СТАТИСТИЧЕСКИЙ АНАЛИЗ КРЕДИТОСПОСОБНОСТИ В УСЛОВИЯХ СКРЫТОЙ МАРКОВСКОЙ ЗАВИСИМОСТИ РЕЙТИНГОВ

Новопольцев А. Ю., Малюгин В. И.
факультет прикладной математики и информатики, кафедра математического моделирования и анализа данных, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: {fpm.novopolc, malugin}@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.

I. Модель данных и постановка задачи

Пусть в моменты (периоды) времени $t = 1, \ldots, T$ регистрируется информация относительно финансового состояния n предприятий одного вида экономической деятельности (отрасли), где T – длина периода наблюдения, выраженная числом кварталов (лет). Каждое предприятие $i\,(i = 1, \ldots, n)$ на конец отчетного периода t характеризуется вектором безразмерных финансовых коэффициентов $x_{i,t} \in \Re^N$ [1] и, по предположению, может быть отнесено к одному из L классов кредитоспособности. Номер класса является дискретной случайной величиной $\nu_{i,t} \in S(L) = \{1, \ldots, L\}$, называемой рейтингом кредитоспособности предприятия. Временной ряд $\{\nu_{i,t}\}\,(t = 1, \ldots, T)$ описывается скрытой (ненаблюдаемой) однородной цепью Маркова (ОЦМ) с вектором вероятностей начальных состояний $\pi_0 = (\pi_{01}, \ldots, \pi_{0L})'$, $\pi_{0l} = P\{\nu_{i,1} = l\} > 0$ $(l \in S(L))$ и матрицей переходных вероятностей (матрицей миграции рейтингов) $P = (p_{rs})$, $p_{rs} = P\{\nu_{i,t+1} = s|\nu_{i,t} = r\}$ $(r, s \in S(L))$ [2]. Распределение случайного вектора $x_{i,t}$ зависит от рейтинга $\nu_{i,t} = l \in S(L)$ и для фиксированных l, t описывается плотностью распределения $f^{(t)}(u, \theta_l)$, $u \in \Re^N$, $\theta_l \in \Theta \subseteq \Re^m$. При сделанных предположениях выборка наблюдений $X = \{x_{i,t}\}$ $(i = 1, \ldots, n,\ t = 1, \ldots, T)$ является выборкой панельных данных (panel data) [3].

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

Задача. Параметры модели, π_0, P, $\{\theta_l\}$, а также рейтинги $\{\nu_{i,t}\}$ не известны. Задача заключается в их оценивании только по наблюдаемым значениям $\{x_{i,t}\}$.

II. Алгоритмы оценивания и классификации

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

Алгоритм 1. Алгоритм анализа панельных данных со скрытой марковской зависимостью классов, позволяющий осуществлять совместное оценивание π_0, P, $\{\theta_l\}$ и $\{\nu_{i,t}\}$ по выборке вида X.

Алгоритм 2. Алгоритм анализа выборки пространственных (одномоментных) данных $Y = \{y_j\}$ $(j = 1, \ldots, m)$, $y_j \in \Re^N$, $m = nT$, полученной на основании выборки X с помощью перенумерации наблюдений $y_j = x_{i,t}$, $j = (i-1)T + t$. Данный алгоритм не учитывает марковскую зависимость классов и предполагает последовательное оценивание $\{\nu_{i,t}\}$ и π_0, P, $\{\theta_l\}$.

Каждый алгоритм включает два шага: классификацию наблюдений из исходной выборки и оценку параметров на первом шаге и дискриминантный анализ новых наблюдений на втором. На первом шаге Алгоритм 1 использует EM-алгоритм, который учитывает скрытую марковскую зависимость номеров классов (EM-HMM,[4]), а Алгоритм 2 – алгоритм L-средних кластерного анализа [5], причем параметры $\{\Sigma_l\}$, P, π_0 в последнем случае вычисляются по классифицированной выборке $X = \{x_{i,t}\}$, полученной в результате обратного преобразования классифицированной выборки Y. На втором шаге Алгоритм 1 использует квадратичный дискриминантный анализ с учетом марковской зависимости классов (КДА-ОЦМ, [6]), а Алгоритм 2 – без учета данной зависимости (КДА, [5]).

Алгоритм 2.1. На первом шаге применяется алгоритм L-средних, а на втором – алгоритм КДА-ОЦМ.

III. Исследование алгоритмов на модельных данных

С помощью статистического моделирования получена выборка наблюдений $X = \{x_{i,t}\}$ ($i = 1, \ldots, n, t = 1, \ldots, T$), $\{x_{i,t}\} \sim \mathrm{N}_2(\mu_l, \Sigma_l)$ ($l \in S(2)$), представляющая собой смесь $n = 300$ однородных цепей Маркова длины $T = 40$ с $L = 2$ состояниями (классами кредитоспособности) и параметрами (1).

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 инициализации всех алгоритмов применяется одна и та же случайная классификация с равновероятным распределением классов (случай отсутствия априорной информации о рейтингах кредитоспособности).

Введем обозначения: C-1 и C-2 – истинные классификации для обучающей и экзаменационной выборок соответственно,полученные в результате моделирования; C1-1 и C2-1 – классификации обучающей выборки, полученные на первом шаге Алгоритмов 1 и 2 соответственно, а C1-2, C2-2 и C21-2 – классификации экзаменационной выборки, на втором шаге Алгоритмов 1, 2 и 2.1. Безусловные ошибки классификации приведены в табл. 1.

Таблица 1 – Ошибки классификации, %

C1-1	C2-1	C1-2	C2-2	C21-2
2.31	5.07	2.00	5.47	2.63

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. Согласно табл. 1, Алгоритм 2.1 на втором шаге также показал высокую точность классификации, сопоставимую с точностью Алгоритма 1.

На основании полученных результатов можно сделать важный для практики вывод: в рамках используемых модельных предположений для классификации исходной выборки и оценки параметров можно применять более простой в вычислительном отношении алгоритм L-средних, используя при этом представление панельных данных в виде пространственной 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.

1. Малюгин, В.И. Исследование эффективности алгоритмов классификации заемщиков банков на основе балансовых коэффициентов / В.И. Малюгин, О.И. Корчагин, Н.В. Гринь // Банковский Вестник. – 2009. – № 10. – С. 26–33.
2. Bhar, R. Hidden Markov models: Application to financial economics1 / R. Bhar., S. Hamori. – Dordrecht : Kluwer Academic Publishers, 2004. – 178 p.
3. Hsiao, C. Analysis of Panel Data / C. Hsiao. – NY : Cambridge University Press, 2002. – 366 p.
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 / Jeff A. Bilmes. – University of California at Berkeley, International Computer Science Institute and Computer Science Division, 1998.
5. Харин, Ю.С. Математические и компьютерные основы статистического анализа данных и моделирования : учеб. пособие / Ю.С. Харин, В.И. Малюгин, М.С. Абрамович. – Мн. : БГУ, 2008. – 455 с.
6. Харин, Ю.С. Обнаружение разладок марковского типа в случайной последовательности многомерных наблюдений / Ю.С. Харин // 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).

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}, \pi_0 = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}, \mu_1 = \begin{pmatrix} 7.0 \\ 1.0 \end{pmatrix}, \mu_2 = \begin{pmatrix} 1.0 \\ 0.1 \end{pmatrix}, \Sigma = \begin{pmatrix} 3.24 & 0.45 \\ 0.45 & 0.16 \end{pmatrix}. \quad (1)$$