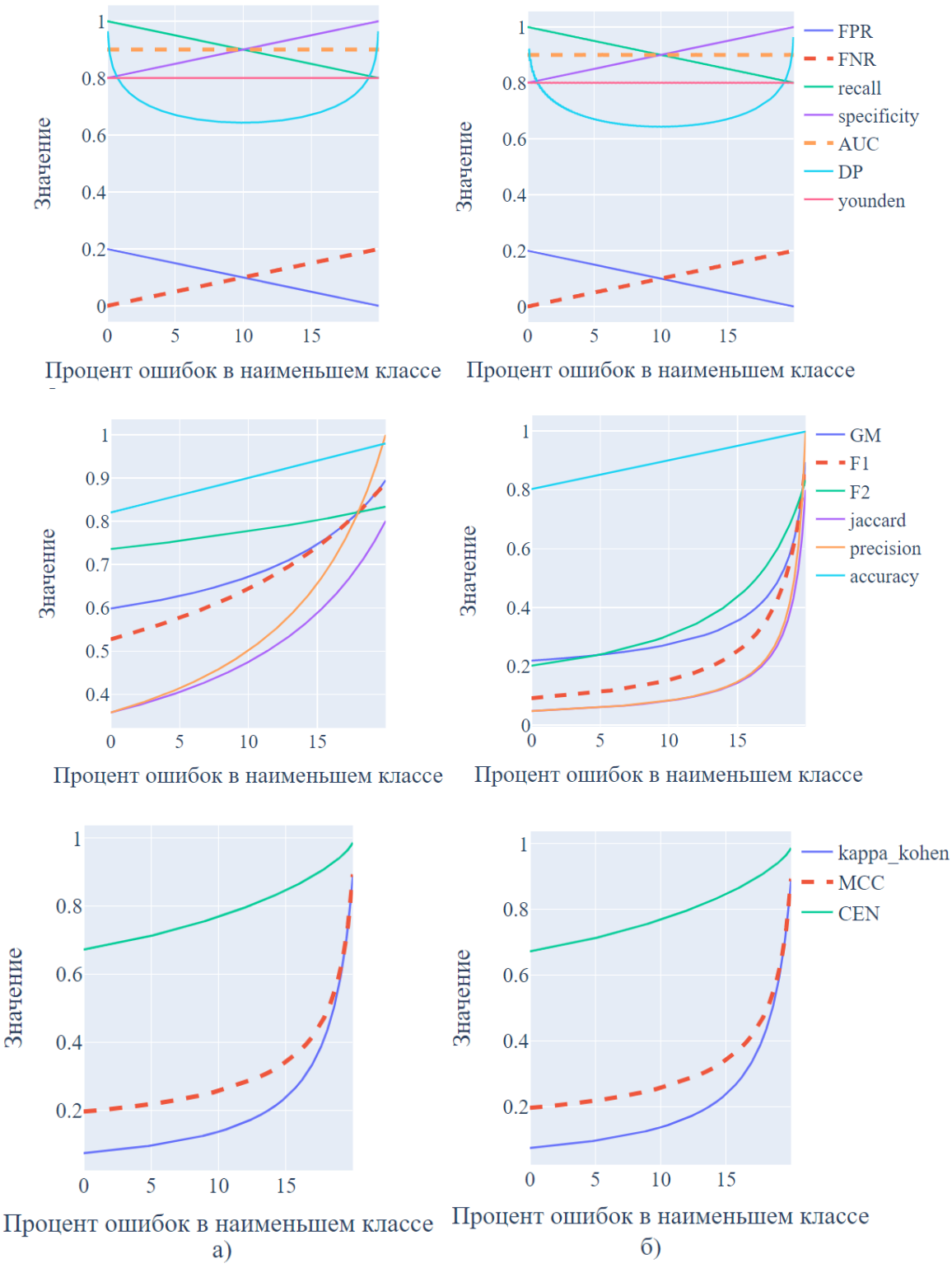


Рис. 2. Распределение ошибок по классам для несбалансированных данных. а) – дисбаланс 0.1, б) 0.01

Функции *FPR*, *FNR*, *sensitivity*, *specificity*, *индекс Юдена* и *AUC* для двух случаев дисбаланса численно равны, следовательно, не зависят от дисбаланса классов. *Индекс Юдена* и *AUC* зависят только от суммарного процента ошибок в обоих классах и не меняются при разном распределении ошибок между классами даже в случае сильного дисбаланса. Функция *D*P является комбинацией функций *sensitivity* и *specificity*, поэтому ее значения также не меняется при дисбалансе классов.

Метрики *GM*, *F1*, *F2*, *индекс Жаккара*, *precision* принимают минимальные значения при наибольшем числе ошибок доминирующего класса и приближаются к единице, если ошибки присутствуют только в наименьшем классе. Это говорит о их сильной чувствительности к дисбалансу классов и плохой применимости. Аналогично метрика *accuracy* для дисбаланса классов при 20% ошибке меньшего класса принимает значение 0,999, что говорит о ее неприменимости для данного рода задач.

Заключение

В статье были рассмотрены и описаны различные метрики, применяемые для задач бинарной классификации. Был проведен сравнительный анализ функций, применяемых для оценки задач классификаций по матрице ошибок и их взаимосвязи.

Проанализированы результаты бинарной классификации сбалансированных и несбалансированных данных. Выявлено, что для оценки сбалансированных данных нецелесообразно применять функцию *CEN*, поскольку при увеличении суммарной ошибки в некоторый момент функция начинает вести себя неадекватно, показывая лучшие результаты.

Все классические функции, такие как *sensitivity*, *specificity*, *precision*, *accuracy*, *F1*, *F2*, *GM* и *индекс Жаккара* очень чувствительны к дисбалансу классифицируемых данных и искажают оценки при большей доли ошибок объектов меньшего класса. Коэффициента корреляции *Мэтьюса* и *каппа Коэна* имеют некоторую чувствительность к дисбалансу. Наглядно показано, что такие функции, как энтропия ошибки (*CEN*), степень разделимости (*DP*), не стоит использовать для оценки бинарной классификации несбалансированных классов. *DP* не зависит от дисбаланса классов, однако неадекватно дает оценку, в случае распределения доли ошибок в один класс.

Индекс Юдена и *AUC* не зависят от дисбаланса классов, и при приведении их к одному числовому диапазону совпадают. В связи с этим *AUC* является наилучшей оценкой для анализа как сбалансированных, так и не сбалансированных данных.

Библиографический список:

1. Д.О. Макиенко, И.А. Селезнев, И.В. Сафонов. Влияние дисбаланса классов в обучающей выборке на качество классификации для задачи определения литотипов по фотографиям полноразмерного керна // Информационные технологии и нанотехнологии (ИТНТ-2020). Сборник трудов по материалам VI Международной конференции и молодежной школы (г. Самара, 26-29 мая): в 4 т. / Самар. нац.-исслед. ун-т им. С. П. Королева (Самар. ун-т), Ин-т систем. обраб. изобр. РАН-фил. ФНИЦ "Кристаллография и фотоника" РАН; [под ред. В. А. Фурсова]. – Самара: Изд-во Самар. ун-та, 2020. – Том 4. Науки о данных. – С. 574-581.
2. Düntsch and G. Gediga, “Confusion matrices and rough setdata analysis,”Journal of Physics: Conference Series, vol. 1229, 2019.
3. Choi S. S., Cha S. H., Tappert C. C. A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics, 2010, vol. 8(1), pp. 43–48.
4. He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons; 2013.
5. Sokolova M., Lapalme G. A . Systematic analysis of performance measures for classification tasks. Information Processing & Management, 2009, vol. 45, no. 4, pp. 427–437.
6. Valverde-Albacete FJ, Pela´ez-Moreno C 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explainsthe Accuracy Paradox; (2014).
7. Ye N, Chai KMA, Lee WS, Chieu HL. Optimizing F-measures: a tale of two approaches. In: Proceedings of the International Conference on Machine Learning; 2012.
8. Powers D. M. What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes, 2015. Уровень доступа: https://arxiv.org/abs/1503.06410.

9. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, vol. 27, no. 8, pp. 861–874.

10. Youden W. J. Index for rating diagnostic tests. Cancer, 1950, vol. 3, no. 1, pp. 32–35.

11. Boughorbel, S. & Jarray, Fethi & El-Anbari, Mohammed. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE. 12. e0177678. 10.1371/journal.pone.0177678.