ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.91

МЕТРИКИ ЭФФЕКТИВНОСТИ МОДЕЛИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Иванов Сергей Александрович,
кандидат технических наук, доцент,
Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова

Кочерженко Андрей Александрович,
студент 4-го курса обучения,
Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова

Иванов Максим Александрович,
Студент 2-го курса обучения,
Астраханский Государственный Технический Университет

Аннотация

Классификация является одной из ключевых задач в области интеллектуального анализа данных, и успешная её реализация требует построения надёжных и точных классификаторов. В данной статье рассмотрены основные методы оценки качества моделей интеллектуального анализа данных. Были исследованы особенности каждого метода в отдельности, а также приведены положительные и отрицательные аспекты влияния на оценку качества исследуемой модели.

Ключевые слова: эффективность классификаторов, методы оценки качества, ROC-анализ, PR-кривая.

PERFORMANCE METRICS OF THE DATA MINING MODEL

Ivanov Sergey Alexandrovich,
Candidate of Technical Sciences, Associate Professor,
Saint Petersburg State Forest Technical University named after S.M. Kirova
kemsit@mail.ru

Kocherzhenko Andrey Alexandrovich,
4th year student,
Saint Petersburg State Forest Technical University named after S.M. Kirova

Ivanov Maksim Aleksandrovich,
2th year student,
Astrakhan State Technical University
Maksfire2001@mail.ru

ABSTRACT

Classification is one of the key tasks in the field of data mining, and its successful implementation requires the construction of reliable and accurate classifiers. This article discusses the main methods for evaluating the quality of data mining models. The features of each method were investigated separately, as well as the positive and negative aspects of the impact on the assessment of the quality of the model under study.

Keywords: the effectiveness of classifiers, quality assessment methods, ROC- analysis, PR-curve.

Основное внимание в методах классификации направлено на исследование характеристик и поведения сформированных групп. С применением алгоритмов классификации можно отнести объекты к заранее известным категориям. Для этого часто используются алгоритмы, которые по небольшому набору данных, принадлежащих к известным классам, определяют новые группы.

Для измерения эффективности классификаторов разработано множество метрик, каждая из которых предоставляет важную информацию о различных аспектах работы модели.

Однако оценка качества подобных моделей, особенно для бинарной классификации, довольна сложна в связи с тем, что целевая переменная в классификации является категориальной (дискретной), и ошибка классификации не может быть выражена числовым значением [1].

Поэтому для оценки качества логистической регрессии используется статистика результатов классификации обучающих примеров. С её помощью вычисляются метрики качества — показатели, которые зависят от результатов классификации и не зависят от внутреннего состояния модели [2].

Ниже рассмотрим несколько основных методов оценки качества.

Меткость

Доля правильных классификаций модели (рис. 2.9)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

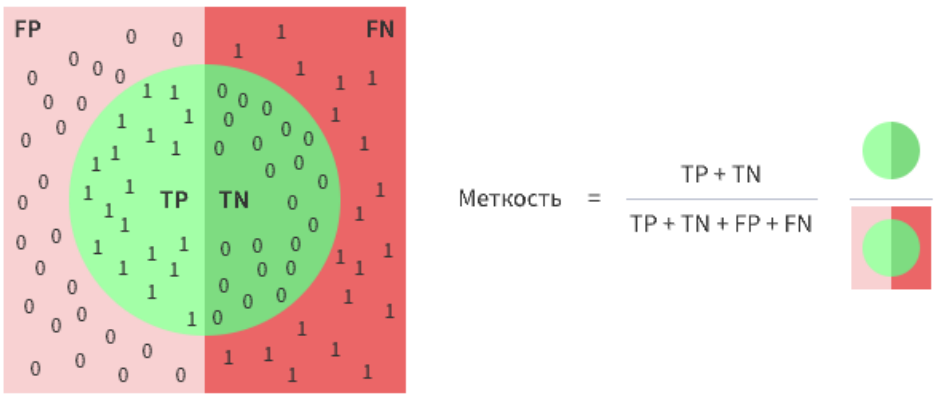

Рисунок 2.9 – меткость, знаменатель – общее число классифицируемых параметров

Данный метод редко используют для определения уровня качества, несмотря на легкость интерпретации, так как в обучающем наборе данных меткость плохо справляется при наличии дисбаланса классов [3].

Точность

Доля действительно (истинно) положительных групп к общему их числу (рис. 2.10).

$$Pr = PPV = \frac{TP}{TP + FP}$$

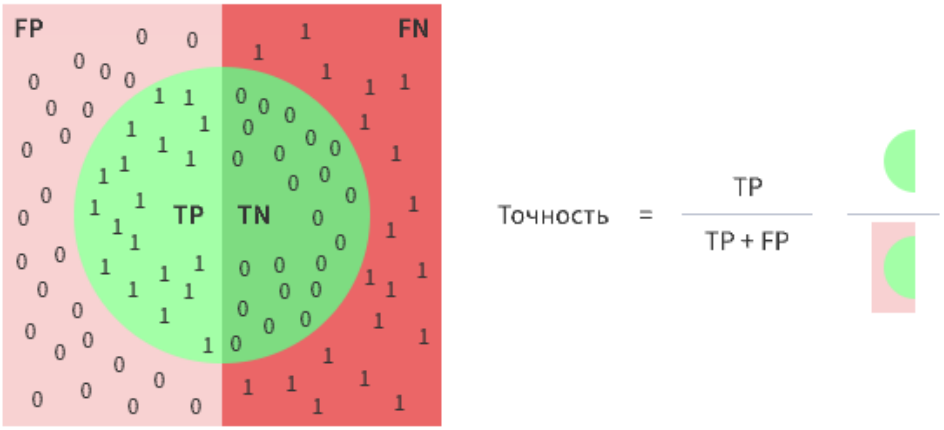

Рисунок 2.10 – точность

Решение объединить объекты в один класс в данном случае приводит к увеличению параметра FP и уменьшению показателя точности.

Полнота

Доля действительно (истинно) положительных примеров по отношению к общей сумме положительных наблюдений (рис.2.12).

$$Re = TPR = \frac{TP}{TP + FN}$$

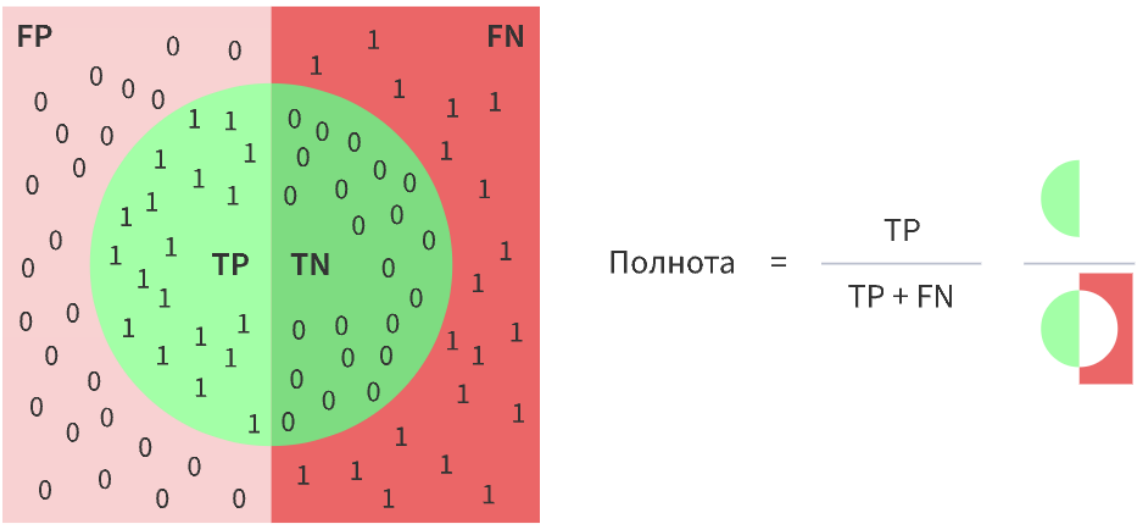

Рисунок 2.11 - Полнота

Специфичность

Доля действительно (истинно) отрицательных классификаций, показывает, насколько хорошо модель распознаёт отрицательные примеры (рис. 2.12).

$$S_p = TNR = \frac{TN}{TN + FP}$$

TNR – истинно отрицательные случаи в общем числе отрицательных примеров.

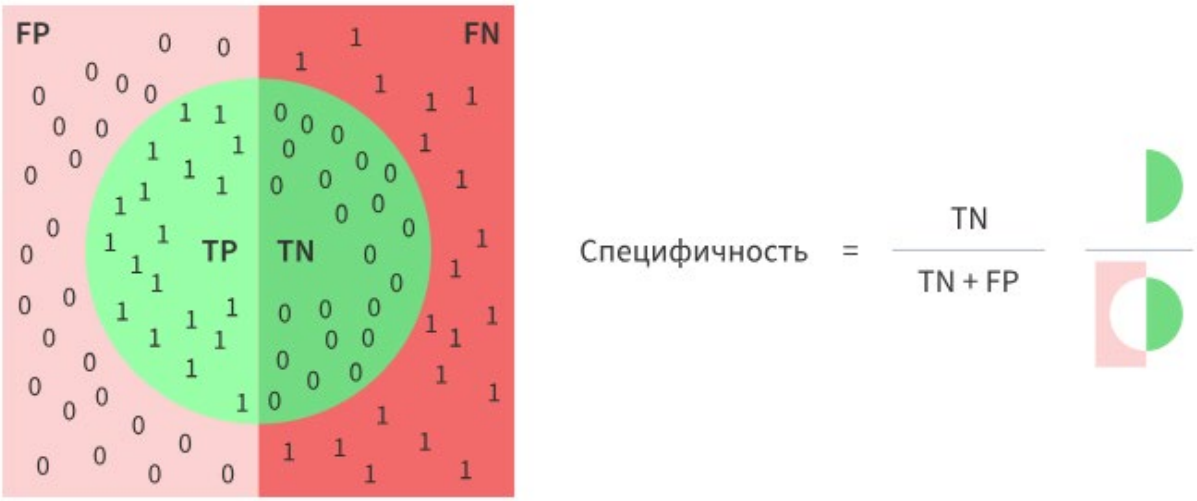

Рисунок 2.12 – Специфичность

Если все отрицательные примеры классифицированы правильно (то есть число ложноположительных случаев равно 0), то TNR будет равен 1.

Площадь под ROC-кривой

Более сложный метод оценки качества, который даёт возможность рассмотреть поведение классификатора при разных уровнях дискриминационного порога [4].

Данная технология используется для оценки качества бинарных классификаций с использованием ROC-кривых и носит название ROC-анализ.

Существуют параметры: TPR (true positive rate) и TNR (true negative rate) классификатора. Рассмотрим значение и расчёт каждого из них.

TPR = 1 в случае, если правильно определены все положительные примеры или отсутствуют ложноотрицательные случаи.

TNR = 1 в случае, если правильно определены все отрицательные примеры или отсутствуют ложноположительные случаи.

Отдельно друг от друга эти характеристики помогают модели определять только один класс, однако вместе они представляют собой метрику, с помощью которой можно задать значение дискриминационного порога и сбалансировать модель для распознавания как положительных, так и для отрицательных примеров единовременно. Это то, что представляет из себя ROC-кривая.

Показатель ложноположительных классификаций 1-TNR вычисляется следующим образом:

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

Если FPR=1, TPR=1, при пороге 1, все примеры определяются как отрицательные.

Если FPR=0, TPR=0, при пороге 0, все примеры определяются как положительные (рис.2.13).

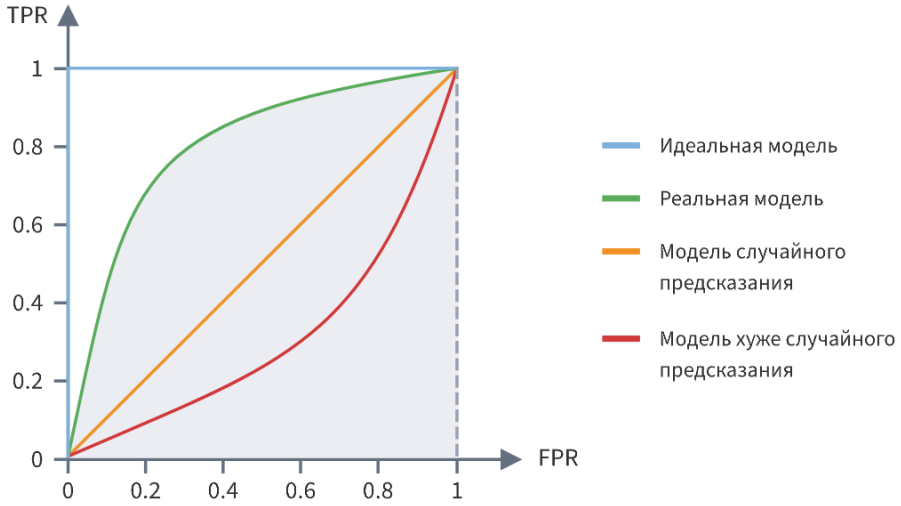

Рисунок 2.13 – примеры ROC-кривых

Представленный рисунок демонстрирует, как для варианта «Идеальная модель» кривая на графике становится ломанной и пересекает точки с координатами (0,0), (0,1), (1,1). В данном случае площадь под ROC-кривой, которая обозначена на рисунке серым цветом, будет составлять 1.

Ситуация, в которой модель в равной степени обучена обрабатывать положительные и отрицательные параметры, это когда оба показателя TPR и TNR равны единице.

Однако идеальная модель является скорее теоретической и, как правило, недостижима на практике [5]. Поэтому обычно приходится работать с ROC-кривыми, которые не проходят через точку (0,1), а лишь приближаются к ней. Соответственно, и значение AUC-ROC оказывается меньше 1.

AUC-PR

PR-кривые аналогичны ROC-кривым, но по оси абсцисс у них откладываются значения полноты, а по оси ординат — точности.

Точность и полнота — ключевые метрики при оценке качества модели бинарной классификации, особенно при несбалансированных классах.

Точность показывает, какая часть параметров, классифицированных как положительные, действительно является положительными, а полнота — какая часть положительных наблюдений была правильно классифицирована.

Если точность равна единице, то ложноположительных классификаций нет, но это не говорит о распознавании всех положительных примеров.

Если полнота равна единице, то все положительные объекты были распознаны, но ничего неизвестно о ложноположительных классификациях.

Поэтому точность и полнота не всегда полезны по отдельности.

Обычно сравнивают значения одной метрики при фиксированном уровне другой или объединяют обе метрики в один показатель, например, F1-мера — среднее точности и полноты [6].

PR-кривые отображают, как выбор порога влияет на точность и полноту классификатора, и помогают выбрать оптимальный порог.

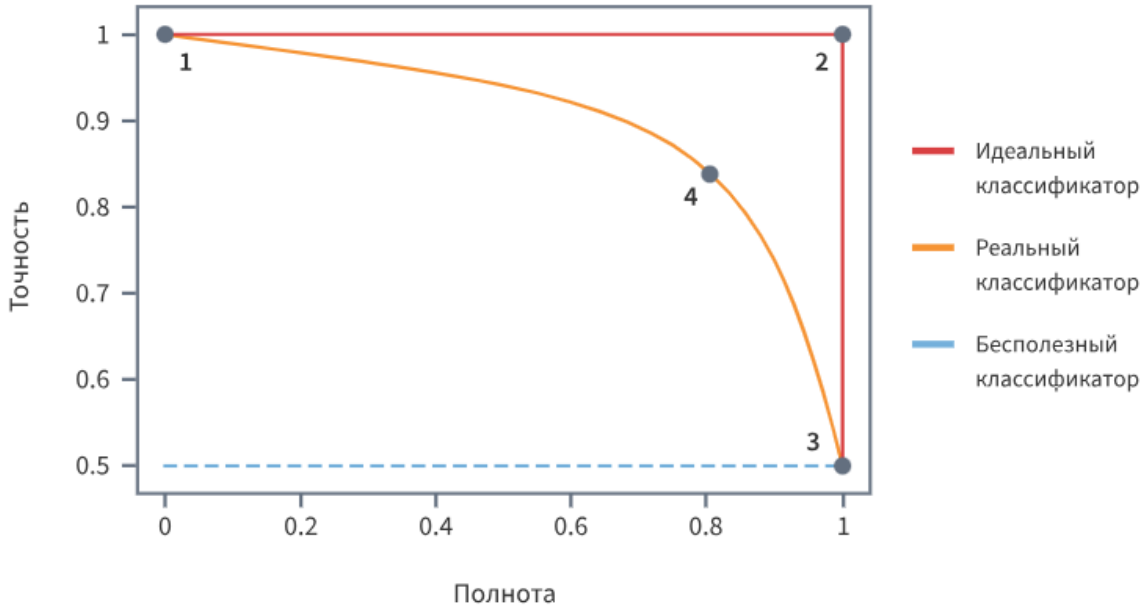

Рисунок 2.14 – Кривая точность-полнота

Каждая точка PR-кривой представляет определенное значение порога и соответствующие ему точность и полноту. Точка 2 на рисунке соответствует идеальному классификатору с координатами (1,1), а точка 4 — оптимальному порогу.

Площадь под PR-кривой (PR-AUC) отражает качество классификатора: чем она больше, тем лучше модель. Пунктирная линия на графике соответствует базовой модели, присваивающей рейтинг 0.5 для любого примера.

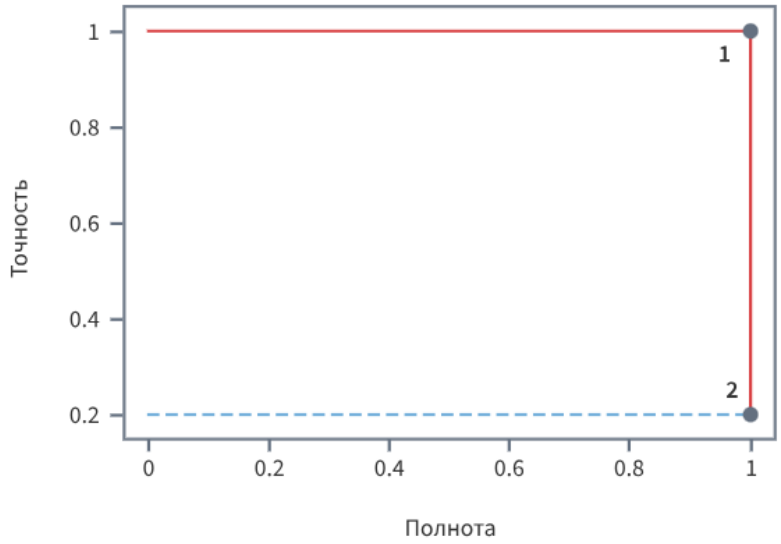

Рисунок 2.15 – Кривая точность-полнота для идеальной модели

Пример идеальной PR-кривой: точка 1 соответствует порогу (0,1], точка 2 — порогу 0, и PR-AUC равен 1.

Для плохой модели PR-кривая расположена ниже линии базовой модели. Улучшить такую модель можно, инвертировав классы. Плохая PR-кривая часто указывает на проблемы в данных, такие как шум или плохую выраженность классов [7].

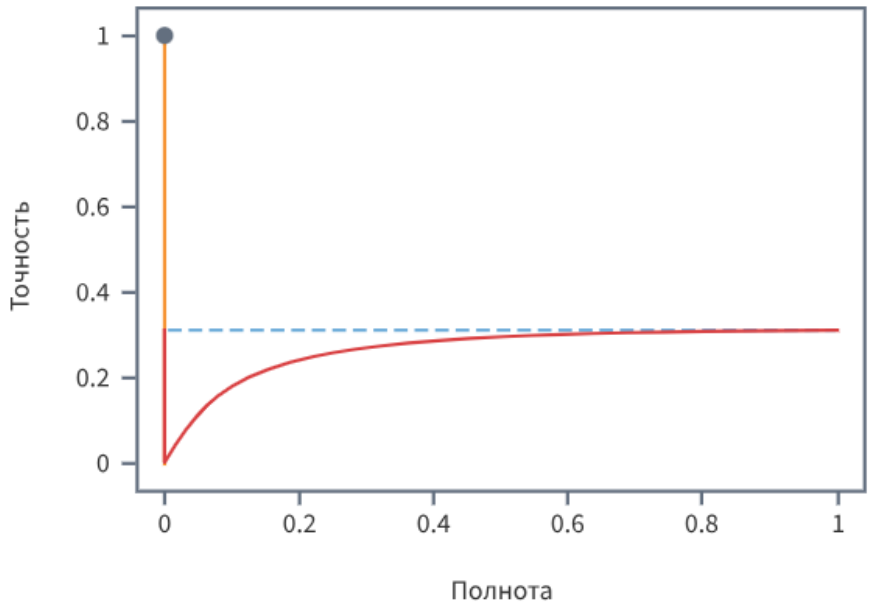

Рисунок 2.16 – Кривая точность-полнота для модели хуже бесполезной

Список литературы:

1. Волков, Д. Качество данных: от стратегии к практике / Д. Волков, А. Незнанов // Открытые системы. СУБД. – 2020. – № 1. – С. 14-18. – DOI 10.26295/OS.2020.86.56.003. – EDN RSHIWC.
2. Грешилов А. А. Математические методы принятия решений; МГТУ им. Н. Э. Баумана - Москва, 2012.
3. Захарова А. А. Математические и инструментальные методы поддержки принятия решений / А. А. Захарова, А. А. Мицель. – Томск : ТУСУР, 2019. – 114 с.
4. Иванов, С. А. Обзор инструментов интеллектуального анализа данных современных ИТ-платформ для решения задач прогнозирования / С. А. Иванов // Информационные системы и технологии: теория и практика : научно-техническая конференция Института леса и природопользования СПбГЛТУ, Санкт-Петербург, 25 февраля 2022 года. Том Выпуск 14. – Санкт-Петербург: Санкт-Петербургский государственный лесотехнический университет имени С.М. Кирова, 2022. – С. 141-143.
5. Яковлев, В. Б. Анализ данных в аналитической платформе Loginom / В. Б. Яковлев. – Saarbrücken: LAP LAMBERT, 2020. – С. 15-62.
6. Шершнев Р.В., Радаев А.В., Коробов А.В., Яцало Б.И. Модуль группового многокритериального анализа решений на основе нечеткого расширения метода TOPSIS.
7. Уткин, Л. В. Анализ риска и принятие решений при неполной информации / Л. В. Уткин; Л. В. Уткин. – Санкт-Петербург: Наука, 2007. 404 с.

References:

1. Volkov, D. Data quality: from strategy to practice / D. Volkov, A. Neznanov // Open Systems. DBMS. – 2020. – No. 1. – pp. 14-18. – DOI 10.26295/OS.2020.86.56.003. – EDN RSHIWC.

2. Greshilov A. A. Mathematical methods of decision-making; Bauman Moscow State Technical University - Moscow, 2012.
3. Zakharova A. A. Mathematical and instrumental methods of decision support / A. A. Zakharova, A. A. Mizel. Tomsk: TUSUR, 2019. 114 p.
4. Ivanov, S. A. Overview of data mining tools of modern IT platforms for solving forecasting problems / S. A. Ivanov // Information systems and technologies: theory and practice: scientific and technical conference of the Institute of Forestry and Environmental Management of St. Petersburg State Technical University, St. Petersburg, February 25, 2022. Volume Issue 14. – St. Petersburg: St. Petersburg State Forestry Engineering University named after S.M. Kirov, 2022. – pp. 141-143.
5. Yakovlev, V. B. Data analysis in the Loginom analytical platform / V. B. Yakovlev. – Saarbrücken: LAP LAMBERT, 2020. – pp. 15-62.
6. Shershnev R.V., Radaev A.V., Korobov A.V., Yatsalo B.I. Module for group multicriteria analysis of solutions based on a fuzzy extension of the TOPSIS method.
7. Utkin, L. V. Risk analysis and decision-making with incomplete information / L. V. Utkin; L. V. Utkin. – St. Petersburg: Nauka, 2007. 404 p.