ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

УДК 004.93
https://doi.org/10.37661/1816-0301-2020-17-1-87-101

Поступила в редакцию 08.01.2020
Received 08.01.2020

Принята к публикации 10.02.2020
Accepted 10.02.2020

Сравнительный анализ оценок качества бинарной классификации

В. В. Старовойтов✉, Ю. И. Голуб

Объединенный институт проблем информатики
Национальной академии наук Беларуси, Минск, Беларусь
✉E-mail: valerys@newman.bas-net.by

Аннотация. Приведены данные аналитического и экспериментального анализов 17 функций, используемых для оценки результатов бинарной классификации произвольных данных. Результаты классификации представлены матрицами ошибок размером 2×2. Исследованы поведение и свойства основных функций, вычисляемых по элементам этих матриц. Анализируются варианты классификации со сбалансированными и несбалансированными классами данных. Показано, что между отдельными функциями существуют линейные зависимости. Многие функции инвариантны к транспонированию матриц ошибок, что позволяет вычислять оценки, не уточняя порядок записи данных в эти матрицы.

Доказано, что все классические функции (Sensitivity, Specificity, Precision, Accuracy, F1, F2, GM, индекс Жаккара) чувствительны к дисбалансу классифицируемых данных и искажают оценки при ошибках классификации объектов меньшего класса. Чувствительность к дисбалансу имеется у коэффициента корреляции Мэтьюса и каппы Коэна. Экспериментально показано, что такие функции, как энтропия ошибки (confusion entropy), степень разделимости (discriminatory power) и диагностическое отношение шансов (diagnostic odds ratio), не стоит использовать для анализа результатов бинарной классификации несбалансированных классов. Две последние функции инвариантны к дисбалансу классифицируемых данных, но плохо оценивают результаты с примерно равным суммарным процентом ошибок классификации.

Доказано, что площадь под ROC-кривой (AUC) и индекс Юдена, вычисляемые по матрице ошибок бинарной классификации, линейно зависимы и являются наиболее подходящими оценочными функциями для сравнения результатов бинарной классификации как сбалансированных, так и несбалансированных данных.

Ключевые слова: бинарная классификация, матрица ошибок, функции точности классификации, площадь под ROC-кривой, индекс Юдена

Для цитирования. Старовойтов, В. В. Сравнительный анализ оценок качества бинарной классификации / В. В. Старовойтов, Ю. И. Голуб // Информатика. – 2020. − Т. 17, № 1. – С. 87–101. https://doi.org/10.37661/1816-0301-2020-17-1-87-101

Comparative study of quality estimation of binary classification

Valery V. Starovoitov✉, Yuliya I. Golub

The United Institute of Informatics Problems of the National Academy
of Sciences of Belarus, Minsk, Belarus
✉E-mail: valerys@newman.bas-net.by

Abstract. The paper describes results of analytical and experimental analysis of seventeen functions used for evaluation of binary classification results of arbitrary data. The results are presented by 2×2 error matrices. The behavior and properties of the main functions calculated by the elements of such matrices are studied. Classification options with balanced and imbalanced datasets are analyzed. It is shown that there are linear dependencies between some functions, many functions are invariant to the transposition of the error matrix,

which allows us to calculate the estimation without specifying the order in which their elements were written to the matrices.
It has been proven that all classical measures such as Sensitivity, Specificity, Precision, Accuracy, F1, F2, GM, the Jacquard index are sensitive to the imbalance of classified data and distort estimation of smaller class objects classification errors. Sensitivity to imbalance is found in the Matthews correlation coefficient and Kohen's kappa. It has been experimentally shown that functions such as the confusion entropy, the discriminatory power, and the diagnostic odds ratio should not be used for analysis of binary classification of imbalanced datasets. The last two functions are invariant to the imbalance of classified data, but poorly evaluate results with approximately equal common percentage of classification errors in two classes.
We proved that the area under the ROC curve (AUC) and the Yuden index calculated from the binary classification confusion matrix are linearly dependent and are the best estimation functions of both balanced and imbalanced datasets.

Keywords: binary classification, confusion matrix, functions of Accuracy classification, area under ROC curve, Youden's index

For citation. Starovoitov V. V., Golub Y. I. Comparative study of quality estimation of binary classification. *Informatics*, 2020, vol. 17, no. 1, pp. 87−101 (in Russian). https://doi.org/10.37661/ 1816-0301-2020-17-1-87-101

Введение. Следует отметить, что распознавание и классификация часто трактуются как синонимы. Например, в работе [1] задача распознавания образов формулируется как классификация заданного множества объектов. На взгляд авторов, это схожие, но отличающиеся понятия. Под классификацией понимается отнесение заданного объекта к одному или нескольким классам, определенным заранее. В данном случае термин «распознавание» можно использовать как синоним классификации. Процесс распознавания имеет самостоятельное значение, если речь идет о выявлении объекта с последующим отнесением его к классу из заданного множества. Например, при распознавании текста или дорожных знаков на изображениях, если большую часть кадра занимает один знак и требуется определить, что это за знак, – это задача классификации изображений как объектов; если же требуется найти знак на изображении и опознать его, – это задача распознавания изображений.

Задачи классификации делятся на бинарные (имеются объекты только двух классов) и многоклассовые. Классификация может быть выполнена с непересекающимися классами и с пересекающимися, когда один объект может принадлежать нескольким классам. Термин «классификация» также можно использовать при диагностике заболевания и определении стадии этого заболевания. Если количество объектов разных классов представлено значениями одного порядка, классы называются сбалансированными [2]. Если же объемы классов различаются на порядок и более, то они называются несбалансированными. Например, в банковской сфере среди огромного числа легальных транзакций по кредитным картам встречается небольшое число (на несколько порядков меньше) мошеннических. Задачи медицинской диагностики также часто содержат несбалансированные классы анализируемых данных, так как здоровых людей больше, чем больных, а больных с начальными стадиями заболевания, как правило, больше, чем с последними.

На сайте kaggle.com в конкурсе по определению мошенничества с поддельными банковскими транзакциями данные представляли собой 284 807 корректных транзакций и 492 ложные (0,172 % от всех операций). Дисбаланс классов составляет 578:1. Тривиальный (необученный) классификатор относит все операции к классу корректных, при этом он имеет высокое значение функции Accuracy (99,827 %) и низкое значение функций Precision и Recall (0,0 %), однако ни одна ложная транзакция не будет выявлена.

Часто результаты работы классификаторов оцениваются по матрицам ошибок (confusion matrix). В табл. 1 представлены объекты верно определенных классов (true) и ошибочно определенных классов (false) для одной из таких матриц.

Таблица 1

Матрица ошибок бинарной классификации

Предсказанный класс	Истинная классификация	
	Класс 1	Класс 2
Класс 1	True Positive (tp)	False Positive (fp)
Класс 2	False Negative (fn)	True Negative (tn)
Число объектов в классе	tp + fn = общее число объектов класса 1	fp + tn = общее число объектов класса 2

Функции оценки результатов бинарной классификации данных. В статье [3] приведены формулы вычисления 76 функций, а в [4] описаны 44 функции оценки результатов бинарной классификации. В обоих работах сравнительный анализ функций и рекомендации по их применению отсутствуют. В статье [5] даны формулы пяти наиболее распространенных функций оценки результатов бинарной классификации, представленных матрицей ошибок, и исследованы некоторые свойства этих функций. Опишем исследуемые в настоящей работе функции оценки результатов бинарной классификации, представленных матрицей ошибок (табл. 2).

Наиболее простой для вычислений и популярной оценкой классификаторов является функция Accuracy. Она имеет парадоксальное свойство (Accuracy Paradox): 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]. Отметим, что функция Accuracy определяет долю правильных ответов. Кратко ее название можно перевести как правильность, и не рекомендуется называть ее точностью. Точностью в переводе с английского называют функцию Precision.

Отметим, что логистическая функция ошибки LogLoss также активно используется для оценки результатов бинарной классификации, но ее невозможно вычислить по матрице ошибок. Она рассчитывается через вероятности принадлежности к заданным классам, поэтому в данной статье не рассматривается.

В литературе встречаются функции, которые являются линейно преобразованными вариантами функций, приведенных в табл. 2. Например, функция Recall (или полнота) идентична функции Sensitivity, Error = 1–Accuracy, коэффициент Gini = 2·AUC–1 [5]. Далее эти функции в настоящей работе не рассматриваются.

Функции 1–7 (табл. 2) отнесем к первой группе. Они используют два-три из четырех элементов матрицы ошибок и в одиночку не дают объективной оценки результатов классификации данных.

Вторая группа – это функции 8–11. Они популярны, просты и используют все четыре элемента матрицы ошибок. Функция F1 – это гармоническое среднее между Recall и Precision, функция GM – геометрическое среднее этих же величин, F2 – вариант функции F1, в котором значение Precision имеет больший вес. Функции F1 и F2 так же, как Accuracy, Recall и Precision, точнее оценивают результаты классификации доминирующего множества при несбалансированных данных [7]. Отметим, что индекс Жаккара и функция F1 связаны следующими нелинейными зависимостями:

$$\text{индекс Жаккара} = F1 / (2 - F1), \quad F1 = 2 \cdot \text{индекс Жаккара} / (1 + \text{индекс Жаккара}).$$

Функция 11 вычисляет площадь под ROC-кривой (ROC – receiver operating characteristic), которую обозначают AUC (area under curve). Ее значения варьируются от 0,5 до 1. Подробнее о свойствах AUC можно прочитать в статье [8]. ROC-кривая строится численно (вычислительной формулы нет) как функция fp от величины tp, значения этих двух параметров могут изменяться от 0 до 1. Кривую можно построить только в случае бинарной классификации данных, фиксируя значения одного из параметров (fp или tp) и вычисляя значение второго параметра.

Таблица 2

Функции оценки результатов бинарной классификации

Наименование	Математическое выражение
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, чувствительность	$\frac{tp}{tp + fn}$
4. Specificity, 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, гармоническое среднее	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ при $\beta = 1$
9. F2, взвешенное гармоническое среднее	$\frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$ β любое
10. Geometric mean (GM), геометрическое среднее	$\sqrt{\text{Precision} \cdot \text{Recall}}$
11. Area under ROC curve (AUC) [5], площадь под ROC-кривой	$\frac{\text{Sensitivity} + \text{Specificity}}{2}$
12. Cohen's kappa [9], каппа Коэна	$\frac{(tp + fp)(tp + fn) + (fn + tn)(fp + tn)}{n^2}$
13. MCC [10], коэффициент корреляции Мэтьюса	Формула (1)
14. Confusion entropy (CEN) [11], энтропия ошибки	Формула (2)
15. Discriminatory power (DP) [12], степень разделимости	$k\left(\log\frac{\text{Sensitivity}}{1 - \text{Sensitivity}} + \log\frac{\text{Specificity}}{1 - \text{Specificity}}\right)$
16. Youden's index [13], индекс Юдена	Sensitivity + Specificity – 1
17. Diagnostic odds ratio (DOR) [14], диагностическое отношение шансов	$\frac{tp}{fn} \Big/ \frac{fp}{tn} = \frac{\text{Sensitivity}}{1 - \text{Sensitivity}} \Big/ \frac{1 - \text{Specificity}}{\text{Specificity}}$

Примечание: $n = tp + fp + fn + tn$, константа k описана в тексте.

Можно построить альтернативную кривую в плоскости Recall/Precision и вычислить площадь под ней. В статье [15] доказана теорема, в которой утверждается, что при бинарной классификации множества объектов существует взаимно-однозначное соответствие между точками ROC-кривой в плоскости fp/tp и кривой, построенной в плоскости Recall/Precision, если эти точки определяют одинаковые матрицы ошибок и Recall $\neq 0$. Однако данный результат нельзя использовать, если итог классификации задан в виде одной матрицы ошибок. Для такой матрицы ROC-кривая состоит из двух отрезков, задаваемых тремя точками с координатами (0,0),

(fpr, tpr), (1,1). Площадь под ROC-кривой можно вычислить как сумму площадей двух треугольников (функция 11 табл. 2). Кривую в плоскости Recall & Precision по одной матрице ошибок построить нельзя, так как известна только одна средняя точка с координатами (Recall, Precision). Поэтому вторая кривая далее не рассматривается. В работе [15] рекомендуется использовать площадь под ROC-кривой для оценки результатов при классификации сбалансированных данных.

Третью группу образуют реже используемые функции, собранные в процессе анализа научной литературы: каппа Коэна (Cohen's kappa) [9], коэффициент корреляции Мэтьюса (Matthews Correlation Coefficient, MCC) [10, 16], энтропия ошибки (Confusion Entropy, CEN) [11], степень разделимости классов (Discriminatory power, DP) [12], индекс Юдена (Youden's index) [13], диагностическое отношение шансов (diagnostic odds ratio, DOR) [14]. Отметим, что в некоторых публикациях индекс Юдена называют Bookmaker Informedness, его можно вычислить через функцию AUC следующим образом:

$$\text{индекс Юдена} = 2\cdot\text{AUC} - 1.$$

Оригинальное определение функции DP дает константу $k = 0{,}5513$, но при этом значение функции может достигать нескольких десятков. Для унификации графического представления функции DP вместе с другими функциями вместо константы k использовался десятичный логарифм от выражения, записанного в скобках в функции 15 табл. 2.

Коэффициент MCC – это дискретный случай коэффициента корреляции Пирсона. Для задач бинарной классификации он вычисляется по формуле

$$\text{MCC} = \frac{(\text{tp}\cdot\text{tn} - \text{fp}\cdot\text{fn})}{\sqrt{(\text{tp} + \text{fp})\cdot(\text{tp} + \text{fn})\cdot(\text{tn} + \text{fp})\cdot(\text{tn} + \text{fn})}}. \quad (1)$$

В статье [16] показано, что для двух классов, случайно сгенерированных и несбалансированных, функции MCC и AUC достаточно устойчивы. Недостатком функции AUC является отсутствие точной формулы ее вычисления в случае многоклассовой классификации. Однако при бинарной классификации и наличии одной матрицы ошибок имеется простая формула вычисления AUC_ROC, представленная в табл. 2.

В работе [17] исследуется зависимость между MCC и CEN в случае многоклассовой классификации, а также приведены формулы вычисления CEN для случая бинарной классификации:

$$\text{CEN} = \frac{(\text{fn} + \text{fp})\cdot\log_2\left((\text{tp} + \text{tn} + \text{fp} + \text{fn})^2 - (\text{tp} - \text{tn})^2\right)}{2(\text{tp} + \text{tn} + \text{fp} + \text{fn})} - \frac{\text{fn}\cdot\log_2\text{fn} + \text{fp}\cdot\log_2\text{fp}}{(\text{tp} + \text{tn} + \text{fp} + \text{fn})}. \quad (2)$$

Если $\text{tp} = \text{tn} = T$ и $\text{fp} = \text{fn} = F$, то верно равенство

$$\text{CEN} = \frac{F}{T+F}\log_2\frac{2(T+F)}{F}.$$

Максимальное значение CEN равно 1,0615 при $T / (T + F) = 0{,}737$ (рис. 1). Это справедливо, например, для матрицы ошибок со значениями [300, 700; 700, 300]. Для такой матрицы MCC = –0,400, что означает низкое качество классификации. При соотношении параметров $T / (T + F) > 0{,}5$ значение CEN > 1. Однако при tp = tn = 0, т. е. в случае полностью неверной классификации, CEN = 1. Эти примеры свидетельствуют о недостаточно корректной оценке функцией CEN матрицы ошибок при плохой классификации данных.

В медицинской статистике известна функция, называемая диагностическим отношением шансов (diagnostic odds ratio, DOR) [14]. Значения данной функции не ограничены сверху при fp или fn, стремящихся к нулю. Поэтому DOR имеет максимальное значение при fp или fn, стремящихся к нулю. Это означает отсутствие ошибок при классификации объектов одного из двух классов, что может привести к плохому разделению пересекающихся классов. Например,

если при определении, болен ли человек, его всегда относить к классу здоровых, то функция DOR в результате такой классификации будет иметь максимальное значение. По этой причине функция DOR не рекомендуется для применения в медицинской диагностике [18]. В настоящей статье данная функция модифицирована следующим образом:

$$\mathrm{DOR}^* = \frac{\log_{10}(\log_{10}(\mathrm{DOR}))}{1{,}4}.$$

Если DOR* > 1, то DOR* = 0,999.

Здесь и далее новые значения параметров будем обозначать значком *.

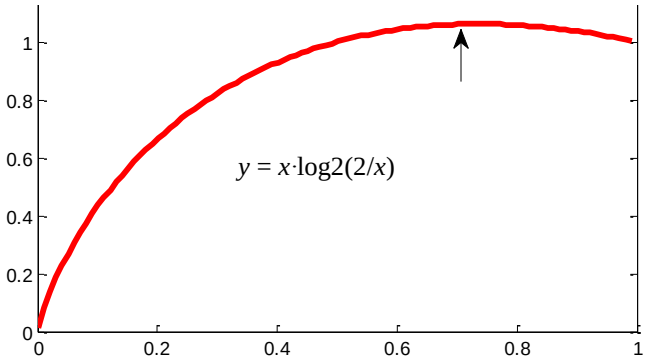


Рис. 1. График значений CEN при изменении $x = T / (T + F)$

Значения функции DOR* при суммарной ошибке классификации не более 49 % всегда находятся в диапазоне [0, 0,999], а при большем проценте ошибок никакие оценки не имеют смысла.

Экспериментальные исследования функций. Оценка результатов классификации, вычисленная по матрице ошибок, должна давать максимально объективную, сбалансированную оценку ошибок при анализе объектов как сбалансированных, так и несбалансированных классов. При решении задачи бинарной классификации такая оценка позволит выбрать объективно лучший метод, учитывающий дисбаланс классов.

Исследования выполнялись в три этапа. На *первом этапе* генерировались матрицы ошибок со сравнимыми размерами двух классов и фиксированным суммарным количеством ошибок, составляющим N % в классах 1 и 2. Например, $N_p = 5$ % ошибочно классифицированных данных класса 1 (positive) и $N_n = (N - N_p) = 95$ % ошибочно классифицированных данных класса 2 (negative). Пусть дисбаланс в размерах классов равен K, тогда размеры классов (tp + fn) = $= K \cdot$ (tn + fp). Затем вычислялись и анализировались значения функций, приведенных в табл. 2, путем изменения N_p от 0 до N.

Оценим изменения функции 1 из табл. 2 при дисбалансе K и процентном соотношении ошибок $N_n : N_p$:

$$\mathrm{FPR}^* = \mathrm{fp}^* / (\mathrm{fp}^* + \mathrm{tn}^*), \quad \mathrm{fp}^* = \mathrm{N_n}\, K \cdot (\mathrm{tn} + \mathrm{fp}) / 100,$$

$$\mathrm{FPR}^* = N_n \cdot K \cdot (\mathrm{tn} + \mathrm{fp}) / (100 \cdot K \cdot (\mathrm{tn} + \mathrm{fp}) = N_n / 100.$$

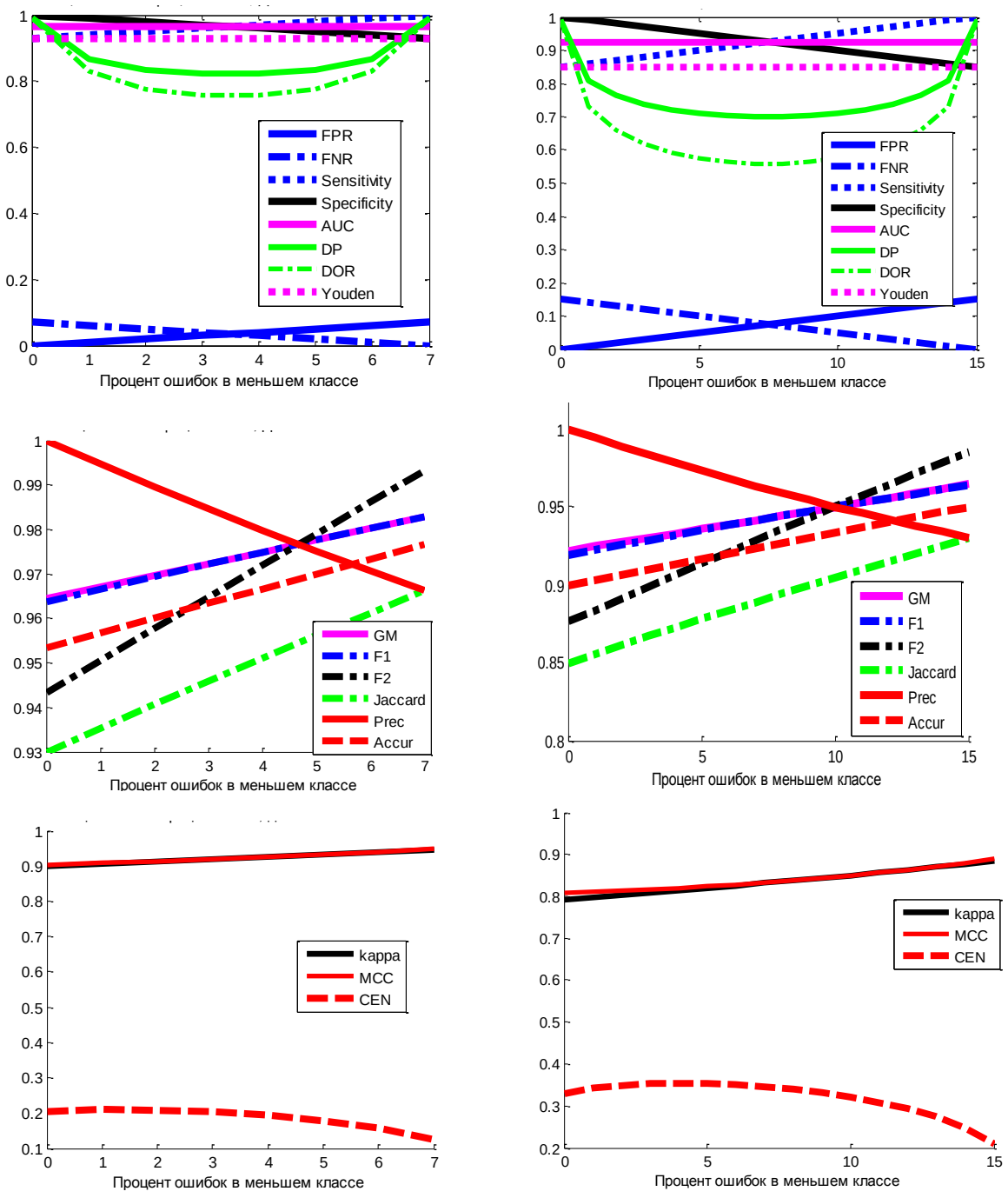
Таким образом, значение функции FPR линейно зависит от процента ошибок в классе 1 и не меняется при дисбалансе классов. Аналогично уточним значения пяти других функций из табл. 2:

$$\mathrm{FNR} = (N - N_p) / 100,$$
$$\mathrm{Sensitivity} = \mathrm{Recall} = (100 - (N - N_p)) / 100 = 1 - (N - N_p)) / 100,$$
$$\mathrm{Specificity} = (100 - N_p) / 100,$$
$$\text{индекс Юдена} = \mathrm{Sensitivity} + \mathrm{Specificity} - 1 = 1 - N,$$
$$\mathrm{AUC} = (\mathrm{Sensitivity} + \mathrm{Specificity}) / 2 = 1 - N / 200.$$

Функции 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 дисбаланса. Функции DP и DOR* являются комбинациями функций Sensitivity и Specificity, поэтому их значения не меняются при дисбалансе классов.

Рассмотрим оценки бинарной классификации с помощью функций из табл. 2 при разных степенях дисбаланса классов.

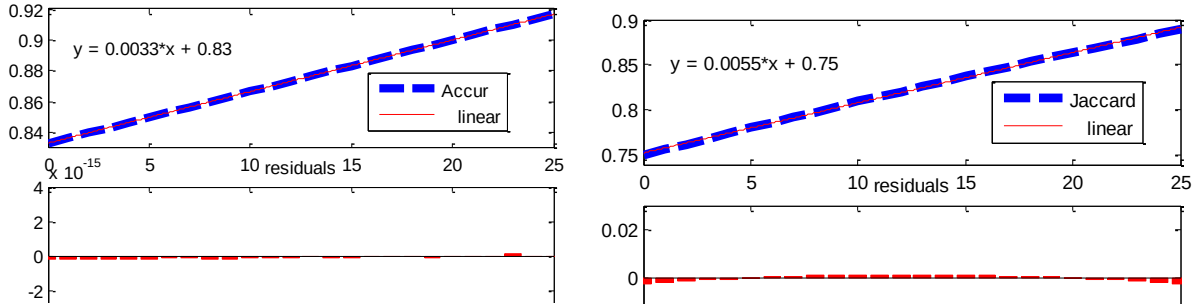
Анализ оценок в случае сбалансированных классов. В настоящей работе данные называются сбалансированными по классам, если отношение K размера большего класса к размеру меньшего находится в пределах от 1 до 5.



а) *б)*

Рис. 2. Графики оценок классификации при 7 % (*а*) и 15 % (*б*) суммарных ошибок в классах с отношением размеров классов $K = 2$

На рис. 2 видно, что шесть функций (FPR, FNR, Sensitivity, Specificity, AUC и индекс Юдена) линейны относительно суммарного процента ошибок. Другие шесть функций (Accuracy, Precision, индекс Жаккара, F1, F2 и GM) визуально почти линейны, но имеют небольшие отклонения от прямых линий (рис. 3).



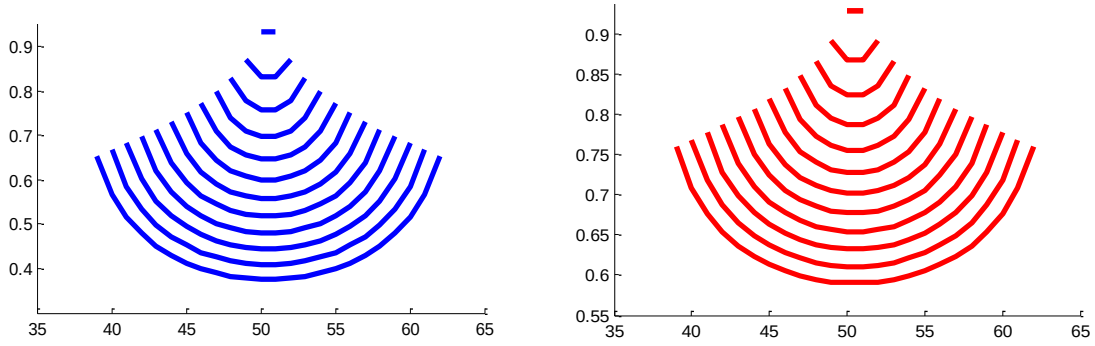
а) *б)*

Рис. 3. Графики функций Accuracy (*а*), индекс Жаккара (*б*) и их аппроксимация прямыми. Внизу представлены графики невязки линейной апроксимации

Значения функций каппа Коэна и MCC нелинейны и очень близки при суммарной величине ошибок до 10 %. При большем проценте ошибок кривая MCC больше изгибается на концах, т. е. сильнее отличается от кривой каппа Коэна при малых процентах ошибок в одном или другом классе. Значения обеих функций растут при увеличении процента ошибок классификации объектов меньшего класса.

Функция CEN при увеличении числа ошибок объектов меньшего класса сначала незначительно растет, а затем убывает (см. рис. 2). Это свойство затрудняет ее использование для сравнения результатов работы разных классификаторов между собой.

На рис. 4 изображены графики функций DOR* и DP. Они подобны, симметричны относительно равного процента ошибок в классах и имеют более высокие значения при меньшем проценте ошибок в одном из классов.



а) *б)*

Рис. 4. Оценки функций DOR* (*а*) и DP (*б*) при суммарной величине ошибок от 2 до 25 % при дисбалансе классов $K = 2$

Анализ оценок в случае несбалансированных классов. На *втором этапе* исследований генерировались матрицы ошибок с дисбалансом K от 10 до 1000 и суммарной ошибкой классификации, равной N % в обоих классах, и оценивались результаты с помощью функций из табл. 2. Графики этих функций показаны на рис. 5 при дисбалансе между классами, составляющем один и два порядка. Функции FPR, FNR, Sensitivity, Specificity, AUC, DP и индекс Юдена имеют такой же вид, как и на рис. 2. Функции DOR* и DP имеют такой же вид и такие же значения, как и на рис. 4. Шесть функций (FPR, FNR, Sensitivity, Specificity, AUC и индекс Юдена) являются линейными относительно суммарного процента ошибок независимо от дисбаланса классов. Другие шесть функций (Accuracy, Precision, индекс Жаккара, F1, F2 и GM) визуально почти линейны при любом дисбалансе, но имеют небольшие отклонения от прямых линий.

При дисбалансе классов K в один-два порядка и большом проценте ошибок многие функции имеют примерно равные значения для широкого диапазона значений процента ошибок в одном классе. Таковыми являются функции GM и F1, Accuracy и индекс Жаккара (рис. 5). Они возрастают почти до максимального значения, равного единице, при уменьшении ошибок в большем классе и увеличении ошибок в меньшем классе. Отмеченное свойство затрудняет использование этих функций для сравнения результатов работы разных классификаторов между собой при выборе лучшего из них для анализа несбалансированных данных.

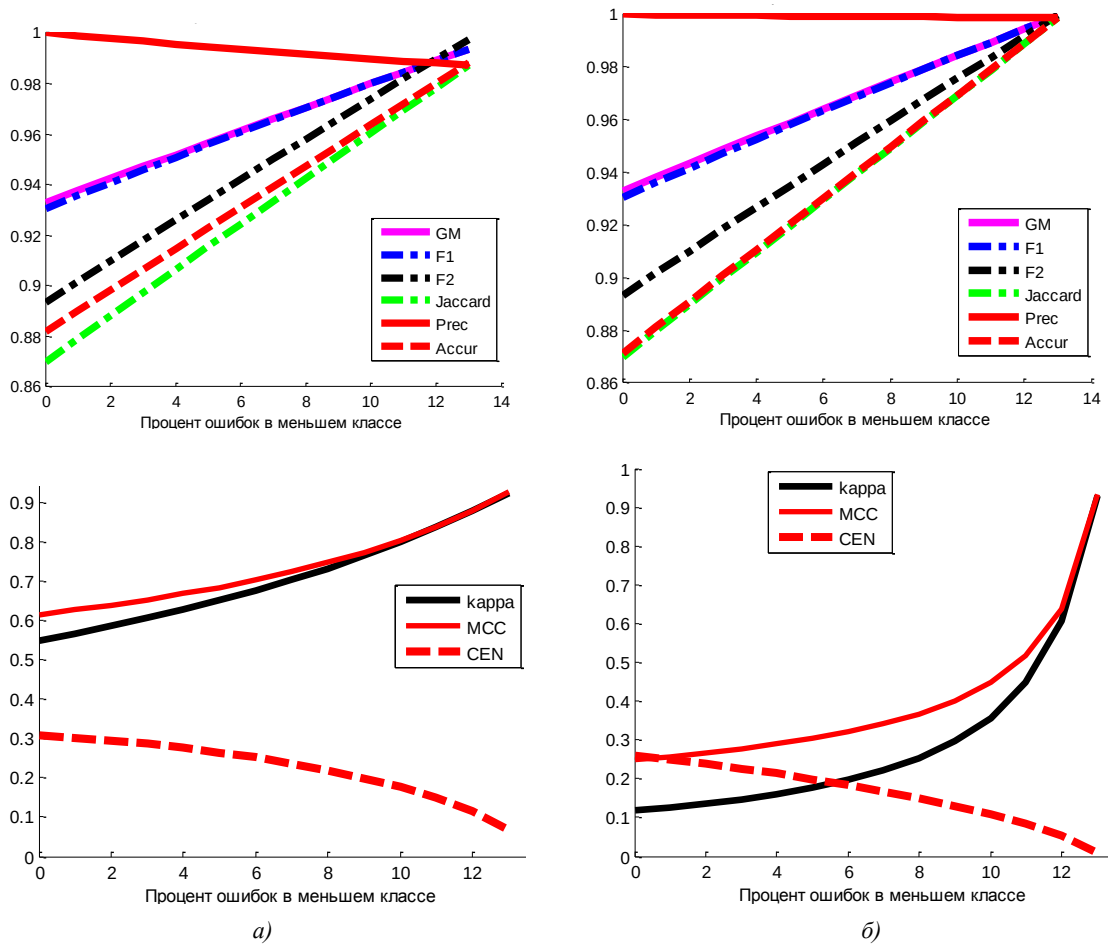


Рис. 5. Суммарная величина ошибок классификации N = 13 %, дисбаланс между классами K = 10 (*а*) и K = 100 (*б*)

При дисбалансе $K \geq 100$ значения функции Precision изменяются от 1 до 0,9977 с увеличением величины ошибок в меньшем классе от 0 до 23 %. Поэтому она не годится для оценки результатов существенно несбалансированных данных.

Функции Accuracy и индекс Жаккара практически совпадают и почти линейно зависят от процента ошибок в меньшем классе:

$$\text{Accuracy} \approx \text{индекс Жаккара} \approx (100 - N + N_1) / 100,$$

где N_1 – процент ошибок в меньшем классе (рис. 5, *б* вверху). На рис. 6 изображены уравнения линейной аппроксимации этих функций и показаны их отклонения от прямой линии. Уравнения линейной аппроксимации функций Accuracy и индекс Жаккара почти совпадают.

Выявлены следующие отношения между значениями функций: (1 – CEN) > MCC > каппа Коэна (см. рис. 5). Все три функции нелинейны, и их меньшие значения соответствуют наименьшему проценту ошибок в меньшем классе. Указанные зависимости проявляются при любом проценте суммарных ошибок и любом дисбалансе классов. При этом CEN имеет наименьший диапазон значений, а каппа Коэна – наибольший при одинаковых ошибках классификации и параметре дисбаланса K.

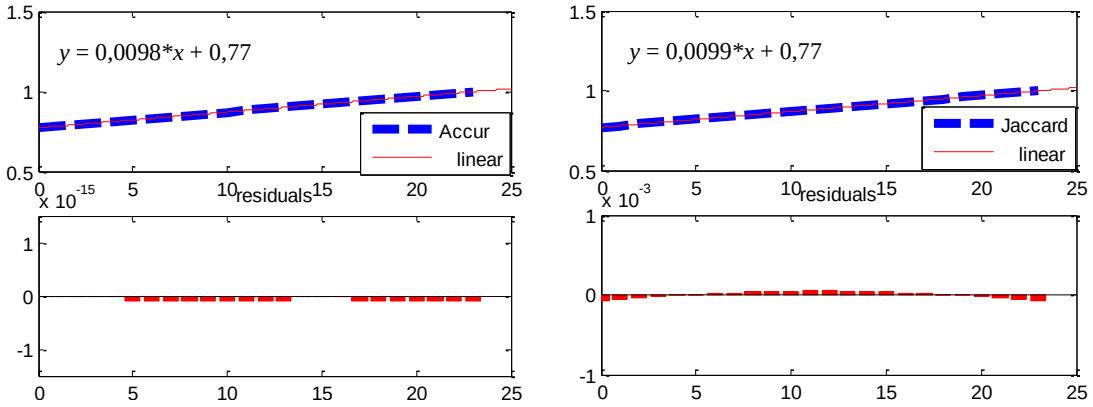


Рис. 6. Суммарная величина ошибок классификации N = 23 % при дисбалансе K = 100

Поведение рассмотренных функций инвариантно к размерам классов и меняется только при их дисбалансе.

Анализ степени разделимости функций DP и DOR. Значения функций DP и DOR для сбалансированных и несбалансированных классов полностью совпадают при одинаковом проценте ошибок N (см. рис. 4 и 7). Значения этих функций минимальны при равном проценте ошибок в обоих классах и максимальны при большем проценте в одном из них. Функции DP и DOR являются хорошими кандидатами в индикаторы ошибок бинарной классификации, инвариантными к дисбалансу классов. Они симметричны относительно равного числа ошибок в классах (рис. 7). Чем больше суммарный процент ошибочной классификации, тем ниже значения DP и DOR.

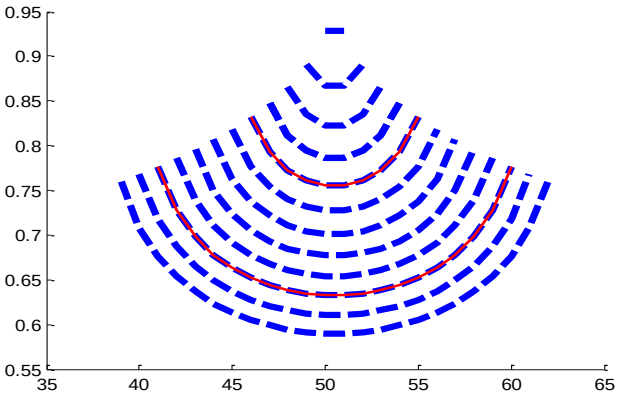


Рис. 7. Оценки DP при суммарной величине ошибок классификации N = 3–25 % при дисбалансе K = 1000 (пунктирная линия) и K = 10 (сплошная красная линия для двух случаев)

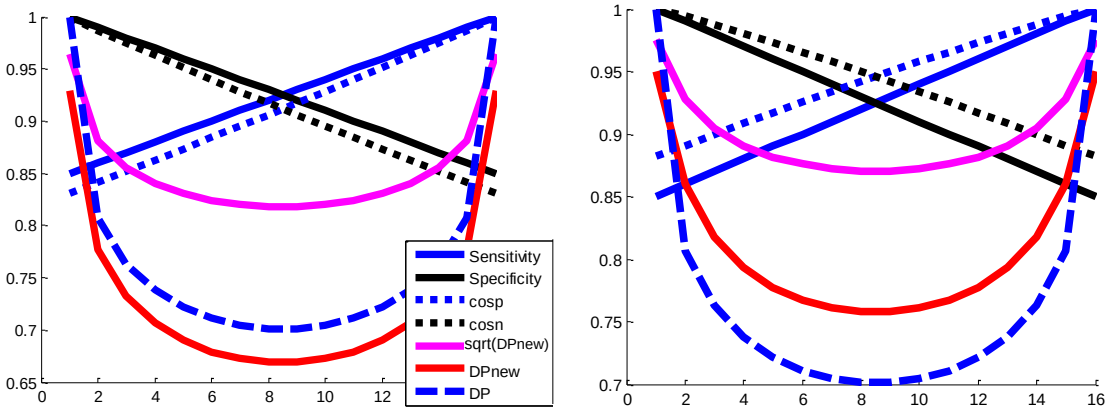
Функции DP и DOR можно скорректировать так, чтобы они принимали значения в более узком диапазоне, что точнее соответствовало бы качеству классификации объектов. На рис. 8 показаны графики модифицированной функции DPnew, в которой Sensitivity и Specificity вычисляются с помощью показателя степени $s > 1$:

$$\text{Sensitivity} = tp / (tp^s+fn^s)^{1/s}, \quad \text{Specificity} = tn / (tn^s+fp^s)^{1/s}. \tag{3}$$

Дополнительно из новых значений DP извлечен квадратный корень, чтобы приблизить минимальные значения модифицированной функции DP к минимальным значениям функций Sensitivity и Specificity. Графики cosp и cosn на рис. 8 – это обобщенные варианты функций Sensitivity и Specificity из формулы (2), которые совпадают с косинусами отношений верно и неверно (true – false) классифицированных объектов классов 1 и 2 при s = 2. Все функции, изображенные на рис. 8, инвариантны к дисбалансу классов. Однако функции Sensitivity, Specificity и их линейные обобщения cosp и cosn оценивают только ошибки в одном из двух

классов. Функции DP, DOR и их вариации оценивают суммарный процент ошибок классификации в обоих классах.

Единственной проблемой функций DP, DOR и их вариантов являются значения функций при fn и (или) fp, равные нулю. В таких случаях при вычислении возникает проблема деления на ноль, а функции принимают значения, бóльшие единицы, при попытке замены нулей константами. Формально в таких ситуациях можно приравнять значения функций DP к константе, близкой к единице.



а) *б)*

Рис. 8. Графики функций DP, DPnew и корень квадратный из DPnew при 15 % суммарных ошибок: *а)* DPnew при s = 0,95; *б)* DPnew при s = 1,1

Анализ абсолютных ошибок классификации несбалансированных данных. На *третьем этапе* исследований строились матрицы ошибок при дисбалансе классов K = 2, 10 и 100, при этом задавалось одинаковое количество ошибок в большем и меньшем классах и вычислялись оценки согласно табл. 2. Результаты экспериментов позволили сделать выводы относительно 17 исследуемых оценочных функций (табл. 3–5). В таблицах курсивом отмечены функции, которые инвариантны к транспонированию матрицы ошибок, подчеркиванием показаны максимальные значения функций, полужирным курсивом – минимальные.

Суммарная величина ошибок, представленных в табл. 3, находится в диапазоне от 1,5 до 15,0 % (в табл. 3 они выделены жирным шрифтом). Очевидно, что большему числу ошибок должны соответствовать меньшие значения оценочных функций, а для функций 1 и 2 – бóльшие значения оценочных функций. Однако функции 1–3 имеют одинаковые минимальные значения для разных пропорций ошибок в табл. 3. Они же и функции 4–6 имеют одинаковые максимальные значения для разных пропорций и количества ошибок классификации, что свидетельствует о ненадежности функций 1–6 в качестве оценок классификации данных. Эти же выводы подтверждают данные табл. 4 и 5.

Обобщим выводы по результатам экспериментов, приведенным в табл. 3–5. Функция Sensitivity при дисбалансе K = 100 имеет почти максимальные значения 0,999 (табл. 5) при ошибке 39 % в меньшем классе. В первую очередь эта функция реагирует на ошибку в большем классе (табл. 3), поэтому она не годится для оценки несбалансированных классов.

Функция Precision при дисбалансе K = 100 имеет почти одинаковые значения (табл. 5) при ошибке 39 и 1 % в меньшем классе, при этом суммарный процент ошибок составляет 39,1 и 1,39 % соответственно. Аналогична ситуация при дисбалансе K = 10 (табл. 4). Поэтому функция Precision не годится для оценки несбалансированных классов.

Функция Specificity при равном проценте ошибок реагирует только на ошибки меньшего класса. Следовательно, она тоже не годится для объективной оценки бинарной классификации данных.

Функция Accuracy в табл. 5 имеет очень высокое значение (0,9960) при 39 % ошибок в меньшем классе и дисбалансе K = 100. В табл. 4 для матрицы ошибок [961, 39; 1, 99] Accuracy имеет значение 0,9636 – почти столько же, как и в табл. 3 для матрицы [190 10; 1 99], где значение Accuracy равно 0,9633. В первом случае дисбаланс K = 10, ошибка в большем классе равна 3,9 %, а в меньшем – 1 %. Во втором случае дисбаланс K = 2, ошибка в большем классе

равна 5 %, в меньшем – 1 %. В табл. 3 для матрицы [199, 1; 10, 90] при $K = 2$ ошибка в большем классе равна 0,5 %, а в меньшем 10 % при таком же значении Accuracy (0,9633). Эта функция учитывает только суммарное количество правильно классифицированных объектов относительно общего числа объектов. Она ни в коем случае не годится для оценки несбалансированных классов.

Таблица 3

Значения оценочных функций при дисбалансе классов $K = 2$

Матрица ошибок	[199, 1; 1, 99]	[198, 2; 1, 99]	[199, 1; 2, 98]	[190, 10; 1, 99]	[190, 10; 10, 90]	[199, 1; 10, 90]	[195, 5; 5, 95]
Суммарная величина ошибок, %	**1,5**	2,0	2,5	6,0	**15,0**	10,5	7,5
1. FPR	***0,0100***	***0,0100***	0,0200	***0,0100***	0,1000	0,1000	0,0500
2. FNR	***0,0050***	0,0100	***0,0050***	0,0500	0,0500	***0,0050***	0,0250
3. Sensitivity	0,9950	0,9900	0,9950	***0,9500***	***0,9500***	0,9950	0,9750
4. Specificity	0,9900	0,9900	0,9800	0,9900	***0,9000***	0,9000	0,9500
5. Precision	0,9950	0,9950	0,9900	0,9948	***0,9500***	0,9522	0,9750
6. *Accuracy*	0,9933	0,9900	0,9900	0,9633	***0,9333***	0,9633	0,9667
7. *Индекс Жаккара*	0,9900	0,9851	0,9851	0,9453	***0,9048***	0,9476	0,9512
8. *F1*	0,9950	0,9925	0,9925	0,9719	***0,9500***	0,9731	0,9750
9. F2	0,9950	0,9910	0,9940	0,9586	***0,9500***	0,9861	0,9750
10. *GM*	0,9950	0,9925	0,9925	0,9721	***0,9500***	0,9733	0,9750
11. AUC	0,9925	0,9900	0,9875	0,9700	***0,9250***	0,9475	0,9625
12. Каппа Коэна	0,9850	0,9776	0,9774	0,9193	***0,8500***	0,9156	0,9250
13. CEN	0,9457	0,9252	0,9252	0,8116	***0,6785***	0,8128	0,8059
14. *MCC*	0,9850	0,9776	0,9775	0,9213	***0,8500***	0,9178	0,9250
15. *DP*	0,9951	0,9633	0,9631	0,8773	***0,7111***	0,8745	0,8201
16. Индекс Юдена	0,9850	0,9800	0,9750	0,9400	***0,8500***	0,8950	0,9250
17. *DOR*	0,9990	0,9886	0,9883	0,8472	***0,5738***	0,8426	0,7530

Таблица 4

Значения оценочных функций при дисбалансе классов K =10

Матрица ошибок	[999, 1; 1, 99]	[998 2 1, 99]	[999, 1; 1, 98]	[990, 10; 1, 99]	[990, 10; 10, 90]	[999, 1; 10, 90]	[995, 5; 5, 95]
Суммарная величина ошибок, %	**1,1**	1,2	2,1	2,0	**11,0**	10,1	5,5
1. FPR	***0,0100***	***0,0100***	0,0200	***0,0100***	0,1000	0,1000	0,0500
2. FNR	***0,0010***	0,0020	***0,0010***	0,0100	0,0100	***0,0010***	0,0050
3. Sensitivity	0,9990	0,9980	0,9990	***0,9900***	***0,9900***	0,9990	0,9950
4. Specificity	0,9900	0,9900	0,9800	0,9900	***0,9000***	***0,9000***	0,9500
5. Precision	0,9990	0,9990	0,9980	0,9990	***0,9900***	0,9901	0,9950
6. Accuracy	0,9982	0,9973	0,9973	0,9900	***0,9818***	0,9900	0,9909
7. Индекс Жаккара	0,9980	0,9970	0,9970	0,9890	***0,9802***	0,9891	0,9900
8. F1	0,9990	0,9985	0,9985	0,9945	***0,9900***	0,9945	0,9950
9. F2	0,9990	0,9982	0,9988	0,9918	***0,9900***	0,9972	0,9950
10. GM	0,9990	0,9985	0,9985	0,9945	***0,9900***	0,9945	0,9950
11. AUC	0,9945	0,9940	0,9895	0,9900	***0,9450***	0,9495	0,9725
12. Каппа Коэна	0,9890	0,9836	0,9834	0,9419	***0,8900***	0,9369	0,9450
13. CEN	0,9831	0,9764	0,9765	0,9369	***0,8912***	0,9375	0,9365
14. MCC	0,9890	0,9836	0,9834	0,9429	***0,8900***	0,9382	0,9450
15. DP	0,9990	0,9990	0,9990	0,9633	***0,8320***	0,9592	0,9158
16. Индекс Юдена	0,9890	0,9880	0,9790	0,9800	***0,8900***	0,8990	0,9450
17. DOR	0,9990	0,9990	0,9990	0,9886	***0,7727***	0,9819	0,9105

Таблица 5

Значения оценочных функций при дисбалансе классов $K = 10$ и $K = 100$

Матрица ошибок	[M1, 0; 0, M2] M1&M2>0	[961, 39; 1, 99], K =10	[999, 1; 39, 61], K =10	[9999, 1; 39, 61], K =100	[9990, 10; 39, 61], K =100	[9961, 39; 1, 99], K =100
Ошибка в большем классе, %	0	3,9	0,1	0,01	0,1	0,39
Ошибка в меньшем классе, %	0	1	***39***	***39***	***39***	1
Суммарная величина ошибок, %	0	4,9	***39,1***	39,01	***39,1***	1,39
1. FPR	0	***0,0100***	0,3900	0,3900	0,3900	***0,0100***
2. FNR	0	0,0390	0,0010	***0,0001***	0,0010	0,0039
3. Sensitivity	1	***0,9610***	0,9990	0,9999	0,9990	***0,9961***
4. Specificity	1	0,9900	***0,6100***	***0,6100***	***0,6100***	0,9900
5. Precision	1	0,9990	***0,9624***	0,9961	***0,9961***	0,9999
6. Accuracy	1	***0,9636***	***0,9636***	0,9960	0,9951	0,9960
7. Индекс Жаккара	1	***0,9600***	0,9615	0,9960	0,9951	0,9960
8. F1	1	***0,9796***	0,9804	0,9980	0,9976	0,9980
9. F2	1	***0,9684***	0,9915	0,9991	0,9984	0,9969
10. GM	1	***0,9798***	0,9805	0,9980	0,9976	***0,9980***
11. AUC	1	0,9755	***0,8045***	0,8050	***0,8045***	0,9930
12. Каппа Коэна	1	0,8121	0,7346	0,7512	***0,7111***	0,8300
13. CEN	1	***0,8450***	0,8541	0,9776	0,9710	0,9765
14. MCC	1	0,8254	0,7591	0,7731	***0,7217***	0,8410
15. DP	>1	0,8921	***0,8665***	0,9849	***0,8665***	1,0060
16. Индекс Юдена	1	0,9510	***0,6090***	0,6099	***0,6090***	0,9861
17. DOR	>1	0,8715	***0,8294***	1,0241	***0,8294***	1,0588

Функции F1, F2, GM, индекс Жаккара и CEN неадекватно реагируют на ошибки при дисбалансе классов (см. табл. 4 и 5).

Функции DP и DOR в табл. 5 при ошибке 0,01 % в большем классе и 39 % в меньшем (суммарная величина ошибок 39,01 %) равны 0,9849 и 1,0241, а при ошибке 0,39 % в большем классе и 1 % в меньшем (суммарная величина ошибок 1,39 %) – 1,0060 и 1,0588 соответственно, т. е. их значения очень близки. В то же время функции DP и DOR в табл. 5 имеют более низкие оценки при ошибке 0,1 % в большем классе и при ошибке 39 % в меньшем классе (суммарная величина ошибок 39,1 %). Значения функций DP и DOR равны 0,8665 и 0,8294 соответственно. Эти функции инвариантны к дисбалансу классов относительно суммарного процента ошибок, однако в силу существенной нелинейности в вычислении они плохо оценивают результаты классификации при близких значениях суммарных процентов ошибок в классах.

Функции AUC, каппа Коэна, MCC и индекс Юдена корректнее других оценивают результаты классификации сбалансированных данных (см. табл. 3 и 4). Однако результаты классификации несбалансированных данных, приведенные в табл. 5, показывают, что каппа Коэна и MCC более чувствительны к дисбалансу классов, чем AUC и индекс Юдена. Две последние функции линейно зависимы друг от друга согласно формуле индекс Юдена = 2·AUC – 1 и дают наиболее корректные оценки результатов классификации (см. табл. 5).

Заключение. В работе выполнен сравнительный анализ 17 функций, применяемых для оценки бинарных классификаторов по матрице ошибок. Показано, что между отдельными оценочными функциями существуют простые линейные зависимости: FPR = 1 – Specificity, AUC = (Sensitivity + Specificity) / 2, индекс Юдена = Sensitivity + Specificity – 1 или 2·AUC – 1, индекс Жаккара = $F1 / (2 - F1)$.

Многие из рассмотренных функций, например Accuracy, F1, GM, AUC, каппа Коэна, MCC, CEN, DP, DOR, инвариантны к транспонированию матрицы ошибок. Это свойство позволяет вычислять их, не уточняя, как записаны данные в матрице ошибок.

Проанализированы результаты бинарной классификации сбалансированных и несбалансированных данных. Показано, что для оценки результатов классификации сбалансированных классов (при соотношении размеров не более пяти) не стоит использовать функцию энтропия ошибки, обозначенную как CEN [11]. В работе [14] также не рекомендуют применять функцию CEN для оценки результатов бинарной классификации.

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 ошибках классификации объектов меньшего класса. Чувствительность к дисбалансу имеется у коэффициента корреляции Мэтьюса и каппа Коэна. Экспериментально показано, что такие функции, как энтропия ошибки (CEN), степень разделимости (DP) и диагностическое отношение шансов (DOR), не стоит использовать для оценки бинарной классификации несбалансированных классов. Две последние функции абсолютно инвариантны к дисбалансу классифицируемых данных, но плохо оценивают варианты с примерно равным суммарным процентом ошибок классификации.

Индекс Юдена имеет диапазон значений [–1, +1], а функция AUC [0, +1]. При приведении диапазона значений индекса Юдена к [0, +1] он совпадает с функцией AUC. В статье [19] функция, использованная для бинарной классификации и совпадающая с AUC, названа сбалансированной правильностью (balanced Accuracy). Таким образом, AUC – это наиболее подходящая из известных оценочная функция для сравнения результатов классификации по матрице ошибок как сбалансированных, так и несбалансированных данных.

References

1. Zhuravlev Y. I. On the algebraic approach to solving problems of recognition and classification. *Problems of cybernetics*, Moscow, Nauka, 1978, vol. 33, pp. 5–68.

2. Haixiang G., Shang J., Mingyun G., Yuanyue H., Bing G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 2017, vol. 73, pp. 220–239.

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. Canbek G., Sagiroglu S., Temizel T. T., Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. *International Conference on Computer Science and Engineering, Antalya, Turkey, 5–8 October 2017*. Antalya, 2017, pp. 821–826.

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 F. J., Peláez-Moreno C. 100 % classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PLoS One*, 2014, vol. 9(1), 10 p. https://doi.org/10.1371/journal.pone.0084217

7. Powers D. M. *What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes*, 2015. Available at: https://arxiv.org/abs/1503.06410 (accessed 17.11.2019).

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

9. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, vol. 20, no. 1, pp. 37–46.

10. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta – Protein Structure*, 1975, vol. 405, no. 2, pp. 442–451.

11. Wei J. M., Yuan X. J., Hu Q. H., Wang S. Q. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 2010, vol. 37, no. 5, pp. 3799–3809.

12. Blakeley D. D., Oddone E. Z., Hasselblad V., Simel D. L., Matchar D. B. Noninvasive carotid artery testing: a meta-analytic review. *Annals of Internal Medicine*, 1995, vol. 122, no. 5, pp. 360–367.

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

14. Glas A. S., Lijmer J. G., Prins M. H., Bonsel G. J., Bossuyt P. M. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 2003, vol. 56, no. 11, pp. 1129–1135.

15. Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning, 25–29 June 2006, Pittsburgh, Pennsylvania, USA*. Pittsburgh, 2006, pp. 233–240.

16. Boughorbel S., Jarray F., El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS One*, 2017, vol. 12(6). https://doi.org/10.1371/journal.pone.0177678

17. 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, e41882. https://doi.org/10.1371/journal.pone.0041882

18. Pepe M. S., Janes H., Longton G., Leisenring W., Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 2004, vol. 159, no. 9, pp. 882–890.

19. Mower J. P. PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics*, 2005, vol. 6, art. 96, pp. 1–15. https://doi.org/10.1186/1471-2105-6-96

Информация об авторах

Старовойтов Валерий Васильевич – доктор технических наук, профессор, главный научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: valerys@newman.bas-net.by

Голуб Юлия Игоревна – кандидат технических наук, доцент, старший научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.

Information about the authors

Valery V. Starovoitov, Dr. Sci. (Eng.), Professor, Chief Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: valerys@newman.bas-net.by

Yuliya I. Golub, Cand. Sci. (Eng.), Associate Professor, Senior Researcher, the United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.