МЕТРИКИ ОЦЕНКИ КЛАССИФИКАТОРОВ В ЗАДАЧАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

П.В. Дудченко
Томский политехнический университет
pv.dudchenko@gmail.com

Введение

В настоящее время методы машинного обучения (Machine Learning, ML) активно используются для решения научных и производственных задач. Классификация – одна из групп алгоритмов ML. Наиболее популярной метрикой для оценки эффективности работы классификационного алгоритма является точность (Accuracy). Но на несбалансированных наборах данных данная метрика часто не показательна и даже ошибочна. Особенно остро этот вопрос стоит в медицинских задачах, когда, например, необходимо предсказать летальный исход.

В данной работе мы рассматриваем метрики качества алгоритмов классификации, применяемые в исследованиях связанных с применением алгоритмов ML в задачах медицинской диагностики.

Метрики качества алгоритмов классификации

Мы отобрали релевантные публикации по запросам «Artificial Intelligence in Cardiology», «Decision Support System in Cardiology», «Expert System for Diagnosis Cardiovascular Disease» в двух библиографических базах PubMed и Web of Science. Рассматривались только научные работы, написанные на английском языке и опубликованные в период 2011 по 2018г. Всего таким критериям отвечало 700 работ. После исключения повторяющихся статей для дальнейшего рассмотрения осталось 437 работ. Проанализировав названия, было удалено еще 370 исследований, не связанных с предметом, 67 тезисов были отобраны, и 26 статей были исключены из окончательного списка. В итоге к рассмотрению было получено 41 исследование. Мы не можем здесь привести ссылки на эти работы из-за жестких ограничений на объем данной публикации. Из рассмотренных работ мы выписали все используемые метрики и методики оценки классификационных алгоритмов и приводим их далее.

Матрица ошибок

Как правило, результаты решения проблемы двоичной классификации помечены как положительные и негативные. Эти решения могут быть представлены в матрице ошибок (Confusion Matrix) (рис. 1), которая содержит 4 ячейки.

- Верно-положительные (TP), объекты, которые были классифицированы как положительные и действительно являются положительными (принадлежащими к данному классу);
- Верно-отрицательные (TN) объекты, которые были классифицированы как отрицательные и действительно отрицательные (не принадлежат к данному классу);
- Ложно-положительные (FP) объекты, которые были классифицированы как положительные, но фактически отрицательные;
- Ложно-отрицательный (FN) объекты, которые были классифицированы как отрицательные, но фактически положительные;

Категория i		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

Рис 1. Матрица ошибок

На основе матрицы ошибок и её значений, рассчитываются различные метрики классификационной способности алгоритма. Следующие метрики чаще всего были указаны в проанализированных нами работах: Accuracy, Точность (precision), Полнота (recall), F-мера (F-measure), Специфичность, Площадь под кривой ошибок (AUC ROC).

Accuracy – широко используемая и легкая для понимания метрика. Это отношение всех правильных прогнозов к общему числу всех предсказанных образцов. В ряде задач accuracy может являться неинформативной. Например, предположим, что только 3% всех пациентов имеют некоторые заболевания. Создадим алгоритм, который будет помечать всех пациентов, как здоровых. В итоге он будет ошибочным только в 3 случаях из 100. Accuracy будет равно 0,97 или 97% и это высокая оценка, но на самом деле алгоритм ничего не делает и абсолютно бесполезен.

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Точность (precision) - это доля прогнозируемых положительных результатов, которые являются действительно верно-положительными результатами для всех положительно предсказанных объектов. Другими словами, точность дает нам ответ на вопрос «Из всех объектов, которые классифицированы как принадлежащие классу, сколько на самом деле принадлежит ему?»

$$\text{Точность} = \frac{TP}{TP + \mathrm{FP}}$$

Полнота (recall) – пропорция всех верно-положительно предсказанных объектов к общему количеству действительно положительных. То есть, полнота показывает сколько образцов из всех положительных примеров были классифицированы правильно. Чем выше значение полноты, тем меньше положительных примеров пропущено в классификации.

$$\text{Полнота} = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

Исследователи в своих работах часто используют такую метрику, как **чувствительность** (**sensitivity**). На самом деле чувствительность и полнота оценивают одно и то же, различие в наименовании возникло из-за принадлежности этих терминов к разным областям науки. Так же встречаются названия **True positive rate** или **TPR** (оценка верно положительных) и **Probability of detection** (вероятность выявления).

F-мера – взвешенное гармоническое среднее полноты и точности. Этот показатель демонстрирует, как много случаев прогнозируется моделью правильно, и сколько истинных экземпляров модель не пропустит. F-мера объединяет в себе информацию о полноте и точности используемой модели.

$$F - \text{мера} = 2 * \frac{\text{точность} * \text{полнота}}{\text{точность} + \text{полнота}}$$

Специфичность – отношение между верно классифицированных негативных экземпляров к числу всех негативных экземпляров.

$$\text{Специфичность} = \frac{TN}{TN + FP}$$

Другой информативной и обобщающей метрикой является **площадь под кривой ошибок,** что буквально означает площадь под ROC-кривой (Receiver Operating Characteristic, рабочая характеристика приёмника) (рис. 2).

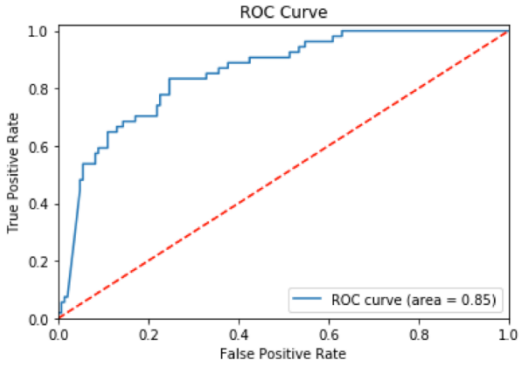

Рис. 2. Пример ROC-кривой

Чтобы получить данную кривую, необходимо вычислить две метрики, которые также получены из матрицы ошибок: уровень верно-положительных предсказанных экземпляров (True Positive Rate, TPR), что тождественно полноте, и уровень ложно-положительных (False Positive Rate, FPR). В свою очередь, FPR - доля негативных объектов, которые по ошибке были классифицированы как положительные, ко всем фактическим негативным. Чем выше FPR, тем больше негативных экземпляров классифицируются не верно.

$$FPR = \frac{FP}{FP + TN}$$

Для объединения FPR и TPR в одну метрику, необходимо вычислить эти метрики, а затем построить их на одном графике с осями FPR и TPR. Результирующая кривая представляет собой кривую ROC, а площадь под кривой является метрикой AUC ROC (рис. 2).

Таблица 1 демонстрирует насколько часто использовались указанные метрики в рассмотренных работах.

Таблица 1. Частота использования метрик в рассмотренных работах

Метрика	Количество публикаций
AucROC	29
Полнота, чувствительность	25
Специфичность	22
Accuracy	18
F-мера	6
Точность	5

Заключение

Наиболее популярным показателем эффективности работы классификаторов в рассмотренных статьях AUC ROC. В то же время, она является самой комплексной и информативной. Полнота и специфичность так же используются часто, потому что эти показатели вычисляются для построения кривой ROC. Accuracy активно применяется в исследованиях, но, как правило, не существует сама по себе, а идет в сочетании с другими метриками. В большинстве случаев показатели качества: полнота, точность и F-мера также приводятся вместе для получения более полной картины о классифицирующей способности классификатора.

Список использованных источников

1. Ohsaki M. et al. Confusion-matrix-based kernel logistic regression for imbalanced data classification //IEEE Transactions on Knowledge and Data Engineering. – 2017. – Т. 29. – №. 9. – С. 1806-1819.
2. Hastie T. Friedman 2009: T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. – 2009.
3. Dudchenko A., Kopanitsa G. Decision Support Systems in Cardiology: A Systematic Review //Studies in health technology and informatics. – 2017. – Т. 237. – С. 209-214.
4. Zhu W. et al. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations //NESUG proceedings: health care and life sciences, Baltimore, Maryland. – 2010. – Т. 19. – С. 67.
5. Flach P. A. The geometry of ROC space: understanding machine learning metrics through ROC isometrics //Proceedings of the 20th International Conference on Machine Learning (ICML-03). – 2003. – С. 19