Вестн. Ом. ун-та. 2016. № 1. С. 14–17.

УДК 512

И.П. Гайдышев

ОЦЕНКА КАЧЕСТВА БИНАРНЫХ КЛАССИФИКАТОРОВ

Рассмотрена оценка качества диагностического метода, представленного в виде модели бинарного классификатора на примере множественной логистической регрессии, через графический анализ соотношения пары «чувствительность – неспецифичность» (ROC-анализ). Предложен алгоритм вычисления оптимального порога отсечения. Оценка статистической значимости метода выполнена на основе площади, отсекаемой ROC-кривой.

Ключевые слова: классификация; регрессия.

Введение

В распознавании образов с обучением (классификации с учителем) возникает ситуация, когда отклик эксперимента представлен в бинарном виде (1 – наличие признака, 0 – отсутствие признака). Для решения задачи может использоваться логит-анализ (множественная логистическая регрессия).

Множественная логистическая регрессия представлена в виде модельной формулы [7]

$$P_j(B) = Logit(X_j B) = \frac{1}{1+e^{-X_j B}},\ j = 1, 2, ..., n,$$

где $P_j(B)$, j = 1,2, …, n – выход модели; $B = \{b_i\}$, i = 1, 2, …, m – вектор-столбец весовых коэффициентов; $X_j = \{x_i\}$, i = 1, 2, …, m; j = 1, 2, …, n – вектор-строка параметров объекта j, измеренных в эксперименте; X_jB, j = 1,2, …, n – множественная линейная регрессия (замечание о модели со свободным членом см. ниже); m – количество измеряемых в эксперименте параметров объекта; n – численность обучающей выборки (число объектов).

Значение $P_j(.)$ может быть интерпретировано как вероятность получения логитом значения 1 при подстановке в уравнение определенного вектора X_j, j = 1, 2, …, n, измеренного в эксперименте.

Оптимальные значения весовых коэффициентов могут быть найдены путем максимизации логарифма функции максимального правдоподобия (далее – ФМП) [3]

$$\ln L = \sum_{j=1}^{n} \left[Y_j \ln P_j(B) + (1 - Y_j) \ln(1 - P_j(B)) \right],$$

где Y_j, j = 1, 2, …, n – выход эксперимента, соответствующий измеренному в эксперименте вектору параметров X_j, j = 1, 2, …, n.

Задача сводится к системе нелинейных алгебраических уравнений

$$\partial \ln L / \partial \ln b_i = 0,\ \ i = 1, 2, ..., m,$$

для решения которой используем стандартный метод Ньютона–Рафсона. Итерационная схема метода записывается формулой [2]

$$B_{k+1} = B_k - \left[H(B_k) \right]^{-1} g(B_k),\ k = 0, 1, ...,$$

где k – номер итерации; $H(.)$ – матрица Гессе (матрица вторых производных) ФМП; $g(.)$ – градиент (вектор производных) ФМП.

Вследствие аналитически доказанной сходимости итерационной схемы на всей области определения аргумента вектор начальных значений можно брать произвольным, для определенности и удобства вычислений – нулевым.

Решение задачи упрощается благодаря известным выражениям вектора градиента и матрицы Гессе ФМП. Градиент ФМП имеет явное представление в виде (опуская номер итерации)

$$g(B) = \sum_{j=1}^{n} (Y_j - P_j(B)) X_j .$$

Матрица Гессе ФМП имеет явное представление в виде (опуская номер итерации)

$$H(B) = -\sum_{j=1}^{n} P_j(B)(1 - P_j(B)) X_j^T X_j .$$

Заметим, что если по условиям задачи требуется логит множественной линейной регрессии со свободным членом, в режиме обучения в массив исходных данных добавляется столбец из одних единиц, а при распознавании необходимо в векторе каждого распознаваемого объекта также установить в данной позиции вектора единицу.

Оценка качества классификатора

Для оценки качества модели бинарного классификатора предложен ряд показателей. Рассмотрим их.

Дальнейшие обозначения проще всего пояснить с помощью таблицы 2 х 2, естественной для бинарных откликов.

Модель	Опыт	
	Положительный исход	Отрицательный исход
Положительный исход	T_P	F_P
Отрицательный исход	F_N	T_N

Суть обозначений ясна из первых букв английских терминов:

1. True – истинно.
2. False – ложно.
3. Positive – положительный.
4. Negative – отрицательный.

Термины «положительный» и «отрицательный» здесь относятся не к объекту исследования, а, скажем, к способности диагностического теста установить диагноз. Так, при исследовании заболевания положительным исходом будет являться наличие заболевания, отрицательным исходом – отсутствие заболевания.

На основе предложенных показателей строится особый график параметрического типа – ROC-кривая [4]. Термин ROC curve (Receiver Operating Characteristic Curve – ROC-кривая) в адекватном переводе, заимствованном из радиотехники, означает кривую соотношений правильного и ложного обнаружения сигналов. Абсцисса и ордината ROC-кривой являются функциями некоторого параметра, произвольно изменяемого или конкретно измеряемого в эксперименте. В исследовательской практике могут иметь место различные сочетания данных функций, что приводит к различным ROC-кривым [9]. Мы строим и анализируем наиболее употребительный тип ROC-кривой, параметрически отображающий величину чувствительности *Se* и величину неспецифичности 1 – *Sp*, где *Sp* – специфичность. Порог чувствительности на графике не отображается, однако каждому (в данном случае) заданному значению порога соответствует пара «чувствительность – неспецифичность». На графике данные величины принято изображать в процентах. Показатели определяются следующими формулами.

Чувствительность показывает долю истинно положительных случаев, т. е.

$$Se = \frac{T_P}{T_P + F_N} .$$

Специфичность показывает долю истинно отрицательных случаев, т. е.

$$Sp = \frac{T_N}{T_N + F_P} .$$

Некоторые авторы величину *Sp* называют частотой истинно отрицательных результатов (true negative rate), а величину 1 – *Sp* называют ценой метода либо частотой ложноположительных результатов (false positive rate, FPR). По аналогии величину *Se* иногда называют частотой истинно положительных результатов (true positive rate, TPR). Некоторые авторы полагают, что в таких терминах ROC-кривая более понятна для чтения. Также условились для построения ROC-кривой использовать показатели в процентах.

Сочетание значений чувствительности и специфичности в дальнейшем анализе может быть выбрано различным в зависимости от требований исследователя. При этом соответствующее значение диагностического параметра называют порогом отсечения. Выбор того или иного оптимального порога отсечения (а также любого другого желаемого порога, в том числе предлагаемой некоторыми авторами величины 0,5) производится на основе требований, предъявляемых исследователем к прогностическим характеристикам модели.

В представленном алгоритме используется критерий Юдена [1], [8], максимизирующий сумму чувствительности и специфичности *Se* + *Sp*.

Рассмотрим алгоритм построения ROC-кривой. Пусть даны исследуемая выборка численностью *n* и стандартная выборка численностью *m*.

Алгоритм ROC-анализа предлагается сформулировать следующим образом:

1. Объединить представленные выборки в массив диагностических параметров численностью *n* + *m* и упорядочить его по убыванию. Максимальный и минимальный эле-

менты объединенного массива будут соответствовать верхней и нижней границам интервала изменения параметра.

2. Используя варианты полученного в предыдущем пункте алгоритма массива диагностических параметров в качестве порогов отсечения, составить на основе исходных выборок для каждой варианты данного массива таблицу 2 х 2. При этом решающее правило имеет вид «параметр ≥ порога».

3. Подсчитать для каждой составленной в предыдущем пункте алгоритма таблицы чувствительность и неспецифичность. Массив чувствительностей численностью $n + m$ будет массивом абсцисс ROC-кривой. Массив неспецифичностей численностью $n + m$ будет массивом ординат ROC-кривой.

4. Построить график ROC-кривой по парам точек «абсцисса–ордината», полученным в предыдущем пункте алгоритма.

5. Подсчитать площадь, отсекаемую ROC-кривой.

Позиции 2, 3, 4 и 5 представленного алгоритма выгоднее выполнять в цикле по всем $n + m$ вариантам массива диагностических параметров.

Объективную оценку качества диагностического метода может показать площадь под ROC-кривой, в литературе кратко называемая AUC (Area Under Curve). Оценка данной площади подсчитывается по формуле трапеций:

$$\hat{A} = \frac{1}{2}\sum_{j=1}^{n+m-1}(Se_i + Se_{i+1})(Sp_i - Sp_{i+1}).$$

При расчете оценки площади условились использовать показатели в долях. Чем выше AUC, тем большую прогностическую ценность имеют представленные данные (представленный метод). Максимальное значение AUC равно 1. При значении AUC, равном 0,5, прогностическая ценность отсутствует. Возможна такая конфигурация исходных данных, что кривая ROC окажется ниже диагонали, а AUC окажется, соответственно, в интервале от 0 до 0,5. В этом случае следует изменить решающее правило (позиция 2 алгоритма) на противоположное: «параметр ≤ порога» – и выполнить алгоритм заново.

Стандартная ошибка оценки AUC подсчитывается по формуле [6]

$$SE(\hat{A}) = \sqrt{\frac{\hat{A}(1-\hat{A}) + (n-1)(Q_1 - \hat{A}^2) + (m-1)(Q_2 - \hat{A}^2)}{n \cdot m}},$$

где для краткости записи обозначено:

$$Q_1 = \hat{A} / (2 - \hat{A}),$$

$$Q_2 = 2\hat{A}^2 / (1 + \hat{A}).$$

Предложен метод сравнения двух ROC-кривых по отсекаемым ими AUC. Для этого используется статистика [5]

$$Z = \frac{\left|\hat{A}_1 - \hat{A}_2\right|}{\sqrt{SE(\hat{A}_1)^2 + SE(\hat{A}_2)^2}},$$

распределенная асимптотически нормально.

В соответствии с алгоритмом для оценки значимости AUC вычисляется статистика Z при сравнении оценки AUC для данной ROC-кривой с величиной AUC, равной 0,5 (случай «бесполезной» классификации). Статистика Z, вычисленная таким образом, позволяет объективно судить о статистической значимости полученной классификации. При этом $SE(0,5)$ вычисляется по показанной выше формуле.

Для вычисления двустороннего доверительного интервала оцениваемой AUC применяется формула:

$$I_{AUC} = (\hat{A} - \psi((1+\beta)/2) \cdot SE(\hat{A}); \hat{A} + \psi((1+\beta)/2) \cdot SE(\hat{A})),$$

где $\psi(.)$ – обратная функция стандартного нормального распределения; β – доверительный уровень, выраженный в долях.

В некоторых публикациях бытует ошибочное изображение ROC в виде кривой гладкой. Это демонстрирует непонимание авторами публикаций самой сути ROC-анализа как графического отображения результатов бинарной классификации. ROC-кривая – не график зависимости одной непрерывной величины от другой непрерывной величины. Параметрическая ROC-кривая может изображаться только в виде лесенки (в этом смысле традиционное название «кривая» – curve не является корректным). Она дискретна по своей природе, меняя значения абсциссы и ординаты скачками даже при непрерывном изменении порога отсечения, и не может быть гладкой.

Оценка значимости классификатора

Статистическая значимость весовых коэффициентов бинарного классификатора может проверяться с помощью статистик Вальда

$$W_i = \frac{\left|\hat{b}_i\right|}{\sqrt{Var(\hat{b}_i)}},\ i = 1, 2, ..., m,$$

где $\hat{b}_i$, $i = 1, 2, ..., m,$ – вычисленные оценки весовых коэффициентов; $Var(\hat{b}_i)$, $i = 1, 2, ..., m,$ – дисперсии оценок весовых коэффициентов.

Дисперсии оценок находятся как диагональные члены матрицы $I^{-1}(B)$, обратной к информационной матрице Фишера $I(B)$. Информационная матрица Фишера в данном случае представляет собой матрицу, элементы которой являются взятыми с обратным знаком элементами матрицы Гессе ФМП

$$I(B) = -H(B),$$

где $B = \{b_i\}$, $i = 1, 2, ..., m,$ – вектор-столбец весовых коэффициентов.

Статистики Вальда имеют стандартное нормальное распределение, что позволяет установить значимость вычисленных оценок весовых коэффициентов модели.

Заключение

В работе на примере множественной логистической регрессии рассмотрена оценка качества бинарных классификаторов. Произведены все необходимые теоретические выкладки, по результатам которых составлена библиотека исходных текстов программ на алгоритмическом языке Си, доступная под свободной лицензией согласно ГОСТ Р 54593-2011. Для пользователей разработанные алгоритмы доступны в виде модулей программы статистического анализа данных AtteStat (зарегистрирована Российским агентством по патентам и товарным знакам 28 июня 2002 г. за №2002611109).

ЛИТЕРАТУРА

[1] *Власов В. В.* Эффективность диагностических исследований. М. : Медицина, 1988. 256 с.

[2] *Носач В. В.* Решение задач аппроксимации с помощью персональных компьютеров. М. : МИКАП, 1994. 382 с.

[3] *Цыплаков А. А.* Некоторые эконометрические методы. Метод максимального правдоподобия в эконометрии : методическое пособие. Новосибирск : ЭФ НГУ, 1997. 129 с.

[4] *Fawcett T.* ROC graphs: Notes and practical considerations for researchers. URL: http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf (дата обращения: 19.09.2015).

[5] *Hanley J. A., McNeil B. J.* A method of comparing the areas under receiver operating characteristic curves derived from the same cases // Radiology. 1983. Vol. 148, № 3. P. 839–843.

[6] *Hanley J. A., McNeil B. J.* The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology. 1982. Vol. 143, № 1. P. 29–36.

[7] *Hosmer D.W., Lemeshow S.* Applied logistic regression. John Wiley & Sons, 2000. 375 p.

[8] *Youden W. J.* Index for rating diagnostic tests // Cancer. 1950. Vol. 3, № 1. P. 32–35.

[9] *Zweig M.H., Campbell G.* ROC Plots: A fundamental evaluation tool in clinical medicine // Clinical Chemistry. 1993. Vol. 39, № 4. P. 561–577.