Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка

Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.
kochedykov@forecsys.ru, andrej_iv@mail.ru, vokov@forecsys.ru
Москва, ЗАО «Форексис»

Принятие решений о выдаче кредитов физическим лицам — это стандартная задача классификации, в которой объектами являются клиенты, а признаки соответствуют полям анкеты, заполняемой клиентом при подаче заявки на выдачу кредита. В простейшем случае клиенты разделяются на два класса — «хорошие» (good) и «плохие» (bad).

Задача имеет следующие особенности: признаки разнотипны; в данных могут присутствовать пропуски и ошибки; объём обучающей выборки может быть крайне мал, в частности, при создании новых кредитных продуктов, при выходе на новые рынки или изменении экономической ситуации. Недоверие кредитных экспертов к «чёрным ящикам» влечёт ещё и требование интерпретируемости: алгоритм классификации должен быть прост и понятен, допускать осмысленную модификацию «вручную», а выдаваемые им решения — иметь логичные объяснения [1]. Этим требованиям удовлетворяют логические классификаторы, в частности, решающий список и взвешенное голосование конъюнкций [2].

Ещё одна важная особенность задачи заключается в том, что наряду с классификацией клиента алгоритм должен выдавать оценку вероятности дефолта (probability of default, PD), т. е. вероятности того, что клиент окажется «плохим» и не вернёт кредит или его часть. Оценки PD заёмщиков необходимы для анализа риска кредитного портфеля банка.

В мировой практике *кредитного скоринга* широко применяется логистическая регрессия, в которой оценки PD получаются естественным образом. Однако, по сравнению с логическими алгоритмами, она хуже интерпретируема, требует заметно бóльших объёмов обучающей выборки и основана на труднопроверяемых вероятностных предположениях.

В данной работе предлагается метод оценивания PD клиента для логических алгоритмов, основанный на понятии переобученности.

Переобученность логических закономерностей

Пусть X — множество объектов (допустимых описаний клиентов), $Y = \{-1, +1\}$ — классы «плохой», «хороший». Будем говорить, что предикат $\varphi\colon X \to \{0, 1\}$ *выделяет* объект x, если $\varphi(x) = 1$. Обозначим через $b(\varphi, U)$ и $g(\varphi, U)$ число объектов классов -1 и $+1$ соответственно, выделяемых предикатом φ из выборки $U \subset X$. Долю объектов класса -1, выделяемых предикатом φ из U, обозначим $\beta(\varphi, U) = \frac{b(\varphi,U)}{b(\varphi,U)+g(\varphi,U)}$.

Закономерность класса y — это предикат $\varphi(x)$, выделяющий достаточно много объектов класса y и достаточно мало объектов класса $-y$. Будем рассматривать логические алгоритмы классификации вида

$$a(x) = \arg\max_{y \in Y} \sum_{t=1}^{T_y} w_y^t \varphi_y^t(x), \tag{1}$$

где φ_y^t — закономерности, w_y^t — веса закономерностей, T_y — число закономерностей класса y. В данном исследовании закономерности φ_y^t строились в виде конъюнкций элементарных (однопризнаковых) пороговых предикатов [2]. Такие закономерности называются *правилами* (rules).

В общем случае для обучения алгоритма $a(x)$ по выборке U применяется некоторый метод поиска закономерностей $\mu\colon U \mapsto \{\varphi_y^t\}_{y\in Y}^{t=1,T_y}$. Закономерности отбираются по критериям $\min_\varphi \beta(\varphi, U)$ для класса $+1$ и $\max_\varphi \beta(\varphi, U)$ для класса -1. В результате оптимизации величина $\beta(\varphi, U)$ оказывается смещённой оценкой PD — заниженной для закономерностей класса $+1$ и завышенной для закономерностей класса -1. Несмещённую оценку $\mathrm{PD} = \beta(\varphi, V)$ можно было бы получить по независимой контрольной выборке V. Однако для надёжного оценивания требуются сотни контрольных объектов, а в условиях малого объёма данных всю выборку хотелось бы использовать как обучающую.

Возникает задача: спрогнозировать $\beta(\varphi, V)$, имея только информацию о закономерности φ, полученную на этапе обучения по выборке U.

Переобученностью закономерности $\varphi \in \mu(U)$ класса y будем называть величину $\delta(\varphi, U, V) = y\beta(\varphi, V) - y\beta(\varphi, U)$.

Эмпирическая методика оценивания переобученности

Заданная выборка $X^L \subset X$ длины L разбивается N способами на обучающую подвыборку длины ℓ и контрольную длины k, $X^L = X_n^\ell \cup X_n^k$, $L = \ell + k$, где индекс разбиения n пробегает значения от 1 до N.

В данной работе использовалась стандартная методика разбиения $t\times q$-fold cross-validation [3]: выборка X^L разбивалась t раз случайным образом на q блоков примерно равной длины и с равными долями классов, и каждый блок поочерёдно становился контрольной выборкой. Таким образом, $N = tq$ и $k \approx \frac{L}{q}$ с точностью до округления.

Пусть $F_n(\varphi) = F(\varphi, X_n^\ell, X_n^k)$ и $Z_n(\varphi) = Z(\varphi, X_n^\ell)$ — две числовые характеристики закономерностей $\varphi \in \mu(X_n^\ell)$. Зависимость F от Z будем оценивать *среднесглаженным* значением характеристики F в точке z:

$$F(z) = \frac{\sum_{n=1}^N \sum_{\varphi\in\mu(X_n^\ell)} \big[|Z_n(\varphi) - z| \leqslant \varepsilon\big] F_n(\varphi)}{\sum_{n=1}^N \sum_{\varphi\in\mu(X_n^\ell)} \big[|Z_n(\varphi) - z| \leqslant \varepsilon\big]},$$

где параметр ε играет роль ширины окна сглаживания. В эксперименте значение ε полагалось равным 2% от размаха распределения Z.

В роли характеристики Z будем рассматривать:

E: число ошибок на обучении, равное $b(\varphi, X_n^\ell)$ для закономерностей класса $+1$ и $g(\varphi, X_n^\ell)$ для закономерностей класса -1;
C: число объектов, выделяемых на обучении, равное $b(\varphi, X_n^\ell) + g(\varphi, X_n^\ell)$;
K: число элементарных предикатов в конъюнкции φ;
I: информативность (энтропию) закономерности φ на выборке X_n^ℓ;
R: рейтинг закономерности φ (номер в порядке убывания информативности) в списке всех предикатов, найденных и оцененных в процессе поиска закономерностей методом μ.

Для каждой характеристики Z из $\{E, C, K, I, R\}$ строится график зависимости среднесглаженной переобученности $F_n(\varphi) = \delta(\varphi, X_n^\ell, X_n^k)$ от Z (на Рис. 1 кривая с заштрихованной областью под ней). На графиках также выводятся среднесглаженные: 90%-й доверительный интервал переобученности (тонкие кривые), частота ошибок на обучении (нижняя жирная кривая) и на контроле (верхняя жирная кривая), а также число правил с данным значением характеристики Z (нижний график).

Эксперименты на 7 задачах из репозитория UCI (две из которых, `crx` и `german` — задачи оценки кредитоспособности), показали, что переобученность правил зависит от характеристик $\{E, C, K, I, R\}$ довольно сложным образом, и в разных задачах по-разному. Анализ графиков переобученности позволяет настраивать поиск закономерностей под конкретную задачу. Некоторые характерные примеры показаны на Рис. 1.

(C): закономерности с малым C переобучены сильнее.

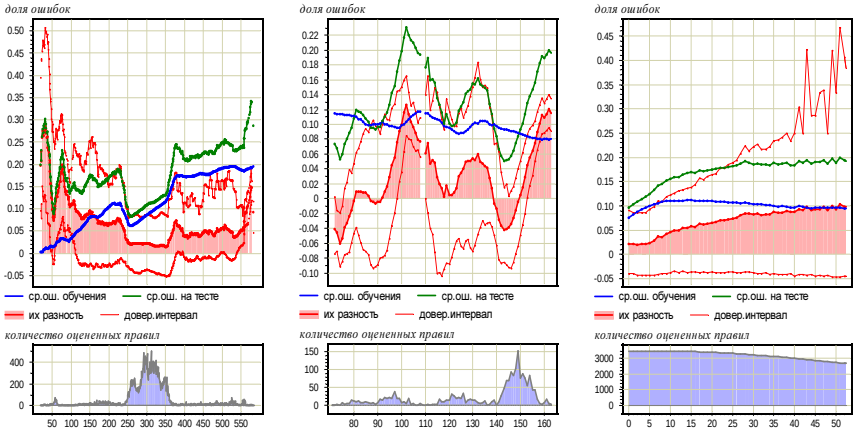
(I): при некоторых «особых» значениях информативности переобученность падает до нуля, что может сопровождаться всплеском числа найденных закономерностей.

(R): оптимизация улучшает качество закономерностей как на обучении, так и на контроле. В таких случаях можно усиливать оптимизацию, расширяя пространство поиска. Но в некоторых случаях зависимость оказывается противоположной, что свидетельствует о переобучении, и тогда пространство поиска следует сокращать.

Было также обнаружено, что число K элементарных предикатов в конъюнкциях практически не влияет на переобученность.

Методика оценивания $\mathrm{PD}(x)$ **для произвольного заёмщика** x

1. По данным скользящего контроля методами непараметрической регрессии оценивается зависимость переобученности $\delta(E, C, K, I)$ от характеристик E, C, K, I, вычисленных по обучающей выборке.
2. На этапе обучения алгоритм $a(x)$ строится по всей выборке X^L.



(C) (I) (R)

Рис. 1. Зависимость среднесглаженной переобученности правил от: (C) числа выделяемых объектов, задача `german`; (I) информативности (энтропии) правил, задача `crx`; (R) рейтинга правил, задача `german`.

3. Для каждой закономерности φ_y^t из (1) вычисляются характеристики E, C, K, I и по ним — оценка переобученности $\delta(\varphi_y^t) = \delta(E, C, K, I)$. При этом характеристики E, C, I приводятся к той длине обучающей выборки ℓ, при которой оценивалась регрессионная зависимость δ.
4. На этапе классификации объекта x оценка PD усредняется по всем закономерностям, выделяющим данный объект:

$$\mathrm{PD}(x) = \frac{\sum_{y,t} w_y^t \cdot \left[\varphi_y^t(x) = 1\right] \cdot \left(\beta(\varphi_y^t, X^L) + \delta(\varphi_y^t)\right)}{\sum_{y,t} w_y^t \cdot \left[\varphi_y^t(x) = 1\right]}.$$

Описанные методы реализованы в системе анализа кредитных рисков и поддержки принятия кредитных решений Forecsys Scoring Solution [4].

Литература

[1] *Соложенцев Е. Д., Степанова Н. В., Карасев В. В.* Прозрачность методик оценки кредитных рисков и рейтингов. — С.-Пб. ун-т, 2005. — 195 с.

[2] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Система кредитного скоринга на основе логических алгоритмов классификации // ММРО-12, Москва: Макс Пресс, 2005. — С. 349–353.

[3] *Webb G. I.* MultiBoosting: A technique for combining boosting and wagging // Machine Learning. — 2000. — Vol. 40, No. 2. — Pp. 159–196.

[4] `http://www.forecsys.ru/creditr.php` — Скоринг, анализ кредитных рисков и поддержка принятия кредитных решений. — 2005–2007.