УДК 519.254

Е. В. Стребков, В. С. Желтухин, И. А. Бородаев

ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ И АНАЛИЗ НАДЕЖНОСТИ СКОРИНГОВЫХ АЛГОРИТМОВ

Ключевые слова: обучающая выборка, скоринг, надежность скорингового алгоритма.

В работе изложен метод определения оптимального объема обучающей выборки, основанный на корреляционной связи между количеством верных предсказаний скорингового алгоритма и общим числом испытаний. Все построения ведутся на примере классификатора Байеса. Полученный результат может быть применен к произвольным скоринговым алгоритмам вне зависимости от их природы.

Key worlds: training selection, scoring, reliability of scoring algorithm.

A method of determination of optimum volume of the training selection is developed. Correlation relationship between quantity of right predictions of scoring algorithm and total number of tests is used as a base of the method. Bayes's qualifier is used as an example of mathematical argumentation. The method can be applied to arbitrary scoring algorithms.

Введение

Сложная ситуация, сложившаяся с кредитованием физических лиц, побуждает кредитные организации более ответственно подходить к выдаче кредитов. Для принятия обоснованного решения о выдаче кредитов применяется скоринг (от англ. scoring – подсчет очков), являющийся инструментом классификации клиентской базы на две группы: клиенты, которым можно выдать кредит и клиенты, кредитование которых рискованно [1]. В основе скоринга лежит предположение о наличии связи добросовестности заемщика с его показателями социального статуса и уровня финансовой обеспеченности (наличие детей, уровень образования, место работы, доходы и др.).

На практике скоринг-тест состоит из двух основных частей. Первая часть представляет собой опросник, отражающий социальные характеристики клиента, например: семейное положение; постоянство и уровень дохода; наличие финансовых обязательств; качество кредитной истории и т.п. (табл. 1) [2-4]. Показатели каждой характеристики раздельно ранжируются по их значимости для кредитоспособности заемщика. Вторая часть скоринг-теста включает метод классификации заемщиков на два класса: клиенты, которым можно выдать кредит; и клиенты, кредитование которых рискованно.

Таким образом, задачу скоринга можно рассматривать как задачу дискриминантного анализа, т.е. построения классификатора, с некоторым уровнем гарантии, на основе имеющейся выборки, который позволяет судить о принадлежности объекта $\mathbf{x}$ определенному классам Y_1 или, Y_2.

Построение классификатора

При наличии в кредитной организации достаточной информационной базы по выданным кредитам, для построения скоринговой системы необходимо определиться с алгоритмом классификации. Известно, что ни один из описанных методов не может быть признан «самым лучшим» во всех случаях. Только сопоставление предсказания и факта может дать оценку эффективности скоринговых моделей [5].

Одним из наиболее простых классификаторов является классификатор Байеса [6]. Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации.

Таблица 1 - Пример опросной анкеты банка

Параметр (Количество допустимых значений)	Допустимые значения
Семейное положение заемщика (2)	Женат (замужем) / одинок(а)
Наличие иждивенцев (2)	Есть / нет
Карьерный уровень должности клиента (4)	Рабочий / служащий / управленец среднего звена / топ-менеджер
Личный доход, тыс. руб. (6)	До 10 / 11-25 / 26-50/ 51-75 / 76-100/ более 100
Семейный доход, тыс. руб. (6)	До 10 / 11-25 / 26-50/ 51-75 / 76-100/ более 100
Наличие у клиента платежных обязательств (2)	Есть / нет
Кредитная история клиента в данном банке (3)	Нет / Положительная / С задержками /
Кредитная история клиента в других банках (3)	Нет / Положительная / С задержками /
Имущество (4)	Нет / дом / квартира / автомобиль
Клиент Банка (2)	Да / нет
Возраст (5)	18-25 / 26-40 / 41-50 / 51-60 / > 60
Своевременный возврат кредита заемщиком (2)	Да / нет

Наивный классификатор Байеса (Naive Bayes Classifier) основан на предположении о независимости отдельных компонент вектора $\mathbf{x}$, описывающего социальные характеристики потенциального заемщика (признаки объекта). Для классификации используется формула вычисления апостериорных вероятностей $P(Y_1|\mathbf{x})$, $P(Y_2|\mathbf{x})$. При этом вероятность $P(Y_i|\mathbf{x})$ принадлежности объекта x классу

Y_i , в силу предположения независимости компонент, вычисляется по формуле

$$P(Y_i|\mathbf{x}) = P(Y_i|x_1)\cdot P(Y_i|x_2)\cdot\ldots\cdot P(Y_i|x_n),$$

где $x_1, x_2, \ldots, x_n$ – компоненты вектора признаков $\mathbf{x}$. Объект $\mathbf{x}$ будет отнесен к тому классу Y_i, для которого выше $P(Y_i|\mathbf{x})$.

Однако, при формальном применении этой формулы для целей кредитного скоринга при незначительном отличии вероятностей $P(Y_1|\mathbf{x})$, $P(Y_2|\mathbf{x})$, велика вероятность ошибки ложного отнесения заемщика к категории надежных. Причин этого может быть несколько: отсутствие независимости компонент вектора признаков объекта, статистический характер корреляционной связи, невозможность полного учета всех факторов, влияющих на возврат кредита заемщиком, и др. В связи с этим построим модифицированный классификатор Байеса на основе обучающей выборки, включающей параметры, перечисленные в табл. 1.

В рамках предоставления кредитных продуктов для кредитора оказывается предпочтительнее не выдать кредит платежеспособному клиенту, чем предоставить неблагонадёжному. В этом случае сравнивается отношение апостериорных вероятностей с некоторой эмпирически заданной величиной (постоянная отсечения) C, то есть рассматривается соотношение: $P(Y_1|x)/P(Y_2|x)<>C$. Например, необходимо, чтобы мера того, что заёмщик благонадежный, в 2 раза превышала меру того, что он неплатежеспособен. Тогда C принимается равной 2.

Качество классификатора зависит от объема выборки. Чем больше объем, тем точнее предсказание. Однако, обработка большой выборки может быть проблематичной с точки зрения объема вычислений. В этой связи встает задача определения минимального объема обучающей выборки и параметра C, при которых достигаются приемлемые результаты классификации.

Анализ надёжности классификатора и подбор параметров

Для оценки качества бинарной классификации день широко применяются ROC-кривые – графики в декартовой системе координат xOy, где по оси Ox откладывается показатель специфичности алгоритма классификации, а по оси Oy – показатель чувствительности. Специфичность алгоритма классификации определяется как доля ошибочных положительных классификаций в общем числе отрицательных событий (ложно отрицательное множество), а чувствительность – как доля верно классифицированных положительных событий в общем количестве положительных классификаций (истинно положительное множество). Очевидно, что ROC-кривая располагается в единичном квадрате (рис. 1).

ROC-кривая, в частности, оценивает качество классификации величиной площади под ней. Чем больше площадь под ROC-кривой, тем выше качество классификатора, для которого она построена. С помощью ROC-кривых можно визуально оценить качество классификатора.

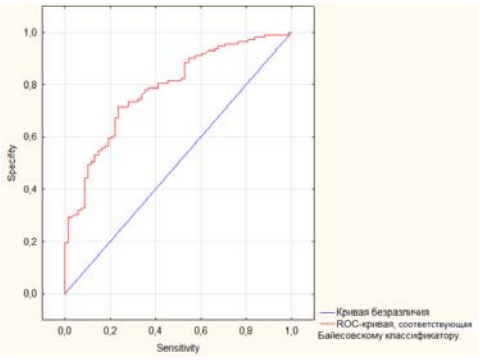

Рис. 1 - Пример ROC-кривой, соответствующей Байесовскому классификатору

Построение ROC-кривой производится в программных пакетах, таких, как R, SPSS и др. [7]. В результате этого исследователь лишен понимания механизма влияния параметров классификатора на ROC-кривую. Поэтому для подбора параметров классификатора и определения необходимого для достижения приемлемых результатов классификации объема обучающей выборки, ROC-кривые, на наш взгляд, не являются оптимальным решением.

В этой связи построим метод, позволяющий осуществлять подбор параметров классификатора на основании на корреляционной связи между количеством верных предсказаний скорингового алгоритма и общим числом испытаний,.

Рассмотрим признаки: X={рекомендация скоринг-теста} с возможными значениями «X+» и «X-» соответственно при положительной и отрицательной рекомендации по выдаче кредита; признак Y= {реальный исход кредитной истории} с возможными значениями «Y+» и «Y-» соответственно при положительной и отрицательной ситуации погашения задолженности заемщиком.

Таким образом, исходные результаты для анализа надежности скоринг-теста можно представить в виде следующей таблицы:

Таблица 2 - Результаты анализа обучающей выборки

$X\backslash Y$	Y+	Y-
X+	a	b
X-	c	d

Здесь $a+b+c+d=n$ – объем обучающей выборки, a, d – количество верно проклассифицированных объектов, b, c –количество ложно проклассифицированных объектов.

Для анализа корреляционной связи между дихотомическими признаками используются коэффициенты контингенции и ассоциации.

Коэффициент контингенции вычисляется по формуле:

$$k_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}. \qquad (1)$$

Коэффициент ассоциации вычисляется по формуле:

$$k_a = \frac{ad - bc}{ad + bc}. \qquad (2)$$

Пусть k- выборочное, а K– генеральное значение коэффициента ассоциации или контингенции. Статистическая значимость коэффициента k при нулевой (H_0: $K=0$) и альтернативной (H_1: $K\neq 0$) гипотезах проверяется на основе t–статистики Стьюдента с фактическим значением:

$$T_{a,k} = k_{a,k}\sqrt{\frac{n-2}{1-k_{a,k}^2}}, \qquad (3)$$

Под устойчивостью коэффициентов к соотношению пропорций a, b, c, d будем понимать независимость коэффициента от значений a и d при фиксированной их сумме. Исследуем устойчивость коэффициентов $k_{a,k}$ при изменении параметра a от 0 до 80 и фиксированных значениях n=100, b=c=10, a+d=80.

Из графика зависимостей $k_{a,k}$ и $T_{a,k}$ от параметра a (рис. 2), видно что:

- при фиксированной сумме a+d изменение соотношения между параметрами a и d приводит к значительным колебаниям коэффициента контингенции – от -0.1 до 0.6 и коэффициента ассоциации – от -1 до 0.88;

- экспериментальное значение статистики T_k изменяется от 1 до 7.4 и T_a – от - ∞ (при k_a=-1) до 18.56.

Для уровня значимости α=0.05 в интервале $5 \le a \le 75$ справедливо неравенство $|T_{k,a}|>T_{cryt}$=1.99, следовательно, значение генерального коэффициента $K_{k,a}>0$ и нулевая гипотеза отвергается.

Таким образом, можно сделать вывод, что коэффициенты ассоциации и контингенции являются недостаточно устойчивыми и в реальных задачах затруднительна однозначная их интерпретация и, следовательно, возможны неадекватные выводы об изучаемом явлении.

Рассмотрим коэффициент согласованности, который имеет вид [8, 9]:

$$k_s = \frac{a+d}{a+b+c+d}, \qquad (4)$$

и равен доле совпадений показаний признаков X и Y.

Для генерального коэффициента согласованности K_s доверительный интервал определяется соотношением:

$$P(k_s - \Delta < K_s < k_s + \Delta) = 2\Phi(t) = \alpha, \qquad (5)$$

где предельная ошибка выборки

$$\Delta = t\sqrt{\frac{k_s(1-k_s)}{n}}. \qquad (6)$$

Здесь k_s – выборочный коэффициент согласованности, α - доверительная вероятность, $\Phi(\tau)$ – значение функции Лапласа.

В равенстве (6) аргумент τ находится по таблице значений функции Лапласа из условия

$$\Phi(\tau) = \alpha/2. \qquad (7)$$

В отличие от коэффициентов контингении и ассоциации, коэффициент согласованности является устойчивым к изменению пропорций параметров a и d. Действительно, нетрудно видеть, что выражение $z=k_s(1-k_s)$ определяет график квадратичной параболы с максимальным значением z=0,25 при k_s=0,5 и, следовательно, из равенства (6) получается оценка

$$n < \tau^2/4\Delta^2 \qquad (8)$$

для необходимого объема выборки n, которое обеспечивает доверительный интервал (5) для генерального коэффициента согласованности K_s при заданных доверительной вероятности α и точности Δ.

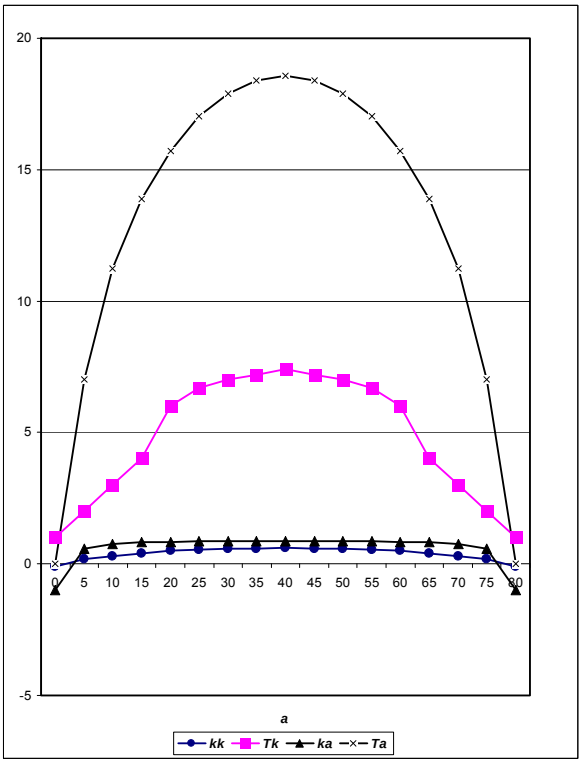

Рис. 2 - Зависимость коэффициента $k_{a,k}$ и статистик $T_{a,k}$ от параметра a

Параметры a, b, c, d связаны с C через алгоритм классификации. Максимизация коэффициента согласованности дает условие для подбора оптимального параметра C. Таким образом, задача отыскания оптимального параметра C имеет вид:

$$k_s \text{ s } \text{s®} \max; \quad C_{\min} \text{ J } C \text{ J } C_{\max}. \qquad (9)$$

Параметр C может быть определен с помощью компьютерного эксперимента, заключающегося в отыскании значений a, b, c, d и расчете коэффициента согласованности при различных C из интервала (C_{min}, C_{max}). За оптимальное принимается то значение параметра C, при котором коэффициент согласованности достигает своего максимального значения.

Выводы

По сравнению с существующими методами анализа надежности скоринговых тестов, метод, основанный на применении коэффициента согласованности в качестве доли (процента) совпадений рекомендаций скорингового алгоритма с фактическим результатом погашения кредита, является универсальным и математически обоснованным мето-

дом исследования надежности скоринговых алгоритмов. Данный метод обладает рядом существенных преимуществ:

1) простотой вычисления и наглядностью интерпретации надежности k_s в качестве доли (процента) совпадений показаний признаков X и Y;

2) устойчивостью при изменении пропорций между параметрами a и d при постоянстве суммы $a+d$;

3) наличием доверительного интервала (5) для оценки надежности K_s на генеральной совокупности и, следовательно, его оценки с необходимыми значениями доверительной вероятности α и точности Δ.

4) возможностью однозначного численного сравнения надежности различных скоринговых алгоритмов.

Литература

1. Э. Мэйз. *Руководство по кредитному скорингу*. Банксис, Москва, 2008. 464 с.
2. А.А. Строев. Расчеты и операционная работа в коммерческом банке. № 6. 28-33. (2004)
3. В.В. Карасев. Риск-менеджмент в кредитной организации, методический журнал. **10.** 2. 97-108. (2013).
4. J. Kruppaa, A. Schwarzb, G. Arminger, A. Ziegler Expert Systems with Applications. **40.** 5125–5131. (2013).
5. Tom Fawcett. HP Laboratories (2004)
6. С.И. Пшеничный. // Экономические науки. 2010. № 63, С. 306-310.
7. К. В. Воронцов, А.С. Инякин, А. В. Лисица. // *Тр. Всеросс. конф. ММРО-13.* «МАКС Пресс», Москва, 2007. С. 577-581.
8. Е. В. Стребков В сб. *Исследования по прикладной математике.* Вып.21. Унипресс, Казань, 1999. С. 228.
9. Стребков Е.В. В сб. материалов международ. научно-практ. конфер. *Логистическая интеграция российских регионов: Институциональные инновации.* Изд-во «Бриг», Казань, 2012. С. 255 -257.

© **Е. В. Стребков -** – к.ф.-м.н., доц. каф. мат. статистики КФУ; **В. С. Желтухин** – д.ф.-м.н., г.н.с. каф. ПНТВМ КНИТУ, vzheltukhin@gmail.com; **И. А. Бородаев** – экономист АКБ «Спурт» (ОАО), студ. КФУ.