Обзорная статья
УДК 004.85+004.052.2
ББК 32.972.133+32.813
М 69
DOI: 10.53598/2410-3225-2022-4-311-52-59

Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения

(*Рецензирована*)

Алексей Андреевич Михайличенко

Институт математики, механики и компьютерных наук, Южный федеральный университет, Ростов-на-Дону, Россия, alexey.a.mikh@gmail.com

Аннотация. *Рассмотрены основные методы оценки и анализа эффективности различных алгоритмов классификации с описанием особенностей этих методов и примеров их использования. Приводится определение матрицы ошибок и обширный набор различных метрик, которые опираются на данные матрицы ошибок. Метрики описаны для случая бинарной классификации, затем дается расширение методики оценки на случай нескольких классов.*

Ключевые слова: *автоматическая классификация, методы оценки классификации, матрица ошибок*

Review Paper

Analytical review of methods for assessing the quality of classification algorithms

Aleksey A. Mikhaylichenko

Institute for Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, Russia, alexey.a.mikh@gmail.com

Abstract*. The article discusses the main methods for evaluating and analyzing the effectiveness of various classification algorithms with a description of the features of these methods and examples of their use. The publication provides the definition of the confusion matrix and an extensive set of different metrics basing on the data of the error matrix. The work describes the metrics for the case of binary classification, and then gives an extension of the estimation technique to the case of several classes.*

Keywords*: automatic classification, classification evaluation methods, confusion matrix*

Введение. Методы классификации находят широкое применение в самых разнообразных областях науки и техники, поэтому важно правильно оценивать эффективность используемого алгоритма при решении поставленной задачи. В простейшем случае в качестве оценки качества классификации можно использовать способность классификатора корректно идентифицировать различные классы, однако в отдельных случаях этого бывает недостаточно, и важно учитывать какие-либо другие критерии. Например, в медицине из-за стремления корректно обнаружить как можно больше больных из общего набора полезной будет склонность классификатора избегать ложноотрицательных срабатываний, или чувствительность к неверной классификации, пусть даже за счет ложноположительных срабатываний.

Существуют различные способы оценки производительности алгоритмов классификации: можно выделить графические (например, использование ROC-кривых [1]) и численные методы оценки (accuracy, sensitivity, specificity [2]). В данной работе вни-

мание уделено численным методам как наиболее распространенным. Большинство этих способов оценки строится на использовании так называемой матрицы ошибок, или *матрицы неточностей* (confusion matrix), которая содержит в себе количество корректно и некорректно классифицированных примеров для каждого класса [3]. На рисунке 1 приводится пример такой матрицы для случая бинарной классификации, где один класс трактуется как *позитивный* (P), а второй – как *негативный* (N).

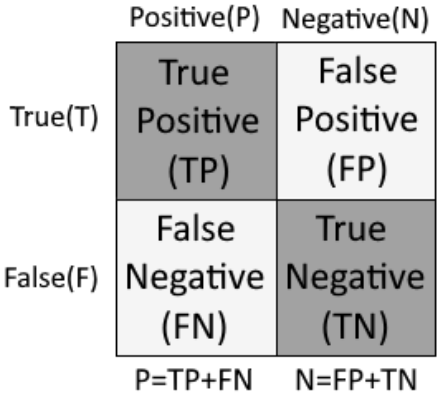

Рис. 1. Иллюстрация матрицы ошибок для бинарной классификации
Fig. 1. An illustration of the confusion matrix for binary classification

Главная диагональ матрицы содержит количество корректно предсказанных образцов для обоих классов, остальные значения – количество ошибочных предсказаний. Так, если позитивный класс был верно предсказан позитивным, он обозначается True Positive (TP); если он был ошибочно распознан как негативный, то обозначается как False Negative (FN), или ошибкой второго типа [2]. Если образец негативного класса был корректно распознан негативным, то он обозначается True Negative (TN), а в случае ошибки, когда он был распознан как позитивный – False Positive (FP), или ошибка первого типа. Упомянутые значения матрицы ошибок используются для вычисления довольно большого количества метрик оценки классификатора.

Accuracy. Наиболее распространенным эмпирическим методом оценки точности классификации является процент корректно классифицированных примеров – *правильность* (или accuracy):

$$\textbf{accuracy} = \frac{TP + TN}{FP + FN + TP + TN}.$$

Данная мера не акцентируется на каком-то конкретном классе, и оценивается ситуация в целом, без детального анализа. В случаях, когда набор данных не сбалансирован (то есть число примеров для каждого класса значительно отличается), данная оценка может вводить в заблуждение.

Например, пусть нам необходимо выполнить бинарную классификацию на наборе данных, в котором один класс представлен 90 образцами, а другой содержит всего 10 примеров. Алгоритм может получить точность 90%, просто если будет предсказывать все подаваемые ему на вход данные как принадлежащие классу 1, что, очевидно, некорректно. Это происходит из-за того, что мы не учитываем распределение тренировочных примеров каждого из классов. Кроме того, данный показатель не дает информации о том, где именно ошибается классификатор.

Последняя причина может быть очень важна в медицинских исследованиях. К примеру, здесь необходимо быть точно уверенным, что классификатор не выдает ложноотрицательные (false negative) результаты – то есть в том, что случай болезни не будет классифицирован ложно при ее наличии, так как это может привести к серьезным

последствиям. В то же время меньше проблем возникнет в случае, если здоровый пациент будет классифицирован как больной – так называемая ложноположительная (false positive) классификация. В данном случае необходимо будет просто провести дополнительную проверку, и эта ошибка не приведет к серьезным последствиям.

Sensitivity и **specificity.** Для оценки эффективности классификатора по разным классам используют метрики *специфичность* (specificity, или true negative rate, TNR) и *чувствительность* (sensitivity, или true positive rate, TPR):

$$\textbf{specificity} = \textbf{TNR} = \frac{TN}{TN + FP},$$

$$\textbf{sensitivity} = \textbf{TPR} = \frac{TP}{TP + FN}.$$

Таким образом, специфичность – это процент примеров негативного класса, которые были корректно распознаны, а чувствительность – часть корректно классифицированных позитивных примеров. То есть в целом чувствительность и специфичность можно рассматривать как метрику *accuracy* для образцов позитивного и негативного классов.

Данные метрики обычно используют в ситуациях ориентации на один класс, когда количество примеров, принадлежащих одному классу, существенно ниже, чем общее количество примеров – в биоинформатике, обработке естественного языка, классификации текстов. Иначе говоря, в случаях, когда среди всех классов есть класс, представляющий особый интерес, а остальные классы либо объединены в один (бинарная классификация), либо оставлены как есть [1].

Positive predictive value и **negative predictive value.** Иногда бывает полезной *предсказательная ценность положительного результата* (positive predictive value, PPV), которая обозначает вероятность того, что результат действительно положительный при предсказании положительного класса, и *предсказательная ценность отрицательного класса* (negative predictive value, NPV), то есть вероятность того, что класс действительно отрицательный при предсказанном отрицательном классе:

$$\textbf{PPV} = \frac{TP}{TP + FP},$$

$$\textbf{NPV} = \frac{TN}{TN + FN}.$$

В приложениях к медицинским исследованиям PPV обозначает вероятность того, что заболевание присутствует при положительном результате теста, а NPV – вероятность того, что болезнь отсутствует при отрицательном результате теста [4].

False positive и **false negative rate.** По аналогии с TPR в некоторых приложениях бывает полезно значение, называемое *долей ложноположительных результатов* (false positive rate, FPR), которое представляет собой отношение неправильно классифицированных отрицательных образцов к общему количеству отрицательных образцов. Другими словами, это доля отрицательных примеров, которые были неправильно классифицированы. Наряду с FNR (false negative rate), которая показывает процент примеров положительного класса, неправильно классифицированного, и дополняет меру sensitivity, FPR дополняет метрику specificity [2]:

$$\textbf{FPR} = 1 - \textbf{TNR} = \frac{FP}{FP + TN} = \frac{FP}{N},$$

$$\textbf{FNR} = 1 - \textbf{TPR} = \frac{FN}{FN + TP} = \frac{FN}{P}.$$

И FPR, и FNR не чувствительны к особенностям распределения данных и поэтому могут использоваться в задачах с несбалансированными датасетами.

Precision и **recall.** Существуют метрики, которые позволяют оценить корректность классификации примеров для разных классов. Наиболее информативными среди них являются *точность* (precision) и *полнота* (recall), которая эквивалентна чувствительности:

$$\textbf{precision} = \frac{TP}{TP + FP},$$

$$\textbf{recall} = \textbf{sensitivity} = \frac{TP}{TP + FN}.$$

Является ли более важным precision или recall – зависит от специфики конкретной задачи классификации. Например, для задачи обнаружения болезни важным является способность классификатора корректно обнаружить как можно больше больных из общего набора, и в этом случае более важным параметром является recall. С другой стороны, важной характеристикой классификатора также можно назвать его чувствительность к неверной классификации, то есть способность обнаруживать как можно больше примеров наличия заболевания даже за счет ложноположительных срабатываний – в таком случае важной характеристикой будет precision. Данные метрики, например, используются в работе [5], которая посвящена автоматической диагностике рака кожи по результатам дерматоскопии, а также в работах, решающих задачу автоматической классификации остеоартрита по рентгенограммам коленного сустава [6, 7]. Однако в большинстве случаев приходится идти на компромиссы при выборе в качестве основной метрики между precision и recall.

Diagnostic odds ratio. В медицинских задачах иногда полезно иметь представление о *диагностическом отношении шансов* (diagnostic odds ratio, DOR [8]). Данный показатель показывает отношение шансов теста быть положительным, если у пациента действительно есть заболевание, к шансам теста быть отрицательным, если у пациента заболевания нет:

$$\textbf{DOR} = \frac{TP / FP}{FN / TN} = \frac{TP \cdot TN}{FP \cdot FN}.$$

Positive likelihood ratio. Также в медицинских задачах иногда используется *отношение правдоподобия положительного* результата теста (positive likelihood ratio, сокр. LR+ [9]), которое представляет собой отношение вероятности положительного результата теста при наличии заболевания (TPR) к вероятности положительного результата теста при отсутствии заболевания (FPR), то есть

$$\textbf{LR}+ = \frac{TPR}{FPR} = \frac{\text{sensitivity}}{1 - \text{specificity}}.$$

Negative likelihood ratio. В пару к этому критерию еще используется *отношение правдоподобия отрицательного результата* теста (negative likelihood ratio, сокр. LR– [9]) – отношение вероятности отрицательного результата теста при наличии заболевания (FNR) к вероятности отрицательного результата теста при отсутствии заболевания (TNR), то есть

$$\textbf{LR}- = \frac{FNR}{TNR} = \frac{1 - \text{sensitivity}}{\text{specificity}}.$$

Индекс Юдена. Индекс Юдена (Youden's index) позволяет оценить возможности классификатора избежать неудачи и представляет собой разницу между долей

истинно положительных результатов (TRP) и долей ложноположительных результатов (FPR):

$$\mathbf{J} = TPR - FPR = \text{sensitivity} - (1 - \text{specificity}).$$

Изначально данный критерий был предложен для сравнения диагностических возможностей двух медицинских тестов [1].

Balanced accuracy. Иногда, в случае несбалансированных классов, используется *сбалансированная точность*, которая является средним значением чувствительности и специфичности и выражается формулой

$$\mathbf{BA} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right).$$

Если в задаче бинарной классификации количество примеров двух классов примерно поровну, то справедливо выражение

$$\text{TP} + \text{FN} \approx \text{TN} + \text{FP} \approx \mathbf{m/2},$$

где **m** – общее количество примеров, и сбалансированная точность примерно будет равна обычному значению accuracy.

F-мера. В некоторых случаях бывает удобно каким-либо образом объединить точность и полноту в одно число, то есть получить своеобразный агрегированный критерий точности работы алгоритма. В этом случае используют F-меру, которая является средним гармоническим точности и полноты, вместо среднего арифметического, что позволяет сглаживать расчеты за счет исключения экстремальных значений. В общем виде F-мера выглядит так:

$$\mathbf{F}_\beta = (1 + \beta^2) \cdot \frac{\text{precision} * \text{recall}}{(\beta^2 * \text{precision}) + \text{recall}}.$$

Параметр β – это вес точности в метрике, и обычно в исследованиях используется значение $\beta=1$, то есть используют F_1-меру:

$$\mathbf{F}_1 = 2 \cdot \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю. Мера F_1 используется в работах по автоматической диагностике заболеваний коленного сустава наряду с метриками *precision* и *recall* [6, 7].

Cohen's Kappa. Еще одним коэффициентом, по которому можно судить о качестве работы классификатора в случае использования несбалансированных датасетов, является коэффициент *Коэна* (Cohen's Kappa [10]). Основной идеей данного коэффициента является перенормировка значений точности при помощи значения точности, которое можно было бы получить случайно. Выражается коэффициент формулой

$$\mathbf{к} = \frac{\text{accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}.$$

При этом *random accuracy* можно вычислить следующим образом:

$$\mathbf{random\ accuracy} = \frac{(TP + FN)(TP + FP)}{FP + FN + TP + TN} \cdot \frac{(FN + TN)(FP + TN)}{FP + FN + TP + TN},$$

где первое слагаемое означает вероятность того, что верно будет угадан класс P, второе слагаемое – вероятность того, что верно будет угадан класс N. В таком случае финальная формула для коэффициента **к** будет выглядеть так:

$$\mathbf{к} = 2 \cdot \frac{(TP \cdot TN - FN \cdot FP)}{(TP + FP)(FP + TN) + (TP + FN)(FN + TN)}.$$

Стоит отметить, что если значение метрики *accuracy* равно единице, то и значение коэффициента **к** будет равно 1. Этот коэффициент полезен в тех случаях, когда точность случайного угадывания достаточно высока (например, классы сильно разбалансированы, и предсказывание просто самого часто встречающегося класса даст высокий результат по сравнению с осмысленной классификацией).

Мультиклассовая классификация. Матрица ошибок может быть составлена не только для бинарного, но и для мультиклассового классификатора. Пример такой матрицы ошибок для случая трех классов A, B и C представлен на рисунке 2.

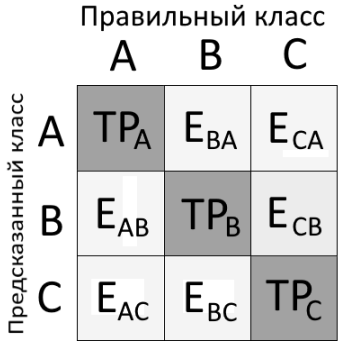

Рис. 2. Иллюстрация матрицы неточностей для мультиклассовой классификации (для случая трех классов)

Fig. 2. An illustration of the confusion matrix for multiclass classification (for the case of three classes)

Здесь, по аналогии с матрицей для бинарного классификатора, главная диагональ (значения TP_A, TP_B, TP_C) обозначает количество верно предсказанных примеров для каждого из классов, оставшиеся ячейки матрицы – ошибочные предсказания. Например, E_{BA} – количество примеров, являющихся классом A, но предсказанных как B; E_{CA} – количество примеров, являющихся классом A, но предсказанных как C и т.д.

Ложноположительной (false positive) ошибкой для класса A будет сумма $FP_A = E_{BA} + E_{CA}$, то есть количество примеров, которые были классифицированы как класс A, но им на самом деле не являются. *Ложноотрицательной* (false negative) ошибкой для класса A будет сумма $FN_A = E_{AB} + E_{AC}$, которая показывает количество примеров класса A, ошибочно классифицированных как класс B или C.

Таким образом, матрица неточностей размера $m \times m$ содержит m ячеек корректных классификаций, и $m^2 - m$ ячеек с ошибочными результатами [2]. Характеристики FN, FP, TN, TP для всего классификатора вычисляются как сумма соответствующих характеристик по всем классам, то есть $FN = FN_A + FN_C + FN_C$ и т.д.

В случае подсчета метрик для мультиклассового классификатора возможны различные варианты подсчета метрик классификатора, не привязанных к конкретным классам:

- **микро**: вычисление глобальных метрик путем подсчета общего количества ложно-негативных (FN), ложно-позитивных (FP) и других параметров (вместо подсчета индивидуальных метрик для каждого класса);
- **макро**: вычисление метрики для каждого класса и определение их невзвешенного среднего – в данном случае не учитывается сбалансированность датасета и вес каждой метрики среди всего набора;
- **взвешенное среднее**: вычисление метрики для каждого класса и определение их взвешенного среднего, опираясь на процентное соотношение количества образцов

каждой метрики среди общего количества примеров в датасете – обычно используется в случае сильно несбалансированных датасетов.

При вычислении метрики для каждого конкретного класса сам класс фиксируется как позитивный, а все остальные – как негативный (то есть ситуация сводится к бинарному классификатору). Кроме этого, при выборе первого способа подсчета (микрометрики) из-за особенностей подсчета параметров матрицы ошибок (TP – общее количество корректно классифицированных классов, FN = FP – общее количество некорректно классифицированных классов) значение метрики *recall* будет совпадать с *accuracy*. Более того, мы можем сказать, что в таком случае справедливо следующее:

$$micro\text{-}F_1 = micro\text{-}precision = micro\text{-}recall = accuracy.$$

Заключение. Приведен сравнительный обзор существующих методов оценки качества классификации для использования в широком круге задач с различными критериями эффективности. Описанные методы могут быть использованы как в машинном обучении, так и в других областях. Для каждой методики оценки приведен способ ее вычисления и дано описание особенностей использования.

В работах [6, 7], посвященных автоматической медицинской диагностике, используются такие индивидуальные метрики оценки точности классификации, как accuracy, precision, recall и F_1-мера. Данный набор метрик является классическим при работе с задачами медицинской диагностики.

В работе была проведена оценка метода на тестовом множестве и построена соответствующая матрица ошибок, на основе которой получены необходимые статистические характеристики для вычисления метрик качества. Метрика accuracy является одним из самых распространенных показателей работы классификаторов и была использована для получения первичной оценки точности методов. Precision и recall использовались для получения сравнительной способности тестируемого алгоритма корректно обнаружить как можно больше больных из общего набора и чувствительность выбранного метода к неверной классификации. В качестве комбинации этих показателей используется F_1-мера.

Ввиду того, что рассматриваемая шкала оценки остеоартрита состоит из нескольких классов, для подсчета метрик был применен метод взвешенного среднего, который позволил учесть несбалансированность используемых для обучения наборов данных – так, для некоторых классов остеоартрита число имеющихся примеров отличалось от остальных в десятки раз. Например, для класса 2 используемой шкалы оценки обучающая выборка состояла из 1504 изображений, а для класса 4 всего из 175 [7]. Использование других способов могло показывать искаженные результаты из-за преобладания одних классов над другими.

Примечания

1. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation // Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science. 2006. Vol. 4304. P. 1015–1021.

2. Tharwat A. Classification assessment methods // Applied Computing and Informatics. 2021. Vol. 17, No. 1. P. 168–192.

3. Stehman S.V. Selecting and interpreting measures of thematic classification accuracy // Remote Sensing of Environment. 1997. Vol. 62. P. 77–89.

4. Корнеенков А.А., Рязанцев С.В., Вяземская Е.Э. Вычисление и интерпретация показателей информативности диагностических медицинских технологий // Медицинский совет. 2019. № 20. С. 45–51.

5. Rezvantalab A., Safigholi H., Karimijeshni S. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms //

ArXiv. 2018. Vol. abs/1810.10348.

6. Mikhaylichenko A., Demyanenko Y. Automatic Grading of Knee Osteoarthritis from Plain Radiographs Using Densely Connected Convolutional Networks // Recent Trends in Analysis of Images, Social Networks and Texts. 2021. P. 149–161.

7. Михайличенко А.А., Демяненко Я.М. Использование блоков сжатия и возбуждения для повышения точности автоматической классификации остеоартрита коленного сустава при помощи сверточных нейронных сетей // Компьютерная оптика. 2022. № 46 (2). С. 317–325.

8. The diagnostic odds ratio: a single indicator of test performance / A.S. Glas, J.G. Lijmer, M.H. Prins [et al.] // Journal of Clinical Epidemiology. 2003. Vol. 56, No. 11. P. 1129–1135.

9. Pauker S.G., Kassirer J.P. Therapeutic decision making: a cost-benefit analysis // The New England Journal of Medicine. 1975. Vol. 293, No. 5. P. 229–234.

10. Chang C.H. Cohen's kappa for capturing discrimination // International Health. 2014. Vol. 6, No. 2. P. 125–129.

References

1. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation // Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science. 2006. Vol. 4304. P. 1015–1021.

2. Tharwat A. Classification assessment methods // Applied Computing and Informatics. 2021. Vol. 17, No. 1. P. 168–192.

3. Stehman S.V. Selecting and interpreting measures of thematic classification accuracy // Remote Sensing of Environment. 1997. Vol. 62. P. 77–89.

4. Korneenkov A.A., Ryazantsev S.V., Vyazemskaya E.E. Calculation and interpretation of indicators of informativeness of diagnostic medical technologies // Medical Council. 2019. No. 20. P. 45–51.

5. Rezvantalab A., Safigholi H., Karimijeshni S. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms // ArXiv. 2018. Vol. abs/1810.10348.

6. Mikhaylichenko A., Demyanenko Y. Automatic Grading of Knee Osteoarthritis from Plain Radiographs Using Densely Connected Convolutional Networks // Recent Trends in Analysis of Images, Social Networks and Texts. 2021. P. 149–161.

7. Mikhaylichenko A.A., Demyanenko Ya.M. Using squeeze-and-excitation blocks to improve an accuracy of automatically grading knee osteoarthritis severity using convolutional neural networks // Computer Optics. 2022. No. 46 (2). P. 317–325.

8. The diagnostic odds ratio: a single indicator of test performance / A.S. Glas, J.G. Lijmer, M.H. Prins [et al.] // Journal of Clinical Epidemiology. 2003. Vol. 56, No. 11. P. 1129–1135.

9. Pauker S.G., Kassirer J.P. Therapeutic decision making: a cost-benefit analysis // The New England Journal of Medicine. 1975. Vol. 293, No. 5. P. 229–234.

10. Chang C.H. Cohen's kappa for capturing discrimination // International Health. 2014. Vol. 6, No. 2. P. 125–129.

Статья поступила в редакцию 09.11.2022; *одобрена после рецензирования 29.11.2022*; *принята к публикации 30.11.2022*.

The article was submitted 09.11.2022; *approved after reviewing 29.11.2022*; *accepted for publication 30.11.2022*.