Инженерный вестник Дона, №11 (2024)
ivdon.ru/ru/magazine/archive/n11y2024/ 9619

Соответствие старых и новых оценок бинарной классификации событий и качества статистических критериев обнаружения

С.С. Звежинский, О.И. Шелухин

Московский технический университет связи и информатики, г. Москва

Аннотация: Рассмотрена связь между «старыми» и «новыми» понятиями / метриками оценки качества статистических критериев обнаружения и бинарной классификации событий. Приведены оценки независимости и состоятельности анализируемых метрик относительно объема/состава исходных входных данных. Уточнены рекомендации по использованию «новых» метрик оценки качества статистических критериев обнаружения и бинарной классификации событий.
Ключевые слова: ошибки 1-го и 2-го рода, достоверность, полнота, специфичность, F-мера, ROC-кривая, интегральная метрика AUC.

Введение

В статистической радиотехнике, математической статистике, радиолокации и других науках давно используется механизм бинарной классификации событий (что равносильно обнаружению) или критерии оценки статистических гипотез при действии на вход решающего устройства-классификатора двух случайных альтернативных воздействий (величин, причин), обычно связанных с присутствием/отсутствием некой угрозы или события нужного класса [1,2]. Это может быть полезный сигнал/шум, цель/фон, аномалия (атака)/трафик и пр. Выходом классификатора является два альтернативных логических решения (условно «1»/«0»), а оценками правильности его работы - два параметра [3,4]:

- ошибка 1-го рода или ложная тревога (как в радиолокации), или ошибочная классификация не угрожающего (неистинного) события как угрожающего (истинного);

- ошибка 2-го рода или пропуск цели (как в охранной сигнализации), или ошибочная классификация истинного (угрожающего) события как не угрожающего (ложного).

При относительно большой статистике входных воздействий принято говорить о вероятности ошибки 1-го рода $P_{ош1} = \alpha$ и вероятности ошибки 2-го рода $P_{ош2} = \beta$, которые в совокупности определяют эффективность решающего устройства - бинарного классификатора. В литературе используются также понятия уровень значимости α и мощность критерия $1-\beta = P_0$, где P_0 - вероятность (правильного) обнаружения (классификации) события; в зарубежной литературе для α и β приняты обозначения - соответственно *FAR* (false acceptance rate) и *FRR* (false rejection rate). Величины α и β в рамках одного классификатора являются взаимозависимыми (стремление к уменьшению одной величины приводит к увеличению другой), однако они являются независимыми относительно общего объема (числа реализаций) и состава входных данных V:

$$V = P + N, \tag{1}$$

при единственном условии их достаточно большого числа:

$$V \gg 1, P \gg 1, N \gg 1, \tag{2}$$

где P – объем входных данных для оценки качества классификатора на выдачу истинно положительных решений для определения величины β; N – объем входных данных для тестирования классификатора на истинно отрицательные решения для определения величины α.

Независимость оценок α и β от объема и соотношения данных P/N указывает на их состоятельность. При увеличении объема выборки они по определению стремятся (по вероятности) к истинным значениям:

$$\mathrm{P}_{\text{ош1}} = \alpha = \lim_{N\to\infty}\left(\frac{n}{N}\right) \approx \frac{n}{N}, N \gg 1, \tag{3}$$

$$\mathrm{P}_{\text{ош2}} = \beta = \lim_{N\to\infty}\left(\frac{p}{P}\right) \approx \frac{p}{P}, P \gg 1, \tag{4}$$

где n – число ложных тревог обнаружителя (классификатора) при его тестировании на объеме данных N; p – число ложных тревог обнаружителя (классификатора) при его тестировании на объеме данных P.

В связи с широким распространением методов интеллектуального анализа данных и машинного обучения в разных областях науки и техники для оценки качества бинарных критериев распознавания (обнаружения, классификации) событий, наряду с указанными параметрами α и β, стали использоваться «новые» термины и метрики: «accuracy» (достоверность), «recall» (полнота), «specificity» (специфичность) и др. [5,6]. Очевидно, это связано с желанием унифицировать введенные метрики для широких областей: медицины, экологии, биологии, экономики, психологии, Интернет-технологий и др. [7,8]. Однако если за понятиями ошибок 1-го и 2-го рода стоял вполне определенный физический смысл (например, обнаружение/пропуск ракеты, ложная тревога о нападении) с соответствующей интерпретацией, то за новыми терминами-метриками часто такого смысла не прослеживается.

Обычно имеющаяся база данных (БД) характеризующаяся объемом $V = P+N$, позволяет получить числовые результаты работы алгоритмов машинного обучения, например, в задачах классификации или кластеризации. Эти результаты обрабатываются (подставляются в соответствующие формулы для метрик), и сравниваются либо с заданными значениями, либо с результатами другого классификатора. Затем может делаться логический вывод типа «лучше/хуже» (по этому параметру) без всякой физической интерпретации.

В связи с вышеизложенным целью работы является проследить связь между «старыми» и «новыми» понятиями/метриками оценки качества статистических критериев обнаружения и бинарной классификации событий, оценить насколько новые метрики являются независимыми и состоятельными относительно объема/состава исходных входных данных и уточнить рекомендации по использованию «новых» метрик.

Структура бинарного классификатора

На рис.1 показана структурная схема бинарного классификатора на входы которого поступают «смешанные» данные в объеме (количестве) V. На 1-й вход поступают данные объемом P, связанные с наличием только истинных событий, подлежащих бинарной классификации (или обнаружению), такими как полезные сигналы, истинные аномалии, вторжения, вирусы и пр. На 2-й вход поступают данные объемом N, связанные с отсутствием истинных событий, подлежащих классификации, а присутствием только ложных событий, таких, как шум, обычный сетевой трафик, невирусная Интернет-среда и пр.

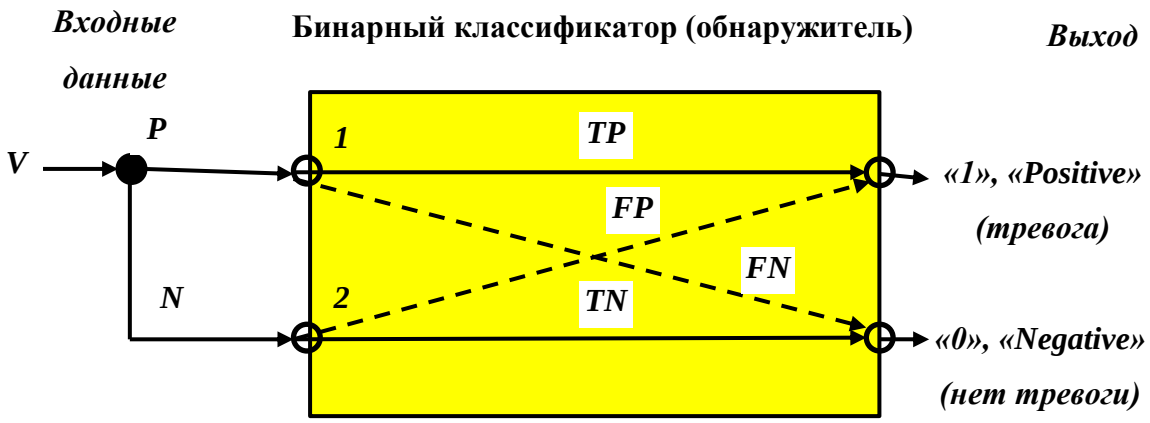

Рис. 1. – Структурная схема бинарного классификатора (обнаружителя)

При бинарной классификации каждому событию (или кванту входных данных), поступающих на входы классификатора, ставится в однозначное соответствие один их 2-х выходов, которые условно обозначены как (рис.1):

– «1» или «положительный» (англ. *Positive*), или «тревога», что соответствует принятию решения о классификации (обнаружении) события *как истинного* (хотя событие может быть и ложным);

– «0» или «отрицательный» (англ. *Negative*), или «нет тревоги», что соответствует принятию решения о классификации события *как ложного* или об отсутствии обнаружения (хотя событие может быть и истинным).

Двум входам (условно *P, N*) и двум выходам (условно «1» и «0») соответствуют четыре возможных результата (перехода) бинарного классификатора, два из которых являются истинными (*true, T*), а два - ложными (*false, F*), как показано на рис.1. При подаче на 1-й вход классификатора данных объемом (числом) *P*, соответствующих истинным событиям, подлежащим классификации (обнаружению), количественный результат работы складывается из двух решений:

$$P = TP + FN, \tag{5}$$

где: *TP* (*true positive*) - число истинно-положительных решений (переход показан сплошной линией); *FN* (*false negative*) - число ложно-отрицательных решений классификатора (переход показан штриховой линией).

При подаче на 2-й вход данных числом *N*, соответствующих ложным событиям, не подлежащим классификации (обнаружению), количественный результат работы классификатора будет:

$$N = TN + FP, \tag{6}$$

где: *TN* (*true negative*) - число истинно-отрицательных решений; *FP* (*false positive*) – число ложно-положительных решений.

Введем в рассмотрение 5 основных «новых» точечных метрик оценки эффективности бинарного классификатора [7,8]:

$$\textit{«recall»}\ (\text{полнота}) \equiv \textit{«sensitivity»}\ (\text{чувствительность}) = \frac{TP}{P}, \tag{7}$$

$$\textit{«specificity»}\ (\text{специфичность}) = \frac{TN}{N}, \tag{8}$$

$$\textit{«precision»}\ (\text{точность}) = \frac{TP}{TP+FP}, \tag{9}$$

$$\textit{«accuracy»}\ (\text{достоверность}) = \frac{TP+TN}{V}, \tag{10}$$

$$\textit{«F-score»}\ (\text{F-мера}) = 2\frac{"precision"\cdot"recall"}{"precision"+"recall"}. \tag{11}$$

Часто в рассмотрение вводится и дополнительные метрики: графическая *ROC*-кривая и интегральная *AUC* [9-11]. *ROC*-кривая (англ.

receiver operating characteristic), также известная как кривая ошибок, - график, позволяющий наглядно оценить качество бинарной классификации. Метрика *AUC* (англ. *area under curve*) – это площадь под *ROC*-кривой, ее количественная интерпретация. Чем выше этот показатель, тем качественнее классификатор; при этом значение *AUC* = 0,5 характеризует абсолютную непригодность выбранного алгоритма, соответствуя случайному гаданию [1,8].

Связь «новых» метрик и «старых» параметров

Проведем анализ введенных в рассмотрение выражений (3)-(4), уравнений связи (5)–(6), метрик (7)-(11) с учетом достаточно «большого» объема БД. Заметим, что выражения (7)-(11) для «новых» метрик справедливы, в принципе, для любого объема данных $V = P+N$. При численной оценке метрик классификатора по этим формулам требуется соблюдать метрологические правила записи выходного результата, чтобы его точность соответствовала объему данных *V*. Например, при объеме $V = 100$ недопустимо представлять результат вычисления до 3-й значащей цифры, соответствующей объему более 1000.

В тоже время случайные величины α и β («старые» показатели - соответственно вероятности ошибок классификатора 1-го и 2-го рода) стремятся к своим истинным значениям при $V \to \infty$, или, по крайней мере, при $V >> 1$ по (2). В этом заключается их принципиальное различие. При оценке ошибок 1-го и 2-го рода, как правило, вводится доверительный интервал, в котором с заданной доверительной вероятностью (типично ряд 0,8; 0,9; 0,95) лежат истинные значения α и β.

Если сравнить формулы (3) и (8) то они, по сути, отражают одно и то же, при очевидном равенстве $n = FP$. С учетом (6) получим:

$$«specificity» = \frac{TN}{N} = \frac{N-FP}{N} = 1 - \frac{n}{N} \approx 1 - \alpha \mid_{N>>1} = 1 - P_л, \quad (12)$$

где $P_л$ – вероятность обнаружения неистинного события на входе классификатора (с выдачей сигнала «1» - *ложная тревога*, рис.1).

Если сравнить формулы (4) и (7) то они, по сути, отражают одно и то же, при очевидном равенстве $p = FN$. С учетом (5) получим:

$$«recall» = \frac{TP}{P} = \frac{P-FN}{P} = 1 - \frac{p}{P} \approx 1 - \beta \mid_{P>>1} = P_0, \quad (13)$$

где P_0 – вероятность обнаружения истинного события на входе классификатора (с выдачей сигнала «1» - *правильная тревога*, рис.1).

Таким образом, при условии (2) «большого» объема БД, метрики *«specificity»* и *«recall»* практически тождественны параметрам соответственно *(1–α)* и *(1–β)*, являются состоятельными, т.е. не зависят от изменений в наборе данных *P/N*. Они имеют ясный физический смысл.

Подставив в (9)-(11) выражения (3)-(6), с учетом $n = FP$ и $p = FN$, нетрудно получить:

$$«precision» = \frac{TP}{TP+FP} = \frac{1}{1+\frac{FP \cdot P \cdot N}{TP \cdot N \cdot P}} = \frac{1}{1+\frac{N}{P}\cdot\frac{n}{N}\cdot\frac{1}{\frac{TP}{P}}} \cong \frac{1}{1+\frac{N}{P}\cdot\frac{\alpha}{1-\beta}}, \quad (14)$$

$$«accuracy» = \frac{TP+TN}{P+N} = \frac{P-FN+N-FP}{P+N} = 1 - \frac{FN}{P\left(1+\frac{N}{P}\right)} - \frac{FP}{N\left(1+\frac{P}{N}\right)} \cong 1 - \frac{\beta}{1+\frac{N}{P}} - \frac{\alpha}{1+\frac{P}{N}}, \quad (15)$$

$$«F\text{-}score» = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \cong \frac{1}{1+\frac{\beta+\alpha\cdot\frac{N}{P}}{2(1-\beta)}}. \quad (16)$$

Получившиеся выражения (14)–(16) для соответствующих метрик – довольно сложные, особенно для *F*-меры, которая по-другому называется «гармоническим средним между точностью и полнотой»; они физического смысла не имеют. Кроме того, данные метрики не являются состоятельными и зависят от соотношения *N/P*. Это значит, что оценивать качество работы классификаторов (по метрикам «точность», «достоверность» и «*F*-мера») по базам с разными распределениями входных данных нежелательно. Нетрудно убедиться, что с увеличением соотношения *N/P* величины метрик «точность» и «*F*-мера» уменьшаются; для метрики «достоверность» такого однозначного

вывода сделать нельзя, все зависит от соотношения параметров *α/β*.

Чтобы результаты оценки качества классификатора по метрикам «точность», «достоверность» и «*F*-мера» были инвариантны относительно распределения данных в БД, нужно зафиксировать величину *N/P* на некотором уровне, и прописать это, например, в руководстве для подразделения, которая занимается разработкой классификаторов. Например, при *N/P* = 1 формулы (14)-(16) упрощаются и имеют вид:

$$\begin{cases} «precision» \cong \frac{1}{1+\frac{\alpha}{1-\beta}} = \frac{1}{1+\frac{P_Л}{P_0}}, «accuracy» \cong 1-\frac{\alpha+\beta}{2} = 1-\frac{P_Л+(1-P_0)}{2}, \\ «F\text{-}score» \cong \frac{1}{1+\frac{\alpha+\beta}{2(1-\beta)}} = \frac{1}{1+\frac{P_Л+(1-P_0)}{2\cdot P_0}}. \end{cases} \quad (17)$$

Метрика *ROC*-кривая, по сути, – это зависимость величины $P_0(\alpha)$ бинарного классификатора в пределах (0…1) при изменении какого-нибудь важного параметра, чаще всего – порога обнаружения (принятия решения) [8]. Чем дальше *ROC*-кривая отстоит от биссектрисы угла, тем лучше алгоритм (обнаружитель) при прочих равных условиях. Несмотря на наглядность, физического смысла метрика *ROC* не имеет.

Метрика *AUC* – это площадь под *ROC*-кривой; чем ближе ее величина к 1, тем выше качество классификатора. Она обычно используется для сравнения эффективности алгоритмов обнаружения/классификации, разработанных при помощи обучающей выборки. Однако эта метрика является весьма чувствительной к шуму [8,11]. Отмечена проблема при ее использовании для сравнения алгоритмов; оказывается, что величине *AUC* присваивается вес больший, чем случайно выбранному негативному решению [12]. Практическая ценность показателя *AUC* также ставится под сомнение (ввиду сложности получения его надежной оценки), что зачастую он вносит больше неопределенности, чем ясности [13]. Вышесказанное дает основание для утверждения, что оценка качества бинарного классификатора

при использовании таких метрик, как *ROC*-кривая и *AUC*, не является однозначной и надо соблюдать определенную «осторожность».

Выводы

В результате работы, согласно выражениям (12)–(16), установлена аналитическая связь между «старыми» показателями (ошибки 1-го и 2-го рода) и «новыми» метриками, предназначенными для оценки эффективности бинарного классификатора (обнаружителя). Выяснено, что при относительно «большом» объеме БД, метрики *«specificity»* и *«recall»* однозначно отображают ошибки 1-го и 2-го рода и являются состоятельными, т.е. не зависят от соотношения *N/P* в наборе данных, имеют физический смысл и ими можно пользоваться без всяких ограничений.

Связь метрик *«precision», «accuracy», «F-score»* с ошибками 1-го и 2-го рода достаточно сложная, физического смысла они не имеют, являются несостоятельными, завися от соотношения данных в БД. Оценивать качество работы бинарных классификаторов по ним, используя базы данных с разным соотношением данных, нежелательно, поскольку это может приводить к ошибкам при сравнении эффективности двух разных классификаторов.

Показано, что использование *ROC*-кривых и метрики *AUC* в ряде случаев может привести к ошибочным результатам.

Литература

1. Шелухин О.И. Сетевые аномалии: Обнаружение, локализация, прогнозирование. М.: Горячая линия-Телеком, 2019. 448 с.

2. Мензелинцева Н.В., Карапузова Н.Ю., Статюха И. М., Попова Е.В. К вопросу о повышении достоверности оценки качества воздушной среды урбанизированной территории // Инженерный вестник Дона. 2020. №3. ivdon.ru/ru/magazine/archive/n3y2020/6352.

3. Исмагилова А.С., Лушников Н.Д. Комплексная биометрическая

аутентификация пользователей информационной системы с применением нейронных сетей // Инженерный вестник Дона. 2024. №1. ivdon.ru/ru/magazine/archive/n1y2024/8961.

4. Магауенов Р.Г. Системы охранной сигнализации: основы теории и принципы построения. М.: Горячая линия-Телеком, 2008. 496 с.

5. Микова С.Ю., Оладько В.С. Оценка качества алгоритма обнаружения сетевых аномалий на основе дискретного вейвлет-преобразования с помощью F-меры // Вестник УрФО: Безопасность в информационной сфере. 2015. №2(16). С.36-40.

6. Zhu W., Zeng N., Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations // Proc. NESUG-2010: Health care and life sciences. 2010. P.1-9.

7. Powers D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC: Informedness, Markedness and Correlation // Machine Learning Technologies. 2011. №2(1). P.37–63.

8. Шелухин О.И., Зегжда Д.П., Раковский Д.И. и др. Интеллектуальные технологии информационной безопасности. М.: Горячая Линия-Телеком, 2023. 384 с.

9. Egan J.P. Signal detection theory and ROC analysis, Series in Cognition and Perception. New York. Academic Press, 1975.

10. Bradley A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms // Pattern Recognition. 1997. №30(7). P.1145-1159.

11. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. №27. P.861-874.

12. Hanley J.A., McNeil B.J. A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases // Radiology.1983. №148(3). P.839-843.

13. Hand D.J. Measuring classifier performance: A coherent alternative to the

area under the ROC curve // Machine Learning. 2009. № 77. P.103-123.

References

1. Shelukhin O.I. Setevye anomalii: Obnaruzhenie, lokalizaciya, prognozirovanje [Net anomalies: detection, localization, forecasting]. M.: Goryachaya Liniya-Telekom, 2019. 448 p.

2. Menzelintseva N.V., Karapuzova N.Yu., Statjukha I.M., Popova E.V. Inzhenernyj vestnik Dona. 2020. №3. URL: ivdon.ru/ru/magazine/archive/n3y2020/6352.

3. Ismagilova A.S., Lushnikov N.D. Inzhenernyj vestnik Dona. 2024. №1. URL: ivdon.ru/ru/magazine/archive/n1y2024/8961.

4. Magauenov R.G. Sistemy okhrannoy signalizatcii: osnovy teorii i printsypy postrojenia [Security alarm systems: theory basic and construction principles]. M.: Goryachaya Liniya-Telekom, 2008. 496 p.

5. Mikova S.Yu., Oladko V.S. Vestnik UrFO: Bezopasnost v informatsionnoj sfere. 2015. №2(16). P.36-40.

6. Zhu W., Zeng N., Wang N. Proc. NESUG-2010: Health care and life sciences. 2010. P. 1-9.

7. Powers D.M.W. Machine Learning Technologies. 2011. №2(1). pp. 37–63.

8. Shelukhin O.I., Zegzhda D.P., Rakovskiy D.I. et al. Intelektualnyje technologii v informatsionnoj bezopasnosti [Intelligent information security technologies]. M.: Goryachaya Liniya-Telekom, 2023. 384 p.

9. Egan J.P. New York. Academic Press, 1975.

10. Bradley A.P. Pattern Recognition. 1997. №30(7). pp. 1145–1159.

11. Fawcett T. Pattern Recognition Letters. 2006. №27. pp. 861-874.

12. Hanley J.A., McNeil B.J. Radiology.1983. №148(3). pp. 839-843.

13. Hand D.J. Machine Learning. 2009. №77. pp. 103-123.