СТАТИСТИЧЕСКИЙ АНАЛИЗ КРЕДИТОСПОСОБНОСТИ В УСЛОВИЯХ СКРЫТОЙ МАРКОВСКОЙ ЗАВИСИМОСТИ РЕЙТИНГОВ

А. Ю. Новопольцев, В. И. Малюгин

Белорусский государственный университет
Минск, Беларусь
E-mail: fpm.novopolc@bsu.by

Предлагаются алгоритмы статистической классификации предприятий на заданное число классов кредитоспособности в пространстве финансовых коэффициентов в условиях ненаблюдаемой марковской зависимости номеров классов (кредитных рейтингов). В предположении гауссовской модели наблюдений и постоянстве параметров модели проводится экспериментальное исследование алгоритмов, учитывающих и не учитывающих зависимость рейтингов для двух вариантов представлений исходной выборки в виде панельных и пространственных данных.

Ключевые слова: кредитные рейтинги, пространственные и панельные данные, кластерный анализ, дискриминантный анализ, алгоритм расщепления смеси распределений, скрытая цепь Маркова.

Модель данных и постановка задачи

Пусть в моменты (периоды) времени $t=1, \dots, T$ регистрируется информация относительно финансового состояния *n* предприятий одного вида экономической деятельности (отрасли), где *T* – длина периода наблюдения, выраженная числом кварталов (лет). Каждое предприятие i $(i=1, \dots, n)$ на конец отчетного периода *t* характеризуется вектором безразмерных финансовых коэффициентов $x_{i,t} \in \Re^N$ [1]. Предполагается, что каждое предприятие *i* в любой период *t* может быть отнесено к одному из *L* классов кредитоспособности. Номер класса является дискретной случайной величиной $\nu_{i,t} \in S(L) = \{1, \dots, L\}$, называемой рейтингом кредитоспособности предприятия. Временной ряд $\{\nu_{i,t}\}\,(t=1,\dots,T)$ описывается скрытой (ненаблюдаемой) однородной цепью Маркова (ОЦМ) с параметрами [2]:

$$\pi_0 = (\pi_{01}, \dots, \pi_{0L})'; \quad P = (p_{rs}),\, r, s \in S(L), \tag{1}$$

где $\pi_{0l} = \mathrm{P}\{\nu_{i,1} = l\} > 0$ – вектор вероятностей начального состояния, *P* – матрица миграции рейтингов, элементы которой $p_{rs} = \mathrm{P}\{\nu_{i,t+1} = s \mid \nu_{i,t} = r\} \geq 0$ – вероятности миграции рейтингов предприятий данной отрасли в течение одного периода. Распределение случайного вектора $x_{i,t}$ зависит от рейтинга $l \in S(L)$ и для фиксированных l, t описывается плотностью $f^{(t)}(u, \theta_l)$ из параметрического семейства:

$$\{f^{(t)}(u, \theta_l)\}\,(u \in \Re^N,\ \theta_l \in \Theta \in \Re^m,\ t=1, \dots, T,\ l \in S(L)). \tag{2}$$

При сделанных модельных предположениях выборка наблюдений $X=\{x_{i,t}\}\,(i=1,\,\dots\,,n,\ t=1,\,\dots\,,T)$ имеет панельную структуру, т. е. статистические данные X являются панельными (*panel data*) [3].

При проведении численных экспериментов на модельных данных используются дополнительные предположения о гауссовской модели наблюдений и постоянстве параметров модели, т. е. полагается, что для фиксированных l,t $(t=1,\,\dots\,,T,\ l\in S(L))$: $f^{(t)}(u,\theta_l)\equiv f(u,\theta_l)$ – плотность N-мерного нормального распределения $\mathbf{N}_N(\mu_l,\Sigma_l)$; $\theta_l\in\Re^m$ – составной вектор параметров, образованный из параметров μ_l,Σ_l при условии, что $\nu_{i,t}\equiv l$.

Задача. Параметры модели, $\pi_0,\ P,\ \{\theta_l\}$, а также рейтинги $\{\nu_{i,t}\}$ не известны. Задача заключается в их оценивании по наблюдаемым значениям $\{x_{i,t}\}$ $(i=1,\dots,n,\ t=1,\dots,T)$.

Алгоритмы оценивания и классификации

Предлагается использовать следующие алгоритмы классификации для двух альтернативных представлений исходной выборки наблюдений.

Алгоритм 1. Алгоритм расщепления смеси распределений в случае гауссовской модели наблюдений и скрытой марковской зависимости классов:

$$X=\{x_{i,t}\}\,(i=1,\dots,n,\ t=1,\dots,T),\ \ x_{i,t}\sim\mathbf{N}_N(\mu_l,\Sigma_l),\ l\in S(L), \tag{3}$$

позволяющий осуществлять совместное оценивание $\pi_0,\ P,\ \{\theta_l\}$ и $\{\nu_{i,t}\}$;

Алгоритм 2. Алгоритм кластерного анализа выборки пространственных (одномоментных) данных, не учитывающий марковскую зависимость классов:

$$Y=\{y_j\}\,(j=1,\dots,m),\ y_j\in\Re^N,\ m=nT, \tag{4}$$

полученной на основании (3) с помощью перенумерации наблюдений:

$$y_j\equiv x_{i,t},\quad j=(i-1)T+t,\ t=1,\dots,T,\ i=1,\dots,n, \tag{5}$$

предполагающий последовательное оценивание $\{\nu_{i,t}\}$ и $\pi_0,\ P,\ \{\theta_l\}$.

Алгоритмы состоят из следующих основных шагов.

Шаг 1. Классификация наблюдений из исходных выборок.

Алгоритм 1 на данном шаге в качестве алгоритма классификации выборки X использует итерационный алгоритм из класса EM-алгоритмов (Expectation-Maximization), учитыващий скрытую марковскую зависимость номеров классов и одновременно оценивающий все параметры модели, $\pi_0,\ P,\ \{\theta_l\},$ а также рейтинги кредитоспособности $\{\nu_{i,t}\}$. Данный алгоритм является частным случаем алгоритма, представленного в работе [4], с тем отличием, что для каждого состояния цепи Маркова возможно появление наблюдения только из одного распределения, а не из смеси распределений. Обозначим этот частный случай алгоритма через EM HMM (*EM for Hidden Markov Model).* Для оценки рейтингов на последней итерации используется алгоритм Витерби [2].

Алгоритм 2 представляет собой алгоритм L-средних [5] кластерного анализа выборки Y. Параметры $\{\Sigma_l\},P,\pi_0$ в данном случае вычисляются по классифицированной выборке $X = \{x_{i,t}\}$, полученной в результате обратного преобразования классифицированной выборки *Y*:

$$x_{i,t} \equiv y_j, \quad i:(i-1)T < j \le iT, t = j-(i-1)T, \ j=1, \dots, m.$$

Для ковариационных матриц $\{\Sigma_l\}$ используются несмещенные оценки, а для P, π_0 – оценки максимального правдоподобия [2].

Пусть $\lambda = \{\mu_l, \Sigma_l, l \in S(L); P, \pi_0\}$, тогда λ^1 и λ^2 – параметры, оцененные с помощью алгоритмов 1 и 2 соответственно на первом шаге.

Шаг 2. Дискриминантный анализ новых наблюдений.

Для классификации новых наблюдений применяются алгоритмы квадратичного дискриминантного анализа с учетом марковской зависимости классов (КДА-ОЦМ, [6]) и без ее учета (КДА, [5]), которые используют параметры λ^1 и λ^2 соответственно.

Для оценки влияния марковской зависимости классов на точность классификации с помощью алгоритма 2 используется следующая специальная модификация данного алгоритма.

Алгоритм 2.1. Данный алгоритм представляет собой комбинацию алгоритмов 1 и 2: на первом шаге к обучающей выборке *Y* применяется алгоритм кластерного анализа, а на втором для классификации новых наблюдений применяется алгоритм КДА-ОЦМ, использующий оценки параметров λ^2 для классификации выборки новых данных вида *X*.

Исследование алгоритмов на модельных данных

В условиях модельных предположений (1)–(3) с помощью статистического моделирования получена выборка наблюдений $X = \{x_{i,t}\}(i=1, \dots, n, \ t=1, \dots, T)$, $x_{i,t} \sim \mathbf{N}_2(\mu_l, \Sigma_l)$ $(l \in S(2))$, представляющая собой смесь $n = 300$ однородных цепей Маркова (ОЦМ) длины $T = 40$ с $L = 2$ состояниями (классами кредитоспособности) и параметрами, определяемыми соотношениями:

$$P = \begin{pmatrix} 0,8 & 0,2 \\ 0,1 & 0,9 \end{pmatrix}, \pi_0 = \begin{pmatrix} 0,4 \\ 0,6 \end{pmatrix}, \ \mu_1 = \begin{pmatrix} 7,0 \\ 1,0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1,0 \\ 0,1 \end{pmatrix}, \Sigma = \begin{pmatrix} 3,24 & 0,45 \\ 0,45 & 0,16 \end{pmatrix}.$$

Таким образом, классы кредитоспособности различаются только средними значениями случайного вектора наблюдений $x_{i,t}$.

Выбор размерности и значений параметров тестовой модели обусловлен достижением определенного сходства модельных и реальных данных по белорусским промышленным предприятиям, а также ориентацией на действующую в республике методику оценки кредитоспособности [7]. Указанная методика основана на анализе значений двух коэффициентов и классификации предприятий на два класса: кредитоспособных (Ω_1) и некредитоспособных (Ω_2).

Для исследования описанных алгоритмов на первом шаге используется неклассифицированная обучающая выборка, для которой $T_1 = 30$, а на втором – экзаменационная выборка, для которой $T_2 = 10$ $(T = T_1 + T_2)$.

Для всех алгоритмов на первом шаге применяется одна и та же случайная начальная классификация с равновероятным распределением классов (случай отсутствия априорной информации о кредитоспособности предприятий). Для алгоритма 1 также используется параметр сходимости $\varepsilon = 0,0001$ с условием остановки:

$$LL^{[k]} \geq LL^{[k-1]} \wedge (LL^{[k]} - LL^{[2]}) < (1+\varepsilon)(LL^{[k-1]} - LL^{[2]}),$$

где $LL^{[k]}$ – значение логарифмической функции правдоподобия на *k*-ой итерации.

Введем обозначения: *C*-1 и *C*-2 – истинные классификации, полученные в результате моделирования, для обучающей и экзаменационной выборок соответственно; *C*1-1 и *C*2-1 – классификации обучающей выборки, полученные на первом шаге алгоритмов 1 и 2 соответственно, а *C*1-2, *C*2-2 и *C*21-2 – классификации экзаменационной выборки на втором шаге алгоритмов 1, 2 и 2.1. Заметим, что результаты алгоритмов 2 и 2.1 на первом шаге совпадают, поэтому рассматриваются только результаты алгоритма 2. На основе перечисленных классификаций были рассчитаны средние (по отрасли) рейтинги, которые представляют собой средние значения номеров классов (рейтингов) в каждый момент времени. Для средних рейтингов приняты соответствующие обозначения *RC*-1, RC-2, *RC*1-1, *RC*2-1, *RC*1-2, *RC*2-2, *RC*21_2.

Для оценивания точности полученных результатов используются статистики: r – безусловная ошибка классификации для рейтингов; *MAPE* – средняя абсолютная ошибка в процентах (*Mean Absolute Percentage Error*) для средних рейтингов; показатели точности оценивания параметров модели (отклонения оценок от истинных значений) $\delta_x = \| \hat{\theta} - \theta \|$, $\delta_P = \| \hat{P} - P \|$, где $\| \cdot \|$ – евклидова норма, а $\hat{\theta}$ и $\hat{P}$ – соответственно составные векторы оценок параметров $\{\mu_l, \Sigma_l\}$ и P.

Результаты численных экспериментов представлены в табл. 1–2, а также на рис. 1.

Таблица 1

Анализ точности классификации

r, %					*MAPE*, %				
$T_1 = 30$		$T_2 = 10$			$T_1 = 30$		$T_2 = 10$		
*C*1-1	*C*2-1	*C*1-2	*C*2-2	*C*21-2	*RC*1-1	*RC*2-1	*RC*1-2	*RC*2-2	*RC*21-2
2,31	5,07	2,00	5,47	2,63	0,44	1,34	0,52	1,76	0,94

Таблица 2

Точность оценки параметров

EM HMM		*L*-средних	
δ_x	δ_P	δ_x	δ_P
0,0879	0,0176	0,6113	0,1617

Оценка вероятности ошибки байесовского решающего правила (с учетом марковской зависимости, [6]) равна 2,23 % для обучающей и 1,97 % для экзаменационной выборки, что немногим лучше результатов Алгоритма 1 (см. табл. 1, *C*1-1 и *C*1-2). Согласно табл. 1 и рис. 1, алгоритм 1 показал самую высокую точность классификации и оценивания параметров. Алгоритм 2.1 на втором шаге также продемонстрировал достаточно высокую точность классификации, практически сопоставимую с Алгоритмом 1 (имеет место увеличение вероятности ошибки всего на 0,63 %). В то время как точность оценивания параметров для алгоритма *L*-средних существенно (почти на порядок) ниже относительно алгоритма ЕМ НММ (см. табл. 2).

На основании полученных результатов можно сделать важный для практики вывод: в рамках используемых модельных предположений для классификации исходной выборки вместо трудоемкого алгоритма EM HMM можно применять более простой в вычислительном отношении алгоритм L-средних, используя при этом представление панельных данных в виде пространственной (одномоментной) выборки. В условиях большой размерности задачи (больших значений L, N, n и T) такая замена алгоритмов может позволить существенно сократить вычислительные затраты при сравнительно малых потерях точности результатов.

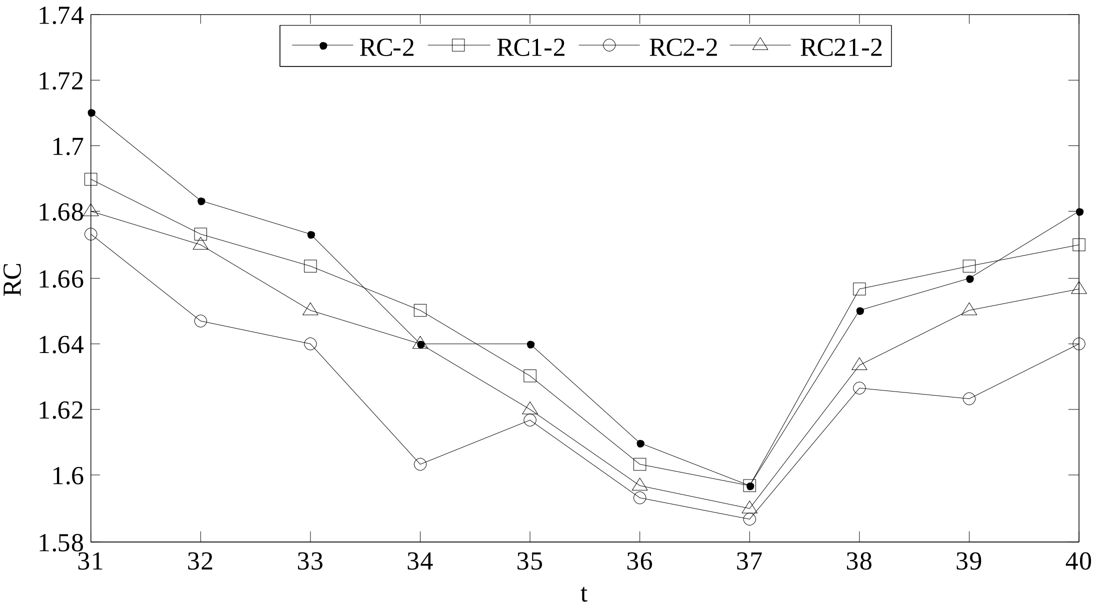

Рис. 1. Средние рейтинги для экзаменационной выборки

Библиографические ссылки

1. *Малюгин В. И., Корчагин О. И., Гринь Н. В.* Исследование эффективности алгоритмов классификации заемщиков банков на основе балансовых коэффициентов // Банковский Вестник. 2009. № 7. С. 26–33.
2. *Bhar R., Hamori S.* Hidden Markov models: Application to financial economics. Dordrecht: Kluwer Academic Publishers, 2004.
3. *Hsiao C.* Analysis of Panel Data. NY: Cambridge University Press, 2002.
4. *Bilmes Jeff A.* A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report TR-97-021. University of California at Berkeley, International Computer Science Institute and Computer Science Division, 1998.
5. *Харин Ю. С., Малюгин В. И., Абрамович М. С.* Математические и компьютерные основы статистического анализа данных и моделирования: учеб. пособие. Минск : БГУ, 2008.
6. *Харин Ю. С.* Обнаружение разладок марковского типа в случайной последовательности многомерных наблюдений // Статистические проблемы управления. Вильнюс. 1984. В. 65. С. 225–235.
7. Инструкция по анализу и конторолю за финансовым состоянием и платежеспособностью субъектов предпринимательской деятельности (в ред. постановления Министерства финансов, Министерства экономики и Министерства статистики и анализа Республики Беларусь от 8 мая 2008 г. № 79/99/50).